NEW PERSPECTIVES ON DIMENSIONALITY REDUCTION AND SELECTIVE INFERENCE

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF STATISTICS
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Anav Sood
June 2025

© 2025 by Anav Sood. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This dissertation is online at: https://purl.stanford.edu/cf629qw1459

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Trevor Hastie, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Art Owen

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Jonathan Taylor

Approved for the Stanford University Committee on Graduate Studies.

Stacey F. Bent, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format.

Preface

This thesis considers topics in two sub-fields of statistics—interpretable dimensionality reduction (Chapter 2 and Chapter 3) and selective inference (Chapter 4, Chapter 5, and Chapter 6).

Interpretable dimensionality reduction concerns the problem of reducing a large dataset into a smaller one, while still ensuring the smaller dataset is readily interpretable. We consider the fundamental dimensionality reduction task of selecting the subset of variables that is most representative of the entire dataset. In Chapter 2, we show that two popular methods for this task, Column Subset Selection (CSS) and Principal Variables, are equivalent. This insight enables us to derive new methods for performing scalable CSS when (1) we have summary statistics in place of unit-level data, or (2) some of the unit-level data is missing and/or censored. In Chapter 3, we present a generative view of both of these methods. Specifically, we show that CSS and Principal Variables can both be interpreted as performing maximum likelihood estimation within a certain semi-parametric model. Within this model, we establish suitable conditions under which the CSS estimate is consistent in high dimensions, specifically in the proportional asymptotic regime where the number of variables over the sample size converges to a constant. Our generative viewpoint also leads to novel methodology for selecting the subset size for CSS/Principal Variables in a principled way.

Selective inference concerns the problem of drawing inferences after selection of a data-dependent inferential question. In Chapter 4, we provide a simple and unifying framework for conditional selective inference that is centered on p-values. Specifically, we introduce selectively dominant p-values, a class of p-values that enable us to easily perform inference after selection. We give an exact characterization of when p-values are selectively dominant and argue that essentially all commonly used p-values are selectively dominant. In Chapter 5, we use our framework to study the inference on winners problem. We offer greater clarity on the behaviors and properties of existing inference on winners methods, while also generalizing them to new settings. We also identify a class of problems for which the conditional selective inference approach attains optimal power, but existing simultaneous inference approaches are powerless. In Chapter 6, we motivate a very broad and general class of rank verification problems that arise in a variety of different domains. We use our framework to derive a new inferential method for these problems that dominates existing simultaneous inference approaches, and we demonstrate our method's efficacy and versatility by applying it to a wide range of real-world problems.

Acknowledgments

The PhD has been a long and rewarding journey, with plenty of ups and downs. There are many people in my life who have helped me reach the start of that journey, and many others who helped me see it through. I will acknowledge but a few of them here.

First, I have to give an immense thank you to my advisor, Trevor Hastie. I've had two great fortunes during my nine years at Stanford. The first was meeting Trevor when I was in undergrad. My friend Mason Swofford and I were in his STAT 315A course, and we must have both asked a million questions during Trevor's lectures and office hours. Trevor was always incredibly receptive, and he never failed to give patient and insightful answers. These conversations were what first sparked my interest in statistics. When I wasn't sure what to do after undergrad, or whether a PhD was even possible, Trevor was the main person who encouraged me to give it a shot. After I returned to Stanford for a PhD, my second great fortune was when Trevor agreed to be my advisor. Working with Trevor has been a huge pleasure. He has completely shaped how I approach, think about, and evaluate research, and his influence on me and my career can truly only be described as invaluable. Trevor was as supportive as I imagine an advisor can be. I always looked forward to our weekly meetings, regardless of whether it had been a good or bad week for me, and he always encouraged me to pursue my interests, even if they didn't always align with his. It's not lost on me how unbelievably lucky I am that this was the case, and I am very grateful to him for it. Also, it has been a huge privilege to work with someone who I, and many others in the department, look up to. Of course, I admire Trevor because he has made more interesting and influential research contributions than I can count. But more importantly, I really look up to Trevor because, even with all his unimaginable success (unimaginable for me at least!), he is such a humble, down-to-earth, fun, and kind person. I think everyone in the department would agree without hesitation that Sequoia Hall is a much better place when Trevor is around.

I would also like to thank some other Stanford faculty, starting with the rest of my thesis committee: Art Owen, Jonathan Taylor, Emmanuel Candès, and Lihua Lei. All of these people have been very helpful and instrumental throughout my PhD. I have to especially thank Jonathan. The biggest luxury of my PhD was getting interested in a field whose pioneer had an office just one floor below me. I would often stop by Jonathan's office unannounced, hoping to discuss ideas or ask questions. Even before we ever wrote anything down concretely, Jonathan would always understand what I was saying (always faster than what I imagined was possible), and he was always happy to discuss research and offer incredibly helpful advice. I would also like to thank Dennis Sun and Gene Kim for inviting me to co-author a new undergraduate probability textbook. Working on that project has been a wonderful opportunity for me, and it's been really exciting to see the book being used in the department's introductory offerings. Lastly, I would like to thank my undergraduate advisor, András Vasy. András taught me MATH 61CM, the first class I took at Stanford.

His lectures were energetic, infectious, and full of passion for mathematics. I, like many others, was nervous about taking such a difficult class. But after one lecture with András, I knew I was in the right place. Following the course, András remained a mentor of mine, and he was incredibly encouraging and supportive when I asked about applying for PhDs. I would not be here without his help and influence.

Briefly, I want to thank my high school crew: Vishie Betapudi, Nabhonil Kar, Matthew McMillan, and Ian Morrison. My academic journey started with this small group of overachievers. I have to give a special thanks to Ian, a friend of mine since I was five years old. We always had a little friendly competition in all of our classes, seeing who could best the other on the next test. I'm lucky to have met one of the smartest people in my life at such a young age, and always trying to measure up to him is what prepared me to get to Stanford and take full advantage of the opportunities I was afforded here.

I have to give a massive thank you to my friends from undergrad. There are too many to list, but I'll do my best: Ethan Shen, Cole Winstanely, Eli Pugh, Sofia Dudas, Eric Kilgore, Vienna Kuhn, Eva Frankel, Viraj Mehta, Marco De Leon, Eric Singh, Veer Shah, Dhruv Medarametla, Ani Jain, Taresh Sethi, Habeeb Jimoh, and Raagavi Ragothaman. If I did forget one of you, I'm really sorry. Let me know and I'll buy you a beer as repentance. The hardest part of the PhD was being away from you all. Thankfully, when the pandemic was still ongoing during my first year, I got to share a house in Boston with Eli, Dhruv, and Marco, with Ethan and Cole living not too far away. That period of time may have been the most fun I've ever had in my life (barring John Duchi's STAT 300B, which was a great class for my academic life, but unsurprisingly not so great for my personal one). I have to especially thank Dhruv, my fellow stats coterm who was always either one, two, or three steps ahead of me throughout all of undergrad. He's a brilliant guy and incredibly fun hang, and the tremendous amount of time and energy I've spent just trying to keep up with him is a big part of what's gotten me to where I am today. I also want to thank Viraj, my RA from undergrad who started his PhD a little before me. He has always offered me helpful advice and looked out for me. I hope one day I can return the favor. And finally, I want to thank my two favorite couples (by a mile!), Emily and Ethan, and Camille and Eli. You're all such wonderful people, and I'm so, so happy that all of you are now in the Bay Area. My happiness has been at least doubled since you all have been around. I hope all five of us (eventually six, I'm working on it) will continue to be in the same place for many years to come.

Next, I want to thank the wonderful PhD students I met in Sequoia Hall. Most people say the best part of doing a PhD in Stanford's Statistics Department is the people, and they're right. Sequoia Hall is full of smart, kind, fun, and interesting people; the exact kind of people you love to work with. I'll go somewhat chronologically, starting with Swarnadip. Swarnadip was a first year PhD student when I was an overly-ambitious undergrad trying to survive David Siegmund's offering of STAT 310B. I was lost right away, and I still vividly remember turning to the person to my right to ask for help. I was lucky the person was Swarnadip. Swarnadip was hugely helpful and incredibly encouraging (even after I made the wise choice of giving up and switching to pass-fail). This remained true all the way until his graduation, and he is a big part of why I ended up in Sequoia Hall myself. Along with Swarnadip, there were many 'older' PhD students I am lucky to call friends— Samir Khan, Isaac Gibbs, Kevin Fry, Yu Wang, and, my office mate, Kevin Guo. I have to give a special thank you to Kevin Guo, who is ridiculously smart and probably the most knowledgeable statistician I've ever met. He was a huge help to me early in my PhD, and gave me invaluable advice on how to approach research and write research papers. Despite my always coming in with twice as many questions as the day before, he was never hesitant to share his opinions, guidance, and time.

My PhD cohort is deserving of a special shout-out. It's been a real privilege to be part of such a diverse, interesting, and talented group. There are a few of you I'd like to thank individually, starting with Yujin Jeong, my first PhD friend and problem set buddy. Coming back from the pandemic, you reminded me how fun of a place Stanford can be. Working and hanging out with you was one of the big joys of my PhD. I have to give a big thank you to John Cherian and Tim Morrison—my main crew and the folks I spent the most time with during this whole ordeal. If I didn't have you two watching sports and movies with me, I think I would've gone insane. I hope our Slack group chat stays alive well after the PhD. I could write a whole acknowledgments section for them both, but I'll say just a couple things about each of them individually. Although Tim probably talks half as much as I do, he has at least twice the smarts and wits. I hugely admire his level-headedness and the way he approaches life, and I think the world would be better if we were all a little more like Tim. He's also a wonderful writer, and I have to thank him for regularly making paragraphs I spent hours writing better in minutes (I probably should've sent him this, but alas). John is an incredibly gifted statistician and, even more so, an incredibly gifted researcher (and even even more so he's a nice person and all that stuff too). Academia as an institution would benefit immensely if it was just full of a bunch of Johns—our students would have wonderful and engaging lectures everyday, and we would be hearing interesting talks and reading insightful papers all the time. If I happen to stumble on some substantial research discovery at some point in my life, I have no doubt it will be because I got to meet and learn from John during my PhD. Last but certainly not least, I have to thank James Yang. I like to think that I have a reasonably well-balanced skill-set. When working with someone who preferred to do theory, I'd try to pick up more of the coding labor. When working with someone who was a strong engineer, I would try and do more math. Unfortunately for me, this didn't pan out when I worked with James, because he was quite a bit better than me at both. Fortunately for me, he was willing to answer my every text and phone call at all hours of the day, and he was a huge help to me throughout my whole PhD (even after he graduated early!). He's had his hand in everything in this thesis, and I'm immensely grateful for him, his help, and his friendship. (Annoyingly, he's really good at piano too, and it's not even his first instrument...)

And to the youngins—Asher Spector, Will Hartog, Michael Howes, Rahul Kanekar, Tim Sidijuno, Yash Nair, Leda Liang, Aditya Ghosh, Etaash Katiyar, Jack Krew, Ginnie Ma, Alexandru Lopotenco, Skyler Wu, Louis Davis, Cherith Chen, Reese Feldmeier—you've all made Sequoia an incredibly fun place to be. It truly feels like it's getting more vibrant (and perhaps rowdy) by the year. I have to thank Leda in particular for always being up for non-work related activities and probably being the only reason I ever left campus. And I have to especially thank Yash. Given that he's younger than me, you might expect that I was the one typically lending a helping hand to him, but things were flipped the other way around more often than not. Getting stuck on a research problem can be very daunting, but it's quite a bit less scary when you know you can ask Yash for some help, and he'll likely solve it for you by the end of the evening. He's a brilliant statistician, and he's very giving and generous with his time, not just to me, but to others as well. Aside from all his brilliance, I'm very lucky to be able to call him a great friend. He's the person in Sequoia with whom I immediately felt most comfortable being myself, and I'm incredibly grateful for that.

I know this section is already abnormally long, but I can't wrap up without thanking the jazz community here at Stanford. Music is my solace, and I don't know if I would've made it through the PhD without it. The first person I have to thank is Jim Nadel, the founder of the Stanford Jazz Workshop. The first phone call I had after I was admitted to Stanford was with Jim. He told me about the vibrant jazz scene at Stanford, and how I'd fit right in. I didn't realize at the time what an integral role Jim would play in my life

for the decade to come. Jim has had an inspirational career here at Stanford, and he'll be leaving behind a legendary legacy and tremendous shoes to fill when he chooses to depart. He's positively impacted countless musicians' lives, and I hope to have a tenth of the impact he's had in whatever endeavors I pursue next. Next, I have to thank Joshua Redman, who started teaching at Stanford during my junior year of undergrad. Anyone familiar with jazz knows that Josh is a giant of the genre. He is a prolific performer and composer, and I never thought for one second that I'd be lucky enough to learn from him during my time at Stanford. I can only describe our overlap as an incredibly fortuitous blessing for me. Like Trevor, despite being at the absolute pinnacle of his field, Josh is one of the most humble and giving musicians I've ever met. I've learned so much from him these past seven years (including many things I never even knew I needed to learn!), and, thanks to him, I have slowly been becoming a musician that I don't think my past self would even recognize. There was a point at the start of my PhD where I was struggling a lot musically, and I was thinking about leaving music for good. It was Josh's words and presence that helped me find the joy in music again, and he's the reason I didn't make the misguided decision to stop playing. I'm indebted to him in a way that I struggle to put into words. Lastly, I have to thank the wonderful musicians I've had the luxury of playing with at Stanford. First, I'll start with Stanford's old guard—Evan Mickas, Josh Payne, Aris Kare, Chris Davis, Sean van Dril, Seamus Lynch. These are some seriously swingin' musicians who I'm lucky to call great friends. You would never guess from their playing that they share an equal expertise in non-musical pursuits as well. And finally, my current band—Michael Hayes, Ethan Buck, and Daiki Nakajima. You're such insanely talented and incredible musicians, and you've opened up a whole new world of music for me. There are times I'm genuinely in disbelief over the quality of music you three write and create together. I really hope we keep playing as a group for a long, long time.

Finally, I have to thank my family. First, my brother, Ayush Sood, who paved the way for me and has consistently made my life easier since the day I was born. Everything I have achieved up to this point has been built on what he started. In high school, my teachers paid extra attention to me because he was such a great student. I learned about Stanford when he was admitted, and he helped me put together my application when the time came. While I was at Stanford, he helped me look for internships and jobs, and he has continuously connected me with all sorts of people and resources that I would have never otherwise had access to. He's also bought me an unreasonable number of very expensive dinners, knowing full well that it is unlikely I will ever pay him back (what a guy). I couldn't ask for a better big brother than him. And most importantly, I have to thank my parents, Ajay and Apra Sood. I really think that they're the best parents in the world. They are incredibly loving, caring, open-minded, and understanding people, and I know that they'll always be there for me no matter the circumstances. The only reason I was able to start and finish a PhD at a place like Stanford is because of how they raised me. Like their parents before them, they championed the importance of education and prioritized their family over themselves. They offered me more opportunity than a kid could ever ask for, and in return asked not that I get perfect grades or first prize in some competition, but only that I gave my 100% effort in everything that I did. They sacrificed so much of their lives and careers for the sake of my and my brother's futures, and they provided us with the freedom and resources to pursue our passions, a luxury that they themselves never had. As a kid, I was oblivious to how much they centered their lives around the two of us, and how great we both had it. I understand that much better now. This thesis is a product of their sacrifice, and I dedicate it to them.

Contents

Preface iv

Acknowledgments v

1 Introduction 1
- 1.1 Interpretable dimensionality reduction . . . 1
  - 1.1.1 Motivation . . . 1
  - 1.1.2 Background . . . 2
- 1.2 Selective inference . . . 3
  - 1.2.1 Motivation . . . 3
  - 1.2.2 Background . . . 5

2 A covariance view of column subset selection 6
- 2.1 Notation and preliminaries . . . 7
- 2.2 The equivalence of CSS and Principal Variables . . . 7
- 2.3 Scalable CSS without unit-level data . . . 9
  - 2.3.1 Algorithms . . . 9
  - 2.3.2 An example from BlackRock . . . 11
  - 2.3.3 Other approaches . . . 12
- 2.4 CSS with missing and/or censored Data . . . 13
  - 2.4.1 A simple simulation study . . . 14
  - 2.4.2 An example with ozone level detection data . . . 14

3 A generative view of column subset selection 16
- 3.1 CSS as maximum likelihood estimation . . . 16
- 3.2 The subset factor model . . . 18
- 3.3 Selecting the subset size . . . 19
  - 3.3.1 A simulated example . . . 22
  - 3.3.2 Reducing survey length via our procedure . . . 22

4 A simple and unifying framework for selective inference 26
- 4.1 Notation . . . 26
- 4.2 Selective dominance . . . 27

4.3 Characterizing selectively dominant p-values and examples . . . 30
4.4 Example applications of selective dominance . . . 31
4.5 Publication-bias aware meta-analysis . . . 35
4.6 Some discussion . . . 36

5 Inference on winners via selective inference **37**
5.1 Conditional inference . . . 38
5.1.1 Confidence regions for the winner . . . 39
5.1.2 More discoveries via the closure principle . . . 41
5.2 Conditioning can gain power . . . 41
5.2.1 Power of the conditional approach . . . 41
5.2.2 Comparison to locally simultaneous inference . . . 43
5.2.3 The (seeming) necessity of conditioning . . . 46
5.3 Hybrid inference . . . 48
5.3.1 Confidence regions . . . 49
5.3.2 Comparison to the union bound . . . 49
5.3.3 Applying the closure principal . . . 50

6 Rank verification via selective inference **52**
6.1 Overview of approach . . . 54
6.2 Main results . . . 56
6.2.1 Getting lower confidence bounds . . . 57
6.2.2 Some special cases . . . 59
6.3 Related work . . . 60
6.4 Applications . . . 61
6.4.1 Ranking large language models from user preference data . . . 61
6.4.2 Ranking ResNet models . . . 63
6.4.3 Ranking clutch basketball players . . . 65
6.5 Rank verification in exponential families . . . 67

A Supplementary material relating to column subset selection **70**
A.1 Additional simulation details and results . . . 70
A.1.1 Missing data simulation setting . . . 70
A.1.2 Simulation setting for selecting subset size . . . 71
A.1.3 Additional simulations for selecting subset size . . . 72
A.1.4 Selecting subset size with a forward-backward algorithm . . . 72
A.2 More on selecting the appropriate subset size . . . 75
A.2.1 A finite-sample test when the unique factors are Gaussian . . . 76
A.2.2 An asymptotic test for the subset factor model . . . 77
A.2.3 Examination of the selected subset . . . 78
A.2.4 The PCSS model . . . 80
A.3 Efficient subset search for other objectives . . . 82
A.4 The high-dimensional consistency of CSS . . . 84

A.4.1 Central argument . . . 85
A.4.2 Establishing tail bounds . . . 87
A.4.3 Estimation with mean . . . 90
A.5 Additional theoretical results . . . 96
A.5.1 Estimation error of regression coefficients . . . 96
A.5.2 Selecting the best subset . . . 98
A.6 L^2 Random variables as a Hilbert space . . . 98
A.7 Proofs and derivations . . . 99
A.7.1 Proof of Lemma A.6.1 . . . 99
A.7.2 Proof of Theorem 12 . . . 100
A.7.3 Proof of Corollary 9 . . . 100
A.7.4 Proof of Proposition 1 and Proposition 2 . . . 100
A.7.5 Proof of Theorem 1 . . . 100
A.7.6 Proof of Theorem 2 . . . 103
A.7.7 Proof of Theorem 3 . . . 109
A.7.8 Proof of Lemma 1 . . . 110
A.7.9 Proof of Lemma A.2.1 . . . 110
A.7.10 Proof of Lemma A.2.2 . . . 111
A.7.11 Proof of Lemma A.2.3 . . . 113
A.7.12 Proof of Theorem 4 . . . 113
A.7.13 Proof of Lemma A.2.4 . . . 113
A.7.14 Proof of Lemma A.2.5 . . . 114
A.7.15 Proof of Proposition A.2.1 . . . 115
A.7.16 Proof of Lemma A.2.6 . . . 115
A.7.17 Correctness of subset search algorithms . . . 116
A.7.18 Proof of Proposition A.5.1 . . . 118
A.7.19 Proofs of finding best subset . . . 120

B Supplementary material relating to selective inference **123**
B.1 Permutation test p-value . . . 123
B.2 F-test p-value . . . 123
B.3 Selective dominance and independence . . . 123
B.4 Data carving for the file-drawer problem . . . 124
B.5 More info on data splitting, carving, fission, and thinning . . . 125
B.6 Post selection inference for the LASSO . . . 126
B.7 Comparing conditional and simultaneous inference . . . 128
B.8 Conditional inference on winners . . . 128
B.8.1 Standard derivation . . . 128
B.8.2 The p-value viewpoint . . . 129
B.8.3 Monotonicity of the selective p-value . . . 130
B.8.4 The conditional LCB and CI for MLR families . . . 130
B.8.5 Distance between winner and conditional LCB . . . 131

B.9 Conditional inference on winners for exponentials . . . 131
B.10 Hybrid inference on winners . . . 132
B.10.1 Standard derivation . . . 132
B.10.2 The p-value viewpoint . . . 133
B.10.3 The hybrid LCB and CI for MLR families . . . 135
B.10.4 Distance between winner and hybrid LCB . . . 136
B.11 Comparing hybrid inference to the union bound . . . 136
B.12 Number of ties in rank verification . . . 138
B.13 Selective p-value for rank verification in exponential families . . . 139
B.14 Rank verification additional details . . . 139
B.15 Selective dominance and one-sided testing . . . 141
B.15.1 MLR Families . . . 141
B.15.2 Exponential families . . . 146
B.15.3 Monotonicity of selective MLR p-values . . . 147
B.16 Selecting multiple p-values for inference . . . 148
B.16.1 Adaptive versions of Fisher's combination test . . . 150
B.16.2 Proof of Corollary 10 . . . 152
B.16.3 Proof of Corollary 11 . . . 153
B.17 Simultaneous approach . . . 154
B.17.1 Tukey's test . . . 154
B.17.2 A refinement of Tukey's HSD . . . 154
B.18 Confidence lower bound for rank verification . . . 156
B.19 Proofs . . . 157
B.19.1 Proof of Theorem 5 . . . 157
B.19.2 Proof of Theorem 6 . . . 159
B.19.3 Proof of Corollary 3 . . . 162
B.19.4 Proof of Corollary 2 and Corollary 4 . . . 163
B.19.5 Proof of Theorem 7 . . . 164
B.19.6 Proof of Theorem 8 . . . 167
B.19.7 Proof of Proposition 3 . . . 171
B.19.8 Proof of Theorem 9 . . . 171
B.19.9 Proof of Theorem 10 . . . 173
B.19.10 Proof of Corollary 5 . . . 175
B.19.11 Proof of Corollary 6 . . . 176
B.19.12 Proof of Corollary 7 . . . 176
B.19.13 Proof of Corollary 8 . . . 176

List of Tables

2.1 The proportion of times the correct subset was selected (% Corr. Subset Selected), the average size of the intersection of the selected subset and correct subset (# Corr. Variables Selected), and the population CSS objective (CSS Obj.) for the thousand simulated trials described in Section 2.4.1. The ± reports one standard error. Bold indicates best performance. 14

3.1 The number of questions attributed to each of the five personality factors for both the original and reduced surveys from Section 3.3.2. 24

3.2 Factor loadings (after rotation) from a exploratory factor analysis on the reduced 19-question survey from Section 3.3.2. As is common in the psychometrics literature, loadings with magnitudes larger than 0.30-0.32 are bolded [Costello and Osborne, 2005]. 25

6.1 The top-1 accuracy (i.e., the fraction of examples where the model's top prediction matches the true label) of the five pretrained ResNet models on the $m = 50000$ sample ImageNet Validation set. The number denotes the number of layers, so a larger number indicates a larger model. The model with the highest observed accuracy is bolded. 53

6.2 Arena scores of the $n = 20$ models from the publicly available LM Arena dataset [Chiang et al., 2024b]. 62

6.3 According to `inpredictable.com` [Beuoy, 2025], the observed clutch field goal percentage and number of clutch field goals taken for LeBron James, Kobe Bryant, Michael Jordan, Stephen Curry, Kevin Durant, and Kawhi Leonard. Player with the highest clutch field goal percentage is bolded. 65

A.1 A time complexity comparison of our algorithms to the naive algorithms that, when features are added or removed from the current subset, recompute the objective as written with no optimizations. The reported time corresponds to one complete run of Algorithm 1 or one iteration of Algorithm 2. McCabe himself pointed out that his first criteria of minimizing $|\boldsymbol{\Sigma}_{-S} - \boldsymbol{\Sigma}_{-S,S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S,-S}|$ is equivalent to maximizing $|\boldsymbol{\Sigma}_S|$, so we report the naive time complexities for both versions of the objective. 82

List of Figures

1.1 The first panel (left) shows the growth of the quantile $z_{1-\alpha_n}$ as a function of the dimension n. The second panel (right) gives the distance between the level $\alpha = 0.1$ conditional LCB and the winner X_W as a function of the gap $X_W - X_R$ between winner and the runner-up. For dimensions $n = 1, 10, 100, 1000$, it also gives the distance from X_W to the level $\alpha = 0.1$ simultaneous LCB. 4

2.1 For the BlackRock example discussed in Section 2.3.2, the average R^2 from the regression of each variable on the selected subset (resp. principal components) versus increasing subset size (resp. number of principal components). 12

2.2 For different amounts of missingness q and subset sizes k, the selected subset's CSS objective value on the fully observed Ozone Level Detection data over the hundred trials described in Section 2.4.2. The red line marks the CSS objective value attained by the subset that was selected using the fully observed data. 15

3.1 For one thousand trials, the size of the selected subset $\hat{S}$ (left), the size of the intersection of $\hat{S}$ and the population subset S (middle), and the population sum of squared canonical correlations between the variables in $\hat{S}$ and S (right) for the different simulation settings described in Section 3.3.1. Color indicates the average population R^2 from the regression of the variables not in S on those in S. 23

3.2 For the five personality sub-surveys, we give Cronbach's coefficient α with 95% bootstrap CIs ($B = 1000$ bootstrap samples) for our survey and the Norwegian [Engvik and Clausen, 2011], Brazilian [Veloso Gouveia et al., 2021], and Serbian [Tucaković and Nedeljković, 2023] 20-question surveys. The x-axis gives our number of selected questions per sub-survey. All other surveys chose four questions per sub-survey. 25

4.1 The first panel (left) depicts an example distribution of a p-value p under the null. The distribution stochastically dominates the uniform distribution. The second panel (middle) depicts the distribution $p|S = 1$ of the same null p-value p given that it was selected for being most $\alpha = 0.1$. This distribution does not stochastically dominate the uniform distribution. The third panel (right) depicts the null distribution $p_{sel}|S = 1$ of Theorem 5's selective p-value p/α given selection. Thanks to Theorem 5's correction, this distribution again is stochastically dominates the uniform distribution. 27

4.2 Scatter plot of the original p_O and replication p_R p-values for 92 psychology studies from the open science collaboration's replication analysis Collaboration [2015]. Note that the x-axis, which ranges from $[0, 0.05]$ is on a different scale than the y-axis, which ranges from $[0, 1]$. . 35

5.1 We plot the level $\alpha = 0.1$ conditional LCB $\hat{\mu}$ for different gaps between the winning value X_W and the runner up value X_R and highlight the distance between $\hat{\mu}$ and X_W in red. The LCB $\hat{\mu}$ is chosen exactly so that the tail probability $P(N(\hat{\mu}, 1) > X_R)$, shaded in blue, is $1/\alpha = 10$ times the tail probability $P(N(\hat{\mu}, 1) > X_W)$, shaded in red (the overlap appears purple). As X_W and X_R get closer, we need to take $\hat{\mu}$ further back for this condition to be satisfied. . . . 38

5.2 Over $B = 1000$ different sub-samples of $n = 2$ failure times from the dataset Molotaliev [2024], the distribution of the conditional and simultaneous LCB for the "winning" parameter. The conditional LCB is always vacuous. 40

5.3 Power of Corollary 1's procedure versus that of locally simultaneous inference ($\alpha = 0.1$, $\beta = \alpha/10$) for the sequence of needle-in-the-haystack alternatives $\mu^{(n)} \in \mathbb{R}^n$ given by $\mu_1^{(n)} = \sqrt{0.8 \log n^\zeta}$, $\mu_i^{(n)} = \sqrt{0.8 \log n^\zeta} - 2\sqrt{\log n}$ for $i = 2, \ldots, \lfloor n^\zeta \rfloor$, and $\mu_i^{(n)} = -n$ for $i = \lfloor n^\zeta \rfloor + 1, \ldots, n$, with $\zeta = 0.9$. Plot shows that LSI becomes powerless as $n \to \infty$, while the conditional method has power approaching one. Shaded region denotes one standard error bands. 44

5.4 For the $n = 20$ dimensional Gaussian problem $X_i \sim N(\mu_i, \sigma^2)$ with largest observation X_W and second largest observation X_R, the standardized distance $(X_W - \hat{\mu})/\sigma$ from X_W to the level $\alpha = 0.1$ hybrid, conditional, and simultaneous LCB $\hat{\mu}$ as a function of the standardized gap $(X_W - X_R)/\sigma$ between the winner and runner-up. For hybrid we take $\beta = 0.01$. The larger of the level $\alpha = 0.09$ conditional LCB and level $\alpha = 0.01$ simultaneous LCB are shown in green (i.e., the union bound LCB). 50

6.1 A screenshot displaying the top ten entries of the LM Arena leader board. Left-most column denotes the rank of the model on the leader board, which is determined by the model's "Arena Score" in the fourth column. The fifth column provides 95% bootstrap confidence intervals for the population arena score, but with no multiplicity correction, so these confidence intervals cannot be used to safely draw inferences about the population ranking. 53

6.2 For each sub-sample size on the x-axis, the five ResNet models from Table 6.1 are evaluated on $B = 100$ different random sub-samples of the size $m = 50000$ ImageNet validation set. The plots give the distribution of the resulting conditional LCBs from Algorithm 4 and simultaneous LCBs from Algorithm 3 (left), along with the proportion of times these LCBs are non-negative (right). 63

6.3 Considering three independent binomials $X_i \sim \text{Bin}(b, s_i)$ with $b = 4$ and $s_i = 1/2$, the first panel (left) depicts $N = 10^6$ draws from the conditional distribution of the p-value p_{12} (used for testing $H_{0,12} : s_1 \leq s_2$) given $W = 1$ and the nuisance statistics $(X_1 + X_2)/2 = 2$, $X_3 = 2$. When $p_{12} < q_{12}^+$, then X_1 is the sole winner and the p-value is selected for inference with probability one, but when $p_{12}^0 \in [q_{12}^+, q_{12}]$ there is a three-way tie and it is selected for inference with probability 1/3. Hence, the p-value's conditional distribution is not uniform on $[0, q_{12}]$ as Hung and Fithian [2019b] implicitly assume. The next panel (right) displays the consequence. Conditional on $W = 1$, Hung and Fithian [2019b] do not maintain Type I error control when testing H_{W2}^0 at level $\alpha = 0.1$ (denoted by horizontal dashed line), whereas our method does. Error bars denote a 99% confidence interval. 69

A.1 Results for the case of Gaussian unique factors. For description of the plots, see the caption of Figure 3.1 73

A.2 Results for the case of non-Gaussian unique factors. For description of the plots, see the caption of Figure 3.1 74

A.3 Replication of the simulations whose results are in depicted Figure 3.1, but now using our forward-backward algorithm to perform subset search. For description of the plots, see caption of Figure 3.1. 75

A.4 In the case of $p = 50$ variables and a size $k = 20$ subset S, the exact finite sample distribution of $T(S)$ for $n = 200$ (left) and $n = 5000$ (right) samples in the cases of both Gaussian and non-Gaussian unique factors, plotted against its limiting $\chi^2_{(p-k)(p-k-1)/2}$ distribution. 79

B.1 For $n = 10$ (left), $n = 100$ (middle), and $n = 1000$ (right) the distance between the hybrid LCB to the winner (dash-dot line) and union bound LCB to the winner (dotted and solid line) with $\alpha = 0.05$ and $\beta = 0.005$ plotted as a function of the gap between the winning and runner-up observation. 136

B.2 For $\mu_2 = 0$ (left), $\mu_2 = -0.5$ (middle), and $\mu_2 = -1$ (right) the empirical power over $N = 10^4$ trials of the hybrid inference approach versus the union bound approach for the needle-in-the-haystack alternative. One standard error bands are also plotted. For the most part, they are so small that they are hardly visible. 138

B.3 For $\ell = 3$ (left), $\ell = 5$ (middle), and $\ell = 10$ (right), power of the $k = 3$ conditional, $\tau = 0.5$ truncated, and original Fisher's combination test for data drawn from $N(\mu, I_{10})$ with $\mu_1 = \cdots = \mu_\ell$ varying according to the x-axis and $\mu_{\ell+1} = \cdots = \mu_n = -2$. Power results from an average over $N = 10^4$ trials, error bands (barely visible) denote one standard error, and the level $\alpha = 0.1$ is denoted by the dashed line. 151

Chapter 1

Introduction

This thesis considers topics in two sub-fields of statistics—interpretable dimensionality reduction (Chapter 2 and Chapter 3) and selective inference (Chapter 4, Chapter 5, and Chapter 6). The material in Chapter 2 and Chapter 3 is taken entirely from Sood and Hastie [2025], and the material in Chapter 4, Chapter 5, and Chapter 6 is taken largely from Sood [2024] and Sood [2025].

The introduction serves to provide some motivation and background on these areas, including a high-level review of related literature. We will continue to mention other related works throughout the thesis when relevant.

1.1 Interpretable dimensionality reduction

In modern data applications, it is common to reduce a large dataset into a smaller one, either for storage efficiency or ease of downstream analysis. As methods become more complex, it is increasingly desirable for this reduced dataset to be interpretable.

The first part of this thesis considers the dimensionality reduction task of selecting a subset of variables that is most representative of the entire dataset. We will establish an equivalence between two popular methods for this task, *Column Subset Selection (CSS)* and *Principal Variables*. Via this equivalence, we provide a generative view of both methods, as well as other extensions and generalizations.

1.1.1 Motivation

The textbook approach to dimensionality reduction is Principal Components Analysis (PCA) [Jolliffe, 2002]. PCA projects the observed data onto a small set of variance-maximizing axes. Ultimately, the reduced dataset consists of linear combinations of the original variables, and these resulting linear combinations are referred to as principal components.

The diverse range of perspectives on PCA has fostered many useful theoretical and methodological insights. For example, although PCA's objective is typically written in terms of unit-level data, it can also be written solely in terms of the data's covariance [Pearson, 1901]. This covariance characterization turns out to provide a more fundamental view of the problem, and Okamoto and Kanazawa [1968] strengthen PCA's theoretical foundation by using it to establish the method's simultaneous optimality for a large class

of objectives. Another example is probabilistic PCA (PPCA), a generative variant of PCA developed in Tipping and Bishop [1999]. PPCA broadens PCA's scope by enabling the application of standard statistical tools (e.g. likelihood ratio testing, expectation-maximization, Bayesian inference).

Unfortunately, the linear combinations that PCA outputs are typically dense and difficult to interpret. The field of interpretable dimensionality reduction has emerged largely in response to this phenomenon, and it is spearheaded by methods that output sparse linear combinations of the original variables [d'Aspremont et al., 2004, Jolliffe et al., 2003, Witten et al., 2009, Zou et al., 2006]. These sparse variants of PCA provide more interpretable results by simplifying the relationship between the reduced and original dataset. Methods like CSS and Principal Variables that select a subset of the original variables simplify this relationship as much as possible and, consequently, can be viewed as PCA's sparsest and most interpretable alternatives.

In the first part of this thesis, we aim to provide for CSS the same diverse set of viewpoints that are available for PCA. In Chapter 2, we show that Principal Variables, which is characterized in terms of covariances rather than unit-level data, solves the exact same problem as CSS. Following this insight, we provide a generative variant of CSS in Chapter 3 that facilitates the application of statistical tools. These covariance and generative viewpoints lead to novel methodological and theoretical insights for several problems in interpretable dimensionality reduction.

1.1.2 Background

To provide the necessary context for Chapter 2 and Chapter 3, we give a detailed description of CSS, Principal Variables, and the relationship (or lack thereof) between their corresponding literatures.

In the CSS problem we aim to choose k of p columns from a data matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ that best linearly reconstruct the rest:

$$\underset{S\subseteq[p]:|S|=k}{\operatorname{argmin}} \min_{\boldsymbol{B}\in\mathbb{R}^{k\times p}} \|\boldsymbol{X} - \boldsymbol{X}_{\bullet S}\boldsymbol{B}\|_F^2 \tag{1.1}$$

Here, $\boldsymbol{X}_{\bullet S}$ denotes the sub-matrix of $\boldsymbol{X}$ containing columns with indices in S. While computing the exact optimum in (1.1) is NP-complete [Shitov, 2021], a vast approximation literature [Boutsidis et al., 2009, Civril and Magdon-Ismail, 2012, Deshpande and Rademacher, 2010, Drineas et al., 2008, Farahat et al., 2015, Ordozgoiti et al., 2018, Tropp, 2009] has produced tractable algorithms with successful applications in epidemiology [Fink et al., 2007, Nowaková et al., 2021], networking [Tripathi and Reza, 2020], business [Boutsidis et al., 2008], imaging [Krömer et al., 2018, Strauch, 2014, Wang et al., 2015], and low-rank matrix approximation [Boutsidis et al., 2014, Dan et al., 2015].

Prior to gaining popularity in computer science, the CSS problem first garnered attention in the numerical linear algebra community. Many foundational works in numerical linear algebra, past and present, study criteria similar to CSS [Baker and DeGroat, 1998, Bischof and Quintana-Ortí, 1998, Chan, 1987, Chan and Hansen, 1994, Chandrasekaran and Ipsen, 1994, Chen et al., 2023a, Foster, 1986, Golub, 1965, Gu and Eisenstat, 1996, Gu and Miranian, 2004, Hong and Pan, 1992, Pan, 2000, Pan and Tang, 1999, Stewart, 1999a, 1998, 1999b]. The articles referenced in the previous paragraph extend this line of research by focusing specifically on the CSS problem and providing either improved theoretical guarantees or superior empirical performance for the objective (1.1).

Principal Variables [McCabe, 1984] puts a non-degenerate $N(0, \boldsymbol{\Sigma})$ model on the observed data and suggests selecting a subset of "principal variables" according to various information criteria. The most

popular of the four criteria suggests selecting the subset S that solves

$$\underset{S\subseteq[p]:|S|=k}{\operatorname{argmin}} \operatorname{Tr}(\boldsymbol{\Sigma}_{-S} - \boldsymbol{\Sigma}_{-S,S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S,-S}) \tag{1.2}$$

where $-S$ is the complement of S and $\boldsymbol{\Sigma}_S$ and $\boldsymbol{\Sigma}_{-S}$ are the principal submatrices of $\boldsymbol{\Sigma}$ whose rows and columns are subseted by S and $-S$ respectively. In practice, an estimate $\hat{\boldsymbol{\Sigma}}$ of the covariance matrix is used in (1.2). Like the CSS literature, the Principal Variables literature has produced approximation algorithms [Brusco, 2014, Cadima et al., 2004, Guo et al., 2002, Masaeli et al., 2010, Wei and Billings, 2006] that have enjoyed application in a number of fields, including psychology [Fehrman et al., 2017, Ladd and Golter, 1988], imaging [Chang et al., 1999], and the natural sciences [Eitrich et al., 2007, Isaac et al., 2014].

Expectedly, each thread of literature has something fruitful to offer: principal variables have strong statistical motivation and practitioners can find them in settings where only covariance or correlation estimates are available. We provide examples of such settings in Section 2.2. The CSS literature places larger emphasis on providing algorithms with theoretical guarantees and minimal computational and storage complexity.

Despite similarities between the methods, the CSS and Principal Variables literatures have existed entirely independently. For example, Masaeli et al. [2010] pose a regularized matrix reconstruction problem in the style of (1.1), but cites Principal Variables as motivation and makes no reference to any CSS literature.

1.2 Selective inference

Selective inference is a sub-field of statistics that allows practitioners to make valid inferences even when the underlying statistical question is chosen by a data-driven selection process. However, many selective methods, which operate by conditioning on the selection event, can be difficult to derive, hard to implement, and exhibit counter-intuitive behavior. To statisticians outside of the sub-field, each selective procedure may seem to come from a different argument or approach.

The second part of this thesis provides a unifying framework for selective inference centered around p-values. So long as a statistician knows how to compute traditional p-values, our framework enables them to easily deliver valid hypothesis tests and confidence intervals, even when the inferential question at hand arose from an arbitrary, albeit known, selection procedure. Our framework (1) can greatly simplify the process of designing new selective methods and (2) results in more natural and general derivations of some existing selective methods, allowing for a deeper understanding of their behavior as well as new variations and extensions. After presenting the framework in Chapter 4, we apply it to derive new methodological and theoretical results for two canonical selective inference questions—inference on winners in Chapter 5 and rank verification in Chapter 6.

1.2.1 Motivation

To motivate the need for a fresh view of selective inference, we recall the age-old task of making inferences from the largest observation. Considering Gaussian data $X \sim N(\mu, I_n)$ with unknown mean μ, the largest observation $W = \operatorname{argmax}_{i\in[n]} X_i$ may well correspond to the largest mean. Therefore, a natural way to verify the existence of a large μ_i is to give a lower confidence bound (LCB) for μ_W, the mean of the winning value.

Traditionally, we provide such an LCB via Sidak's simultaneous approach. Let $z_{1-\alpha}$ denote the $1-\alpha$

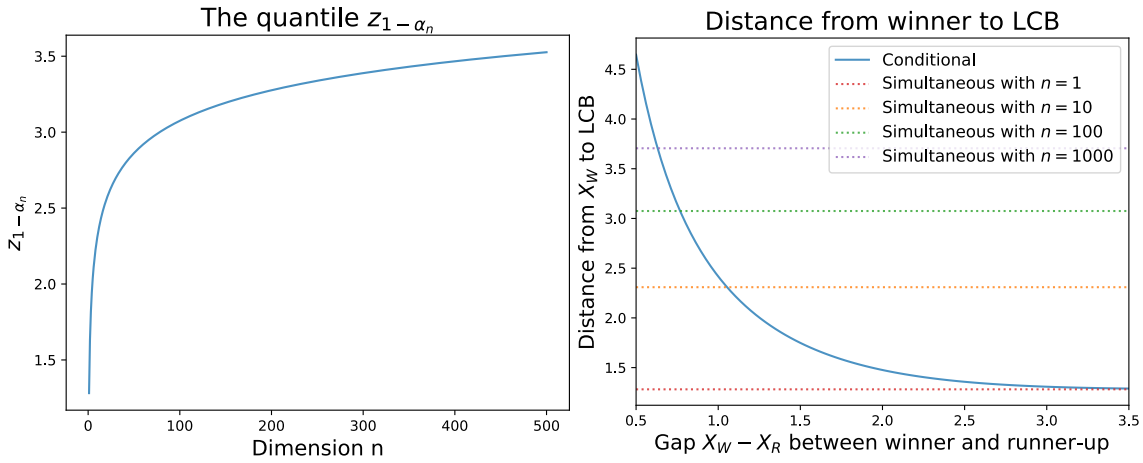


Figure 1.1: The first panel (left) shows the growth of the quantile $z_{1-\alpha_n}$ as a function of the dimension n. The second panel (right) gives the distance between the level $\alpha = 0.1$ conditional LCB and the winner X_W as a function of the gap $X_W - X_R$ between winner and the runner-up. For dimensions $n = 1, 10, 100, 1000$, it also gives the distance from X_W to the level $\alpha = 0.1$ simultaneous LCB.

quantile of a standard normal distribution and define $\alpha_n = 1-(1-\alpha)^{1/n}$. Sidak's approach tells us that the lower bounds $\mu_i > X_i - z_{1-\alpha_n}$ all simultaneously hold with probability $1-\alpha$. Therefore, $\hat{\mu}_{simul} = X_W - z_{1-\alpha_n}$ is a valid lower bound for μ_W that holds with probability at least $1-\alpha$. Performing simultaneous inference on n means, however, comes at a cost. As n grows, the quantile $z_{1-\alpha_n}$ grows as well (depicted in the left panel of Figure 1.1), and the distance from the winning observation to the LCB correspondingly increases.

The more modern selective approach suggests providing an LCB that is valid conditionally on W. This approach provides inferences solely for the winning mean μ_W, so we may hope that it avoids the simultaneous inference's curse of dimensionality. Following the recipe of Fithian et al. [2017] (see Appendix B.8 for details), one finds that the conditional LCB $\hat{\mu}_{cond}$ is the infimum of the confidence region

$$\{\mu_0 : \mu_0 > X_W - \text{Quantile}(1-\alpha, TN(0,1,X_R-\mu_0,\infty))\}, \tag{1.3}$$

where X_R is the runner-up (second largest) observation and $TN(\mu,\sigma^2,a,b)$ is a $N(\mu,\sigma^2)$ distribution truncated to lie in the interval (a,b).

The conditional LCB (1.3) is near impossible to parse at first glance, but it turns out to have some very interesting properties. As plotted in right panel of Figure 1.1, the distance $X_W - \hat{\mu}_{cond}$ between the winner and the conditional LCB is purely a function of the gap $X_W - X_R$ between the winner and runner-up. If the gap between X_W and X_R is large, then the conditional LCB for μ_W will be roughly $X_W - z_{1-\alpha}$, i.e., what we expect in a one-dimensional inference problem. But as the runner-up gets close to the winner, the conditional LCB explodes quickly to $-\infty$ and gives much worse inferences than the classical approach. In summary, the conditional method appears to avoid the curse of dimensionality when the winner is well separated, but when it is not, the consequences can be tremendous.

The motivation for our framework comes from the following fact: the conditional LCB becomes shockingly simple to parse once written in terms of p-values. Suppose we wanted just to check whether each LCB was non-negative, i.e., we used each LCB solely to verify the existence of a positive mean. Typically, we verify the existence of a positive mean by using the p-values $p_i = 1-\Phi(X_i)$ to test the nulls $H_{0,i} : \mu_i \leq 0$. It turns

out that the simultaneous LCB verifies the existence of a positive mean when smallest of these p-values $p_{(1)}$ is at most α_n. In contrast, the conditional LCB does so when the ratio of the two smallest p-values $p_{(1)}/p_{(2)}$ is at most α.

At least when the p-values p_i are uniform under the null, there appears to be a simple and intuitive proof for the conditional procedure's validity (much simpler than our derivations in Appendix B.8): under the null, the smallest p-value $p_{(1)}$ should be uniform on $[0, p_{(2)}]$, so we should control error if we reject when $p_{(1)} \leq \alpha p_{(2)}$. But the question remains, does this argument go through when p-values are not exactly uniform? If not, why does it work in the Gaussian case?

The framework we develop in Chapter 4 gives a complete answer to these questions. It identifies a broad class of p-values for which the sort of argument we have outlined above goes through. Beyond inference on winners, it turns out to encompass a comprehensive set of selective inference problems, and it provides a clearer view on how selective methods behave and are derived. As we illustrate throughout the article, our framework serves to both deepen our understanding of existing methods in the selective inference literature, as well as enable easy design of new selective inference procedures.

1.2.2 Background

The field of modern conditional selective inference was started by the seminal work Lee et al. [2016b] and has been the subject of much academic research over the last decade [Benjamini et al., 2019, Chen et al., 2023b, Chen and Witten, 2023, Fithian et al., 2015, 2017, Gao et al., 2024, Hyun et al., 2021, Lee and Taylor, 2014, Loftus and Taylor, 2015, Panigrahi et al., 2024, Taylor and Tibshirani, 2018, Taylor et al., 2014, 2016, Tian and Taylor, 2018, Tian et al., 2018, Tibshirani et al., 2016]. Alongside this contemporary body of work, there is an earlier line of work that studies drawing inferences from data-dependent questions, particularly as it pertains to rank verification Bofinger [1985a], Fabian [1962], Gutmann and Maymin [1987], Hsu [1981a], Maymin and Gutmann [1992]. Outside of conditional selective inference, there is a class of selective methods that avoid conditioning on the selection event (e.g, [Romano et al., 2014, Zrnic and Fithian, 2024]), but Perry et al. [2024] suggest empirically that such approaches may be less powerful than their conditional counterparts, and we provide some theoretical justification to this claim.

We point out a few articles that are most relevant to our framework. For continuous data (i.e., non-discrete data) Hung and Fithian [2020], Zhao et al. [2019] both identify similar classes of p-values and correct for selection identically to our approach. Their results, however, only apply when the selection mechanism selects the p-value for being below a fixed threshold. Our framework is fully general and applies (1) when the data is non-continuous (including when data is discrete) and (2) for all selection mechanisms, including those involving randomized selection. Hung and Fithian [2020], Zhao et al. [2019]'s work is done for the sake of a one-off application. In contrast, we contextualize our more general theory as the backbone that supports all of modern conditional selective inference, and demonstrate its applicability across a range of problems. While we assume the selection procedure is known, Andrews and Kasy [2019], Markovic et al. [2019] consider settings where the selection mechanism is unknown, and provide either (1) methods for learning the selection mechanism or (2) theoretical results regarding to what extent the selection mechanism is identified. When we treat the procedures we derive as global null testing procedures and close them, we get sequential selective hypothesis testing procedures that are similar to those in Fithian et al. [2015], although the procedures in Fithian et al. [2015] specifically test two-sided nulls, while our nulls are arbitrary.

Chapter 2

A covariance view of column subset selection

In this chapter, we establish a mathematical equivalence and conceptual connection between CSS and Principal Variables. This equivalence is not mathematically difficult to establish, but it unlocks new perspectives that enable us to design novel and useful methodology for several problems in interpretable dimensionality reduction. Particularly, it leads to two applications.

Scalable CSS without unit-level data: We introduce fast algorithms for performing CSS when only summary statistics are available. To demonstrate our algorithms' efficacy, we provide a real world example where BlackRock, our industry affiliate, wants to perform CSS but only has access to a covariance estimate. Our insights enable us to perform CSS in this setting, and our novel algorithms enable us to run the appropriate experiments, which would take days using current methods, in less than three minutes.

CSS with missing and/or censored data: The equivalence between CSS and Principal Variables suggests a natural workflow for performing CSS when data is missing and/or censored: model the missingness and/or censorship, produce a covariance estimate that is reasonable under this model, and perform CSS using this covariance estimate. As an illustrative example, we use this workflow to propose a simple and novel procedure for performing CSS with missing-at-random data. In a simulation study where we sample data from Chapter 3's generative model and omit entries at random, our method always selects the correct subset while the best competing method selects it less than half the time. We also confirm that our method outperforms existing methods on a real dataset.

To start, we introduce some notation and preliminaries. The same notation will be used in Chapter 3.

2.1 Notation and preliminaries

We use the following linear algebra notation. All matrices are bolded capital letters. We use $||\boldsymbol{A}||_F$ to denote the Frobenius norm of the matrix $\boldsymbol{A} \in \mathbb{R}^{p\times q}$. We denote the Moore-Penrose inverse, or pseudo-inverse, of $\boldsymbol{A}$ by $\boldsymbol{A}^+$, and the ith row and jth column of $\boldsymbol{A}$ by $\boldsymbol{A}_{i\bullet}$ and $\boldsymbol{A}_{\bullet j}$ respectively. For subsets $S \subseteq [p] = \{1, \ldots, p\}$ and $T \subseteq [q]$, $\boldsymbol{A}_{S,T}$ is the $|S| \times |T|$ sub-matrix of $\boldsymbol{A}$ with rows and columns indexed by S and T respectively. If we want to index just the rows (resp. just the columns), we use $\boldsymbol{A}_{S\bullet} \in \mathbb{R}^{|S|\times q}$ (resp. $\boldsymbol{A}_{\bullet T} \in \mathbb{R}^{p\times|T|}$). Indexing always precedes matrix operations, e.g., $\boldsymbol{A}^+_{S,T} = (\boldsymbol{A}_{S,T})^+ \neq (\boldsymbol{A}^+)_{S,T}$. When $\boldsymbol{A} \in \mathbb{R}^{p\times p}$ is square, we let $\boldsymbol{A}_S \in \mathbb{R}^{|S|\times|S|}$ denote the sub-matrix with rows and columns both indexed by S. We denote the determinant of a square matrix $\boldsymbol{A}$ by $|\boldsymbol{A}|$. For symmetric square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we write $\boldsymbol{A} \succeq \boldsymbol{B}$ (resp. $\boldsymbol{A} \succ \boldsymbol{B}$) if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite (resp. positive definite), and we denote the set of positive-semi definite matrices as $\mathbb{S}^{p\times p}_+$. The $p \times p$ identity matrix is $\boldsymbol{I}_p$. For a vector $x \in \mathbb{R}^p$, we use $\|x\|_2$ to denote the L^2 norm of x. The vector $x_S \in \mathbb{R}^{|S|}$ is the sub-vector of x with entries indexed by S. When appropriate, 0 and 1 (resp. $\mathbf{0}$ and $\mathbf{1}$) refer to a vector (resp. matrix) of zeros and ones.

We use the following statistics notation. We typically denote random vectors as unbolded capital letters, e.g., $X \in \mathbb{R}^p, Y \in \mathbb{R}^q$, and $X \sim (\mu, \boldsymbol{\Sigma})$ means that X has mean μ and covariance $\boldsymbol{\Sigma}$. Generally $\boldsymbol{\Sigma}_Y$ is $\mathrm{Cov}(Y)$, but we sometimes write $\boldsymbol{\Sigma}$ for $\mathrm{Cov}(X)$ specifically (when unambiguous). Samples drawn from a distribution are assumed independent and identically distributed (i.i.d). A model refers to a family of distributions.

We will assume that all random vectors $X \in \mathbb{R}^p$ we consider satisfy $E[\|X\|_2^2] < \infty$.

2.2 The equivalence of CSS and Principal Variables

In this section, we establish a mathematical equivalence and conceptual connection between CSS and Principal Variables. This equivalence is not mathematically difficult to establish, but it unlocks new perspectives that enable us to design novel and useful methodology for several problems in interpretable dimensionality reduction. Henceforth, selecting principal variables using $\boldsymbol{\Sigma}$ refers to the following subset search:

$$\operatorname*{argmin}_{S\subseteq[p]:|S|=k} \mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet S}\boldsymbol{\Sigma}_S^+\boldsymbol{\Sigma}_{S\bullet}). \tag{2.1}$$

The subset search (2.1) generalizes the original Principal Variables criterion (1.2) by allowing $\boldsymbol{\Sigma}$ to be singular. When $\boldsymbol{\Sigma}$ is non-singular, Cauchy's interlacing theorem [Bhatia, 1997] implies that $\boldsymbol{\Sigma}_S$ is always also non-singular, so $\boldsymbol{\Sigma}_S^+ = \boldsymbol{\Sigma}_S^{-1}$ and our new search problem (2.1) reduces to the original search problem (1.2).

Proposition 1 establishes an exact mathematical equivalence between CSS (1.1) and Principal Variables (2.1). It tells us explicitly how we can perform CSS by instead searching for principal variables, and vice-versa.

Proposition 1 (CSS and Principal Variables are equivalent)**.** *Consider a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ and define $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}^\top\boldsymbol{X}/n$. For any size-$k$ subset $S \subseteq [p]$, the CSS objective (1.1) with $\boldsymbol{X}$ and the Principal Variables objective (2.1) with $\hat{\boldsymbol{\Sigma}}$ are equal:*

$$\min_{\boldsymbol{B}\in\mathbb{R}^{k\times p}} \frac{1}{n}\|\boldsymbol{X} - \boldsymbol{X}_{\bullet S}\boldsymbol{B}\|_F^2 = \mathrm{Tr}(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\bullet S}\hat{\boldsymbol{\Sigma}}_S^+\hat{\boldsymbol{\Sigma}}_{S\bullet}).$$

One important consequence of Proposition 1 is that performing CSS on a centered data matrix $\boldsymbol{X}$ is identical to selecting principal variables using the sample covariance. Conceptually, this suggests viewing CSS as an estimation problem where the estimand is the population principal variable set (i.e., the principal variable set one would select using $\boldsymbol{\Sigma}$). This draws a clear parallel between CSS and PCA: both methods seek to estimate some function of the population covariance, and the typical unit-level data characterization of each method does so by plugging in the sample covariance as an estimate.

In what follows, we fully develop this point-estimation view of CSS. Consider the typical CSS use case where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a data matrix whose rows are centered i.i.d samples of a p-dimensional random vector $X \sim (\mu, \boldsymbol{\Sigma})$. In this setting, it is immediate that the CSS solution (1.1) is the plug-in estimate for the estimand

$$\underset{S \subseteq [p]: |S| = k}{\operatorname{argmin}} \min_{\boldsymbol{B} \in \mathbb{R}^{p \times k}} E[\|(X - \mu) - \boldsymbol{B}(X_S - \mu_S)\|_2^2]. \tag{2.2}$$

Proposition 2 confirms that the population CSS subset search (2.2) is equivalent to selecting principal variables using $\boldsymbol{\Sigma}$. Furthermore, it demonstrates that, even when X does not follow a non-degenerate Gaussian distribution, the subset selected by Principal Variables is characterized by an interpretable and desirable optimality criterion: optimal linear reconstruction of the remaining variables (after centering).

Proposition 2 (The CSS estimand)**.** *Consider a p-dimensional random vector $X \sim (\mu, \boldsymbol{\Sigma})$. For any size-$k$ subset $S \subseteq [p]$, the population CSS problem (2.2) has the same objective as the Principal Variables objective with $\boldsymbol{\Sigma}$:*

$$\min_{\boldsymbol{B} \in \mathbb{R}^{p \times k}} E[\|(X - \mu) - \boldsymbol{B}(X_S - \mu_S)\|_2^2] = \operatorname{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet S} \boldsymbol{\Sigma}_S^{+} \boldsymbol{\Sigma}_{S \bullet}).$$

Characterizing CSS in terms of covariances and reframing it as an estimation problem yields a number of benefits and insights:

1. A more general class of estimators for the population CSS solution (2.2) results from solving (2.1) with any estimate $\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$, not just sample covariance. Henceforth we will refer to solving (2.1) with $\hat{\boldsymbol{\Sigma}}$ as performing CSS with the covariance $\hat{\boldsymbol{\Sigma}}$. proba

2. It is immediately clear that practitioners can perform CSS in settings where they cannot access unit-level data but can access covariance or correlation estimates. For example, only the correlation matrix was available for four of the nine real data examples in Cadima et al. [2004], and in genome wide association studies practitioners often work with correlation estimates in place of inaccessible unit-level data [He et al., 2022, Mak et al., 2017].

3. Characterizing CSS as an estimation problem stresses the importance of selecting a subset that generalizes well. This is in line with recent work from Ordozgoiti et al. [2019] that emphasizes CSS's out-of-sample performance and examines the benefits of performing CSS with ℓ_2 regularization. Our equivalence suggests performing CSS with regularized covariance estimates as an alternative regularization scheme.

4. Similarly, a practitioner may be interested in the quality of their regression coefficient estimates from the regression of the remaining variables on the selected subset (i.e., the minimizing $\boldsymbol{B}$ in (1.1)). Using our covariance view, we show in Appendix A.5.1 that the error of these coefficient estimates is well-behaved, even when the practitioner selects a subset via a complex CSS algorithm.

5. The NP-completeness of CSS implies the NP-completeness of solving (2.1). Equivalently, if $X \sim N(0, \mathbf{\Sigma})$, $\mathbf{\Sigma} \succeq 0$, finding a size k subset $S \subset [p]$ that minimizes $\mathrm{Tr}(\mathrm{Cov}(X|X_S))$ is NP-complete.

6. If $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ has rank $r \ll \min(n, p)$, we can find $\boldsymbol{X}_r \in \mathbb{R}^{r \times p}$ such that $\boldsymbol{X}_r^\top \boldsymbol{X}_r = \boldsymbol{X}^\top \boldsymbol{X}$ via the singular value decomposition (SVD) of $\boldsymbol{X}$. Because $\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{X}_r^\top \boldsymbol{X}_r$, Proposition 1 tells us that the CSS objective value for any subset S is the same for $\boldsymbol{X}$ and $\boldsymbol{X}_r$, and thus performing CSS on $\boldsymbol{X}_r$ is equivalent to performing CSS on $\boldsymbol{X}$. The time complexity of finding $\boldsymbol{X}_r$ is $O(npr)$, and computing and then performing CSS on $\boldsymbol{X}_r$ may be much faster than performing CSS directly on $\boldsymbol{X}$.

2.3 Scalable CSS without unit-level data

Our first application is scalable CSS without unit-level data. We provide novel algorithms that allow us to efficiently perform CSS when we can only access covariance estimates and a real-data example that illustrates these algorithms' usefulness. These algorithms are made possible by the equivalence drawn in Section 2.2. We also discuss some other reasonable approaches that perform surprisingly poorly on our real data example.

2.3.1 Algorithms

As it stands, it is unclear how to efficiently perform CSS with just a covariance $\mathbf{\Sigma}$. The Principal Variables literature suggests some sophisticated approximation algorithms, but it lacks efficient implementations of them. The CSS literature provides efficient approximation algorithms, but it is unclear how to apply them when we only have covariance estimates. Proposition 1 implies that we could find $\boldsymbol{X}$ such that $\mathbf{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}/n$ and apply existing CSS algorithms to $\boldsymbol{X}$. Unfortunately, factoring $\mathbf{\Sigma}$ in this way has larger time complexity than some CSS algorithms themselves.

Hence, we derive versions of the greedy algorithm from Farahat et al. [2015] and the swapping algorithm from Ordozgoiti et al. [2018] that take a covariance $\mathbf{\Sigma}$ as input. These two algorithms are the empirically best-performing and fastest algorithms in the CSS literature, and they can be easily modified to perform subset search according to other objectives as well, which will be crucial for our application later in Section 3.3. Working directly with $\mathbf{\Sigma}$ allows for a significantly simpler presentation, derivation, and modification of these algorithms, and our new algorithms maintain the same time complexity as the originals when $n \geq p$. For the interested reader, we discuss how to modify our algorithms to perform subset search according to the other three Principal Variables criteria in Appendix A.3. We also provide some stylized settings where we can guarantee that our algorithms find the optimal subset in Appendix A.5.2.

The key insight in deriving our algorithms is writing the CSS objective in terms of residual covariance matrices. For two mean-zero random vectors $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ we define the residual from the regression of Y on X as

$$R(Y, X) = Y - \boldsymbol{B}^* X \qquad \boldsymbol{B}^* \in \operatorname*{argmin}_{\boldsymbol{B} \in \mathbb{R}^{q \times p}} E[\|Y - \boldsymbol{B}X\|_2^2].$$

The discussion in Appendix A.6 establishes that $R(Y, X)$ is well defined.

Lemma 1 summarizes what we need to know about residual covariances. In its statement, we adopt the convention that $0 \cdot \infty = 0$.

Lemma 1 (Efficient residual covariance update)**.** *Consider a random vector* $X \sim (0, \mathbf{\Sigma})$ *and a subset* $U \subset [p]$ *that does not contain the* i*th variable. Then the residual covariance for the subset* U *is given by* $\mathbf{\Sigma}_{R(X,X_U)} = \mathbf{\Sigma} - \mathbf{\Sigma}_{\bullet U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U\bullet}$*, and the residual covariance for the subset* $U \cup \{i\}$ *is given by*

$$\mathbf{\Sigma}_{R(X,X_{U\cup\{i\}})} = \mathbf{\Sigma}_{R(X,X_U)} - \frac{\beta\beta^\top}{\beta_i} \cdot I_{\beta_i>0}$$

where $\beta = (\mathbf{\Sigma}_{R(X,X_U)})_{\bullet i} = \mathbf{\Sigma}_{\bullet i} - \mathbf{\Sigma}_{\bullet U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,i}$*. That is, we can update* $\mathbf{\Sigma}_{R(X,X_U)}$ *when adding variables to* U *in* $O(p^2)$ *time, and removing variables from* U *in* $O(p^2)$ *time if* $\mathbf{\Sigma}_U^+$ *is known.*

Lemma 1 illustrates that CSS amounts to finding a subset S that minimizes $\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_S)})$, and it also allows us to quickly update $\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_S)})$ when we add or remove variables from S. Define the function

$$f(i, \boldsymbol{A}) = -\|\boldsymbol{A}_{\bullet i}\|_2^2/\boldsymbol{A}_{ii} \cdot I_{\boldsymbol{A}_{ii}>0}. \tag{2.3}$$

Considering some $U \subset [p]$, computations in Appendix A.7.17 show that any $i \notin U$ which minimizes $f(i, \mathbf{\Sigma}_{R(X,X_U)})$ also minimizes $\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_{U\cup\{i\}})})$. This is sufficient for establishing the correctness of the greedy and swapping algorithms we present below.

Greedy subset selection: Our greedy algorithm (Algorithm 1) starts with the empty subset $S^{(0)} = \{\}$ and, on iteration t, finds $S^{(t)}$ by adding the variable to $S^{(t-1)}$ that minimizes $\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_{S^{(t)}})})$. It terminates after k iterations and returns $S^{(k)}$. Alternatively, we can allow it to iterate until $\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_{S^{(t)}})})$ is sufficiently small. Algorithm 1's time complexity is $O(p^2k)$. A nice property of greedy subset search is that greedily selected subsets of different sizes are nested: upon finding the greedily selected size-k subset, we have necessarily found the greedily selected subsets of size $k-1, \ldots, 1$ in the process.

Algorithm 1 Greedy Subset Selection

Input: covariance matrix $\mathbf{\Sigma}$, number of columns to select k

procedure GREEDY_SUBSET_SELECTION($\mathbf{\Sigma}, k$)
  $S^{(0)} \leftarrow \{\}$
  $\mathbf{\Sigma}_{R(X,X_{S^{(0)}})} \leftarrow \mathbf{\Sigma}$
  for $t = 1, \ldots, k$ **do**
    $i^* \leftarrow \mathrm{argmin}_{i \notin S^{(t-1)}} f(i, \mathbf{\Sigma}_{R(X,X_{S^{(t-1)}})})$ ▷ f is from (2.3), arbitrarily break ties
    $S^{(t)} \leftarrow S^{(t-1)} \cup \{i^*\}$
    Compute $\mathbf{\Sigma}_{R(X,X_{S^{(t)}})}$ from $\mathbf{\Sigma}_{R(X,X_{S^{(t-1)}})}$ using Lemma 1
  return: $S^{(k)}$

Subset selection via swapping: On each iteration of our swapping algorithm (Algorithm 2) we loop through our currently selected subset S and swap each selected variable out for the variable that most reduces $\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_S)})$. If no swap reduces $\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_S)})$ for a particular selected variable, then we keep that variable in S. This is reminiscent of coordinate descent, a procedure that searches for a minimum of $f : \mathbb{R}^q \to \mathbb{R}$ by iteratively optimizing each coordinate while holding the others fixed. Algorithm 2 terminates after an iteration where no swaps are made, at which point the algorithm has converged to a kind of local

optimum. Since each swap strictly reduces the objective value, Algorithm 2 must converge in finite time. Its time complexity is $O(p^2k)$ per iteration. In practice, we run the algorithm multiple times with different random initializations and take the best resulting subset.

Algorithm 2 Subset Selection by Swapping

Input: covariance matrix $\mathbf{\Sigma}$, number of columns to select k, initial selected subset $S \in [p]^k$

procedure SWAPPING_SUBSET_SELECTION($\mathbf{\Sigma}, k, S$)
  $S_{copy} \leftarrow 0 \in \mathbb{R}^k$
  $\mathbf{\Sigma}_S^+ \leftarrow$ compute pseudo-inverse of $\mathbf{\Sigma}_S$
  $\mathbf{\Sigma}_{R(X,X_S)} \leftarrow \mathbf{\Sigma} - \mathbf{\Sigma}_{\bullet S}\mathbf{\Sigma}_S^+\mathbf{\Sigma}_{S\bullet}$
  while $S \neq S_{copy}$ **do**
    $S_{copy} \leftarrow S$
    for $j = 1, \ldots, k$ **do**
      $U \leftarrow S_{-j}$ ▷ $S_{-j} \in \mathbb{R}^{k-1}$ is S with S_j removed
      $\mathbf{\Sigma}_U^+ \leftarrow$ compute $\mathbf{\Sigma}_U^+$ from $\mathbf{\Sigma}_S^+$ using [Petersen and Pedersen, 2008, 3.2.7] ▷ Takes $O(k^2)$ time
      $\mathbf{\Sigma}_{R(X,X_U)} \leftarrow$ compute $\mathbf{\Sigma}_{R(X,X_U)}$ from $\mathbf{\Sigma}_{R(X,X_S)}$ using Lemma 1
      if $S_j \notin \operatorname{argmin}_{i \notin U} f(i, \mathbf{\Sigma}_{R(X,X_U)})$ **then** ▷ f is from (2.3)
        $i^* \leftarrow \operatorname{argmin}_{i \notin U} f(i, \mathbf{\Sigma}_{R(X,X_U)})$ ▷ Arbitrarily break ties
        $S_j \leftarrow i^*$ ▷ This alters S in place
        $\mathbf{\Sigma}_S^+ \leftarrow$ compute $\mathbf{\Sigma}_S^+$ from $\mathbf{\Sigma}_U^+$ using [Petersen and Pedersen, 2008, 3.2.7] ▷ Takes $O(k^2)$ time
        $\mathbf{\Sigma}_{R(X,X_S)} \leftarrow$ compute $\mathbf{\Sigma}_{R(X,X_S)}$ from $\mathbf{\Sigma}_{R(X,X_U)}$ using Lemma 1
  return: S

2.3.2 An example from BlackRock

We provide a real world application where BlackRock, an industry affiliate of ours, seeks to perform CSS and only has access to a covariance estimate. BlackRock is currently the largest asset manager in the world. For proprietary reasons, the data has been anonymized.

BlackRock has a collection of $p = 774$ thematic investment portfolios and would like to find a small subset of them that are representative of the remaining. Specifically, they are interested in performing CSS. Thematic portfolios invest in assets that are tied together by an interpretable common thread, e.g., renewable energy companies or start-ups in India. It is important to select a subset of portfolios (rather than find linear combinations of them) so that this thematic structure is maintained.

BlackRock does not have unit-level returns for each portfolio, but they do have a covariance estimate. After standardizing the returns of each portfolio to have unit variance, we perform CSS using our greedy and swapping algorithms for $k = 1, \ldots, 30$. For our swapping algorithm, we try 25 random initializations. The results are presented in Figure 2.1, where we plot the (estimated) average R^2 from the regression of each individual portfolio on the selected subset. For the size six subset found by our swapping algorithm, this average R^2 is above 90%. The swapping algorithm consistently outperforms the greedy one, and for smaller subset sizes it is able to get the same performance with one less variable. For sake of comparison, we also plot the average R^2 from the regression of each portfolio on the top k principal components. The principal components maximize this R^2 among all possible linear combinations of the portfolios. We find

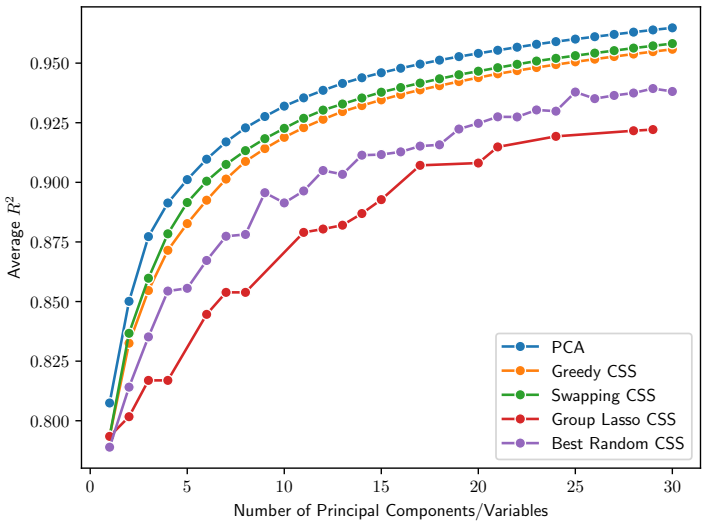


Figure 2.1: For the BlackRock example discussed in Section 2.3.2, the average R^2 from the regression of each variable on the selected subset (resp. principal components) versus increasing subset size (resp. number of principal components).

that the performance gap between PCA and CSS is well worth the trade-off in interpretability. For smaller subset sizes, CSS achieves the same performance as PCA while using a subset size that is only one larger than the number of principal components.

It is important to use efficient algorithms for problems of this scale. The Principal Variables literature, which mainly considers $p < 100$, suggests recomputing the objective from scratch for each new considered subset [Cadima et al., 2004]. For the BlackRock data, using Algorithm 1 to greedily select a size-30 subset takes less than a tenth of a second, while this naive approach takes more than six minutes. For subset selection via swapping, Algorithm 2 typically runs in less than half a second for one initialization, while this naive approach often takes over an hour (depending on the number of iterations until convergence). We remark that although the swapping algorithm finds better subsets, the greedy algorithm has comparable performance and is notably faster, especially considering it only needs to be run once for the largest k of interest.

2.3.3 Other approaches

Another approach to finding approximate CSS solutions is formulating a convex relaxation of the original objective [Balzano et al., 2010, Masaeli et al., 2010, Wang and Singh, 2017]. The group lasso [Yuan and Lin, 2006] provides one such natural convex relaxation:

$$\underset{\boldsymbol{B}\in\mathbb{R}^{p\times p}}{\operatorname{argmin}} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\|_F^2 + \lambda \sum_{i=1}^{p} \|\boldsymbol{B}_{i\bullet}^T\|_2 \tag{2.4}$$

For large values of λ, the solution to (2.4) has many zeroed-out rows, i.e., many variables are not used in the reconstruction. This suggests the following procedure for CSS: tune λ in (2.4) to yield exactly k non-zero

rows and select the variables corresponding to these rows. Another approach from Masaeli et al. [2010], Convex Principal Feature Selection (CPFS), solves (2.4) but with an ℓ_∞ penalty on the rows of $\boldsymbol{B}$ instead of a ℓ_2 one.

We can characterize (2.4) solely in terms of covariances and therefore also apply this group lasso approach to BlackRock's data. Defining $\hat{\boldsymbol{\Sigma}} = \boldsymbol{X}^T\boldsymbol{X}/n$, (2.4) is equivalent to

$$\underset{\boldsymbol{B}\in\mathbb{R}^{p\times p}}{\operatorname{argmin}} \operatorname{Tr}(\hat{\boldsymbol{\Sigma}} - 2\boldsymbol{B}^T\hat{\boldsymbol{\Sigma}} + \boldsymbol{B}^T\hat{\boldsymbol{\Sigma}}\boldsymbol{B}) + \frac{\lambda}{n}\sum_{i=1}^{p}\|\boldsymbol{B}_{i\bullet}^T\|_2$$

Surprisingly, on BlackRock's data, this group lasso approach performs worse than selecting subsets at random! Figure 2.1 displays the group lasso's performance as well as the best performance from 25 randomly selected subsets. The subset chosen by the group lasso performs notably worse than the best random subset. We were unable to tractably apply CPFS to the whole dataset, so we tried it on a random sub-sample of one hundred portfolios. On this sub-sample, CPFS did even worse than the group lasso (results not shown). Empirically, we found that the group lasso and CPFS systematically pulled in variables that were highly correlated with those that had already been selected. Our results are another example of convex variable selection methods performing poorly in settings with high correlation [Freijeiro-González et al., 2022].

2.4 CSS with missing and/or censored Data

The equivalence drawn in Section 2.2 suggests a natural workflow for performing CSS when some data is missing and/or censored: model the missingness and/or censorship, produce a covariance estimate that is reasonable under this model, and perform CSS using this covariance estimate.

As an illustrative example of this workflow, we use it to suggest a procedure for performing CSS with missing-at-random data. Suppose we observe i.i.d samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ and some values are missing-at-random. Let the set $\mathcal{I}_r$ be the indices of samples for which the rth variable is not missing, and $\mathcal{I}_{rs}$ be the indices of samples for which the r and sth variables are not missing. To get a reasonable covariance estimate $\hat{\boldsymbol{\Sigma}}$, we estimate the pairwise covariances using the not-missing values and then project onto the positive semi-definite cone:

$$\hat{\boldsymbol{\Sigma}} = \underset{\boldsymbol{M}\in\mathbb{S}_+^{p\times p}}{\operatorname{argmin}} \|\boldsymbol{M} - \hat{\boldsymbol{\Psi}}\|_F^2 \qquad \hat{\boldsymbol{\Psi}}_{rs} = \frac{1}{|\mathcal{I}_{rs}|}\sum_{i\in\mathcal{I}_{rs}}\left(x_r^{(i)} - \frac{1}{|\mathcal{I}_r|}\sum_{j\in\mathcal{I}_r}x_r^{(j)}\right)\left(x_s^{(i)} - \frac{1}{|\mathcal{I}_s|}\sum_{j\in\mathcal{I}_s}x_s^{(j)}\right)$$

Projecting a symmetric matrix onto the positive semi-definite cone amounts to setting its negative eigenvalues to zero. Finally, we apply the algorithms from Section 2.3.1 to $\hat{\boldsymbol{\Sigma}}$.

Using both simulated and real data, we demonstrate that our procedure outperforms the two existing methods for CSS with missing data. Initial work on this problem by Balzano et al. [2010] uses a greedy Block Orthogonal Matching Pursuit (BOMP) algorithm to approximately solve a CSS-like objective that ignores the missing values. In follow-up work, Wang and Singh [2017] mask the missing values and solve a group lasso penalized problem similar to (2.4).

Briefly, we remark that one could put additional parametric assumptions on the generative model we will introduce in Chapter 3 and then apply the expectation-maximization (EM) algorithm to handle missingness. We found this to be too sensitive to the parametric assumptions and the EM algorithm's initialization to

Method	% Corr. Subset Selected	# Corr. Variables Selected	CSS Obj.
Our Method	**1.000**	**4.000 ± 0.000**	**2.400 ± 0.000**
BOMP	0.456	2.191 ± 0.056	3.223 ± 0.025
Group Lasso	0.000	1.713 ± 0.016	5.806 ± 0.037
Random	0.001	0.804 ± 0.023	6.065 ± 0.045

Table 2.1: The proportion of times the correct subset was selected (% Corr. Subset Selected), the average size of the intersection of the selected subset and correct subset (# Corr. Variables Selected), and the population CSS objective (CSS Obj.) for the thousand simulated trials described in Section 2.4.1. The ± reports one standard error. Bold indicates best performance.

work well in practice.

2.4.1 A simple simulation study

We run a simulation study to evaluate the quality of each method. In each trial, we draw $n = 200$ samples from a distribution that has $p = 20$ unit-variance variables. We initially sample a set of $k = 4$ variables, and the remaining variables are linear combinations of these four plus a small amount of Gaussian noise (this is a preview of Chapter 3's generative model). The simulation setting is constructed so that this ground truth set of $k = 4$ variables solves the population CSS subset search (2.2), and therefore can be treated as the gold standard subset. Prior to running any search algorithms, we randomly omit each observed value with probability 0.05. The distribution we sample from is described in full detail in Appendix A.1.1.

For each trial, we select a size-four subset using each method. We also select a random size-four subset as a benchmark. For our method, we use our swapping algorithm and try ten random initializations. For the group lasso, we find the regularization strength that selects a size-four subset via a binary search. We run one thousand trials and use three metrics to measure the quality of each method: (1) whether or not the selected subset $\hat{S}$ is exactly equal to the ground truth subset S; (2) the size of the intersection of $\hat{S}$ and S; and (3) the selected subset's population CSS objective $\mathrm{Tr}(\mathbf{\Sigma} - \mathbf{\Sigma}_{\bullet\hat{S}}\mathbf{\Sigma}_{\hat{S}}^{-1}\mathbf{\Sigma}_{\hat{S}\bullet})$, where $\mathbf{\Sigma}$ is the population covariance. Recovery of the ground truth subset S is indeed a reasonable performance metric as, in our example, S achieves the lowest population CSS objective among all size-4 subsets.

Table 2.1 documents the results from our simulation. Our method selects the correct subset every time while BOMP selects it less than half the time and group lasso fails to select it even once. As expected, our method also achieves the lowest average CSS objective value by a significant amount.

2.4.2 An example with ozone level detection data

We consider the Ozone Level Detection Dataset from the UCI machine learning repository [Dua and Graff, 2017]. The dataset has $n = 2535$ samples of $p = 73$ variables. So that we can compute a ground truth CSS objective value for any selected subset, we drop samples with already missing values and are left with a sample size of $n = 1847$.

One hundred times, we randomly omit observed values with probability $q = 0.05, 0.1, 0.2$ and select a size $k = 5, 10, 20$ subset using our method and BOMP. For our method, we use our swapping algorithm and only try one random initialization. We neglect the group lasso approach because it is too slow for a problem of this scale. Given its noted bad performances, we expect it would have done poorly. To measure a selected subset's quality, we compute the CSS objective value it would have attained on the fully observed dataset.

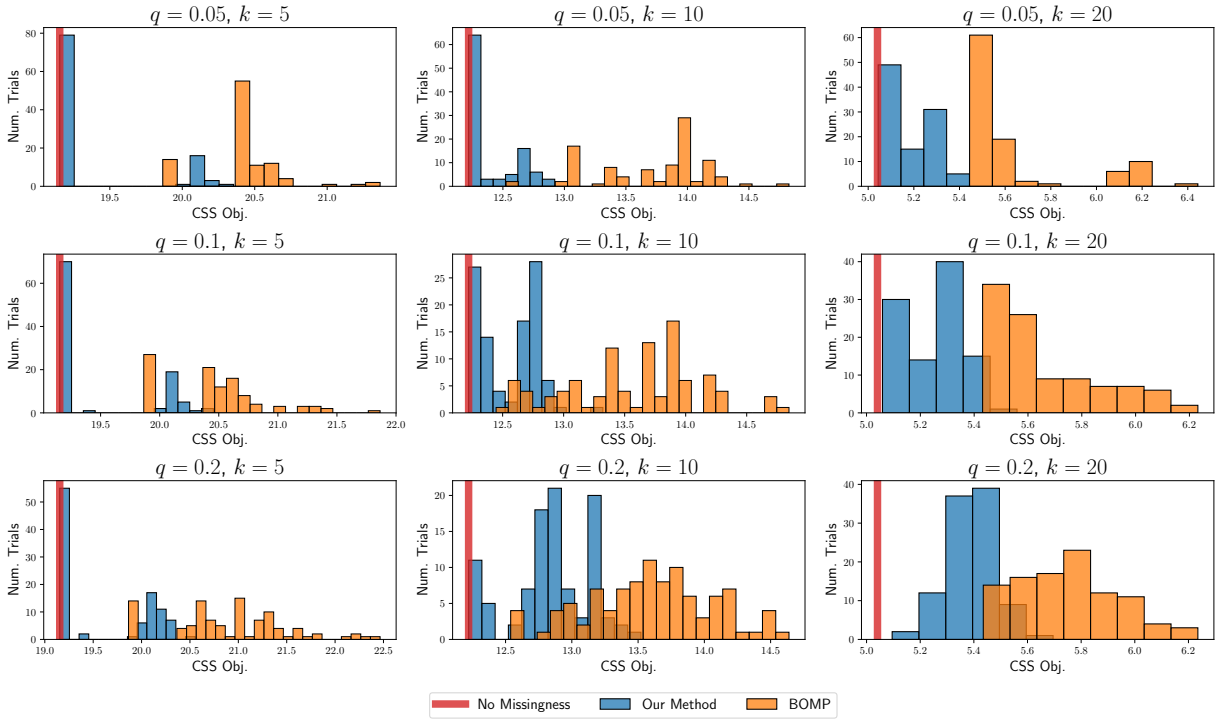


Figure 2.2: For different amounts of missingness q and subset sizes k, the selected subset's CSS objective value on the fully observed Ozone Level Detection data over the hundred trials described in Section 2.4.2. The red line marks the CSS objective value attained by the subset that was selected using the fully observed data.

As a benchmark for how well we can expect to do with no missing data, we run our swapping algorithm on the fully observed dataset with one hundred different random initializations and record the best result.

We provide the results from our experiment in Figure 2.2. Our method significantly outperforms BOMP in every setting and performs similarly to our no-missing-data benchmark when k and q are not too large. Also, our method's performance appears to be less variable than BOMP's.

Chapter 3

A generative view of column subset selection

In this chapter, we show that CSS and Principal Variables can both be viewed as maximum likelihood estimation within a certain semi-parametric model. We call this viewpoint *Probabilistic Column Subset Selection (PCSS)*. As the name suggests, PCSS provides a generative view of CSS the same way that PPCA does for PCA.

Aside from offering a deeper conceptual understanding of CSS and providing a theoretical statistical justification for the CSS problem, the generative model underlies an important application in psychology and psychometrics.

Selecting the subset size for CSS: Supposing we observe data drawn from a generalization of the PCSS model, we propose a novel and theoretically motivated procedure for selecting an appropriate subset size. We first verify that our procedure performs well in some difficult simulated settings, and we then demonstrate how it can be used to shorten surveys in an automated way. As an example, we apply our procedure to the Big Five Inventory (BFI), a 44-item personality questionnaire [John et al., 1991]. In less than a tenth of a second, it suggests cutting the survey by more than half. Importantly, the questions our procedure selects appear to sufficiently capture the relevant information in the survey.

3.1 CSS as maximum likelihood estimation

In what follows, we show that CSS and Principal Variables can both be viewed as maximum likelihood estimation within a certain semi-parametric model. We call this viewpoint *Probabilistic Column Subset Selection (PCSS)*. As the name suggests, PCSS provides a generative view of CSS the same way that PPCA does for PCA.

We introduce the k-dimensional *PCSS model*. It posits that each data vector $X \in \mathbb{R}^p$ arises from the

following two-step sampling procedure:

$$\begin{aligned} X_S &\sim F \\ X_{-S} \mid X_S &\sim N(\mu_{-S} + \boldsymbol{W}(X_S - E_F[X_S]), \sigma^2 \boldsymbol{I}_{p-k}). \end{aligned} \tag{3.1}$$

In words, we first sample the k principal variables X_S from some distribution F, and then generate the remaining variables X_{-S} as linear combinations of these principal variables plus spherical Gaussian noise. The parameters of this model are the set of principal variables S, their distribution F, the remaining variables' mean μ_{-S}, the regression coefficients $\boldsymbol{W}$, and the residual variance $\sigma^2 > 0$. The size of the principal variable set, k, is fixed and not a parameter in the model.

The relationship between PCSS and CSS is explained by Theorem 1. It says that, in our PCSS model, any CSS solution is a maximum likelihood estimate (MLE) for S over a non-parametric family of principal variable distributions F.

Theorem 1 (CSS solution is an MLE)**.** *Let $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ be samples with sample covariance $\hat{\boldsymbol{\Sigma}}$. For any $M < \infty$, let $\mathcal{P}_M$ be the set of distributions in the k-dimensional PCSS model (3.1) where F admits a probability density bounded by M. Then, so long as no k variables achieve perfect in-sample linear reconstruction of the remaining variables, any CSS solution*

$$\hat{S} \in \operatorname*{argmin}_{U \subset [p], |U| = k} \operatorname{Tr}(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\bullet U} \hat{\boldsymbol{\Sigma}}_U^{+} \hat{\boldsymbol{\Sigma}}_{U\bullet})$$

is a maximum likelihood estimator for S in the model $\mathcal{P}_M$. Moreover, each $\hat{S}$ is a maximum hybrid likelihood estimator (see [Owen, 2001, Equation 9.1]) for S in the original model (3.1) where F is completely unrestricted.

The density bound $M < \infty$ is only in place to ensure that the likelihood is finite. Otherwise every subset S results in infinite likelihood. Likewise, if some size-k subset of variables achieves perfect in-sample linear reconstruction of the remaining (and thus attains the minimum CSS objective value of zero), then the likelihood for this subset tends to ∞ as we send $\sigma^2 \to 0$ and technically no MLE exists.

Theorem 1 tells us that any CSS solution is a MLE under the model (3.1), but since the model (3.1) is not fully parametric, it is not immediately clear that the MLE is a favorable estimate. The usual desirable properties of the MLE, e.g., consistency, are not automatically guaranteed.

Theorem 2 alleviates these concerns by establishing that, in the high-dimensional regime where tools like CSS are typically applied, the MLE subset $\hat{S}$ from Theorem 1 is consistent for the population subset S. Further, it tells us that this population subset minimizes the population CSS objective (2.1).

Theorem 2 (High-dimensional consistency of CSS)**.** *Suppose we observe n samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ from a distribution P in the k-dimensional PCSS model, and consider the proportional asymptotic regime where p/n converges to a constant as $n, p \to \infty$ while k remains fixed. If, as $n, p \to \infty$, the data generating distribution satisfies*

1. *(Light tails) The sub-Gaussian norm (see [Vershynin, 2018, Definition 2.5.6]) of each variable is bounded above by a constant independent of p and n,*
2. *(Imperfect reconstruction) The minimum eigenvalue of the principal variables' covariance $\operatorname{Cov}_P(X_S)$ and the residual variance σ^2 are both bounded below by a positive constant independent of p and n,*

3. *(Important principal variables) For some $\delta > 0$, at least $\Omega(p^{1/2+\delta})$ entries in each column of the regression coefficients $\boldsymbol{W}$ are bounded away from zero by some constant independent of p and n,*

then the MLE subset $\hat{S}$ from Theorem 1 is consistent for the population subset S (i.e., $\lim_{n\to\infty} P(\hat{S} = S) = 1$), which eventually minimizes the population CSS objective (2.1).

A closer examination of Theorem 2's proof actually tells us that, under the three conditions, the probability that $\hat{S}$ does not equal S shrinks to zero at a super-polynomial rate. The first two conditions of the theorem are standard regularity conditions. They ensure that the in-sample CSS objective is well behaved for all size-k subsets. The third condition is a strong identifiability condition for our problem. It ensures that, compared to other size-k subsets U, the population subset S results in lower reconstruction error for sufficiently many (but still a vanishing proportion) of the variables in $-S \setminus U$. Without some such condition, the population subset S may not even be unique. For the interested reader, Appendix A.4 identifies a more general high-dimensional setting where the CSS estimate (1.1) is consistent for population CSS solution, but at the expense of giving a less interpretable statement.

3.2 The subset factor model

By generalizing the probabilistic model (3.1) in which the CSS solution is the MLE, we can motivate subset versions of other existing probabilistic methods.

In this vein, we introduce the k-dimensional *subset factor model*, a generalization of our k-dimensional PCSS model that drops all distributional assumptions and allows the covariance of $X_{-S} \mid X_S$ to be a diagonal matrix $\boldsymbol{D} \succ \boldsymbol{0}$:

$$\begin{aligned} &X_S \sim F \qquad \epsilon \sim (0, \boldsymbol{D}) \qquad \epsilon_1 \perp, \ldots, \perp \epsilon_{p-k} \qquad \mathrm{Cov}(X_S, \epsilon) = \boldsymbol{0} \\ &X_{-S} = \boldsymbol{W}(X_S - E_F[X_S]) + \mu_{-S} + \epsilon. \end{aligned} \tag{3.2}$$

Our k-dimensional subset factor model has a close relationship with the k-dimensional factor model [Mardia et al., 1979, Section 9.2.1]

$$\begin{aligned} &Z \sim (0, \boldsymbol{I}_k) \qquad \epsilon \sim (0, \boldsymbol{D}) \qquad \epsilon_1 \perp, \ldots, \perp \epsilon_p \qquad \mathrm{Cov}(Z, \epsilon) = \boldsymbol{0} \\ &X = \boldsymbol{W}Z + \mu + \epsilon, \end{aligned} \tag{3.3}$$

where again $\boldsymbol{D} \succ \boldsymbol{0}$. In the k-dimensional factor model, each variable X_i is a linear combination of latent common factors $Z \sim (0, \boldsymbol{I}_k)$ plus a mean-zero unique factor ϵ and a non-random shift μ_i. In our k-dimensional subset factor model, an unknown subset of the observed variables act as common factors (once centered), and the remaining variables are similarly linear combinations of these common factors plus a unique factor and non-random shift. Moreover, by dropping distributional assumptions and allowing a previously spherical conditional covariance to be diagonal, our subset factor model generalizes our PCSS model the same way the factor model generalizes the model used for PPCA.

The subset factor model (3.2) arises naturally from a hybridization of PCA and Factor Analysis that aims to maintain strengths from both methods. To see how, we must first discuss the advantages and disadvantages of PCA and Factor Analysis. Factor Analysis enjoys two major benefits over PCA. First, the factor model (3.3) provides an interpretable generative model for the data. In contrast, PCA has no

associated generative model. Second, in the factor model, the regression of $X \in \mathbb{R}^p$ on its lower dimensional representation $Z \in \mathbb{R}^k$ has uncorrelated residuals, meaning that Z accounts for all the common variation in X. If these residuals were correlated, i.e., they jointly varied about their means in some structured way, that would mean there may be additional meaningful trends in X that Z does not account for. This is the case in PCA, where the regression of X on its top k principal components typically yields correlated residuals. Due to the fact that the lower dimensional representation Z is latent and unobserved, however, Factor Analysis suffers from a slew of unidentifiability and indeterminancy issues [Anderson and Rubin, 1956, Guttman, 1955, Steiger, 1994]. Proponents of components-based methods are particularly critical of factor indeterminacy, the fact that the common factors are not uniquely determined, even up to rotation [Schönemann and Steiger, 1976, 1978, Steiger, 1979, Velicer and Jackson, 1990]. In contrast, components (linear combinations of the observed variables) are always observed and determined, and they therefore do not suffer from any of the same issues.

A natural attempt to hybridize PCA and Factor Analysis in a way that inherits strengths from both methods is to assume a factor model where the factors are themselves components. Then, via an interpretable generative model, we can recover determined factors that account for all the common variation in X. This idea has been of great historical interest, but it has repeatedly been emphasized that it is impossible to do so when all the unique factors ϵ_i are required to have positive variance [Guttman, 1955, Schönemann and Steiger, 1976, Wilson, 1928]. However, if we allow some unique factors to have zero variance, Theorem 3 tells us that, surprisingly, this is equivalent to assuming a subset factor model.

Theorem 3 (The subset factor model compromise)**.** *Suppose $X \sim (\mu, \mathbf{\Sigma})$, $\mathbf{\Sigma} \succ \mathbf{0}$ follows the k-dimensional factor model (3.3) where the unique factors ϵ_i are allowed to have zero variance. Then the common factors Z can be written as components (i.e., $Z = \boldsymbol{B}(X - \mu)$ for some $\boldsymbol{B} \in \mathbb{R}^{k \times p}$) if and only if X follows the k-dimensional subset factor model (3.2).*

3.3 Selecting the subset size

We will use the above generative models to determine the smallest subset size needed to capture some observed data's underlying structure. More broadly, the procedure we propose showcases how this chapter's generative view facilitates the application of statistical tools in subset selection problems.

Formally, we aim to find k^*, the smallest non-negative integer k for which the k-dimensional subset factor model (3.2) holds. Recall that the k-dimensional subset factor model consists of distributions satisfying (3.2) for some size-k factor set S:

$$\begin{aligned} &X_S \sim F \qquad \epsilon \sim (0, \boldsymbol{D}) \qquad \epsilon_1 \perp, \ldots, \perp \epsilon_{p-k} \qquad \mathrm{Cov}(X_S, \epsilon) = \mathbf{0} \\ &X_{-S} = \boldsymbol{W}(X_S - E_F[X_S]) + \mu_{-S} + \epsilon. \end{aligned} \tag{3.2, revisited}$$

We focus on the subset factor model over the PCSS model (3.1) because it is more general and hence more representative of the structure we expect to see in real data. An interested reader can find the analogous discussion for the PCSS model in Appendix A.2.4. Prior to proceeding, we point out that subset factor models of increasing subset size are nested. This is because any distribution satisfying (3.2) for a set S also satisfies (3.2) for any superset of S.

Our approach is to output $\hat{k}$, the smallest k for which we cannot falsify that the data generating distribution belongs to the k-dimensional subset factor model. Formally, $\hat{k}$ is the smallest k for which we fail to reject the null $H_{0,k} : |S| \leq k$ that the data is drawn from a distribution satisfying (3.2) for some size-k set S.

To start, we present an idealized procedure that guarantees a notion of error control. This ideal procedure controls the probability that the suggested subset size is too large, i.e., that $\hat{k} > k^*$. Unfortunately, the procedure is computationally intractable, and we ultimately use it to motivate a tractable procedure that performs well in practice. A more technical and detailed discussion of this ideal procedure can be found in Appendix A.2.

For a fixed k, our idealized procedure rejects the null $H_{0,k} : |S| \leq k$ when the statistic T_k is large:

$$T_k = \min_{U \subset [p]:|U|=k} T(U), \qquad T(U) = n \log \left(\frac{|\mathrm{Diag}(\hat{\mathbf{\Sigma}}_{-U} - \hat{\mathbf{\Sigma}}_{-U,U} \hat{\mathbf{\Sigma}}_U^+ \hat{\mathbf{\Sigma}}_{U,-U})|}{|\hat{\mathbf{\Sigma}}_{-U} - \hat{\mathbf{\Sigma}}_{-U,U} \hat{\mathbf{\Sigma}}_U^+ \hat{\mathbf{\Sigma}}_{U,-U}|} \right) \tag{3.4}$$

The statistic T_k is the generalized likelihood ratio test statistic for testing the null that the data is drawn from a Gaussian distribution in the k-dimensional subset factor model (see Appendix A.2.1 for details). Although the test statistic T_k is designed for Gaussians, we will be able to make use of it even in a totally non-parametric setting. Intuitively, T_k is large when the residual covariance from the regression of X_{-U} on X_U is highly non-diagonal for every size-k subset U. If this is the case, the k-dimensional subset factor model cannot possibly hold.

In what follows, we determine an appropriate critical value for the aforementioned test. By definition, the null $H_{0,k} : |S| \leq k$ is true exactly when $k \geq k^*$. Bearing this in mind, fix some $k \geq k^*$. First, we imagine restricting the subset factor model so that the unique factors ϵ_j are Gaussian and independent of the principal variables X_S, which we assume have a density. In this simpler setting, whenever the data generating distribution satisfies (3.2) for some size-k set S, the statistic $T(S)$ has distribution exactly

$$n \sum_{j=2}^{p-k} \log \left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-k-j}} \right), \tag{3.5}$$

where $\{\chi^2_\ell\}, \{\tilde{\chi}^2_\ell\}$ are mutually independent chi-squared random variables with degrees of freedom specified by their subscript. Thus, rejecting $H_{0,k} : |S| \leq k$ when $T_k = \min_{U:|U|=k} T(U)$ is larger than the critical value

$$Q_{n,p,k}(1-\alpha) = \mathrm{Quantile} \left\{ 1-\alpha, n \sum_{j=2}^{p-k} \log \left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-k-j}} \right) \right\} \tag{3.6}$$

results in exact finite sample type I error control at level α. Surprisingly, even when we make no parametric assumptions, this same critical value guarantees asymptotic type I error control.

Theorem 4 establishes that, by using the above critical value, our ideal procedure does not suggest a subset size that is too large with high probability.

Theorem 4 (Error control). *Consider $n > p$ samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ from a distribution P and let k^* be the smallest non-negative integer k for which P belongs to the k-dimensional subset factor model (3.2). If $\hat{k}$ is the smallest k for which the test statistic T_k from (3.4) is not strictly larger than the critical value*

$Q_{n,p,k}(1-\alpha)$ defined in (3.6), then

$$\limsup_{n\to\infty} P(\hat{k} > k^*) \leq \alpha.$$

That is, our ideal procedure suggests a subset size that is too large with probability at most α in large samples. Furthermore, if we restrict the subset factor model (3.2) to have Gaussian unique factors ϵ_j that are independent of the principal variables X_S, then the above conclusion holds in finite samples, i.e., $P(\hat{k} > k^*) \leq \alpha$.

With the result of Theorem 4 in mind, we make two remarks about the output $\hat{k}$ of our ideal procedure. First, if the test statistic T_k typically exceeds the critical value $Q_{n,p,k}(1-\alpha)$ whenever $k < k^*$ (i.e., the test is powerful for the nulls $H_{0,k} : |S| \leq k$, $k < k^*$), our ideal procedure should return k^* with high probability. Second, whenever our ideal procedure returns $\hat{k} > k^*$, the outputted $\hat{k}$ should seldom be much larger than k^*. Since T_k minimizes $T(\cdot)$ over all size-k subsets, the proposed critical value becomes increasingly conservative when the data generating distribution satisfies (3.2) for many size-k subsets. As k exceeds k^*, the proportion of size-k subsets for which (3.2) is satisfied grows quickly to one. Also, for k up to $\lfloor (k^* + p)/2 \rfloor$, there exist subsets S for which (3.2) is satisfied with progressively larger symmetric differences, and their corresponding statistics $T(S)$ are accordingly less positively correlated. These facts combined suggest that the probability of rejecting $H_{0,k} : |S| \leq k$ shrinks quickly as k increases past k^*.

Since computing the test statistic T_k^* requires an exhaustive subset search, our idealized procedure is admittedly computationally intractable. Still, we can use it as motivation to design a tractable procedure that works well in practice.

Our practical approach is to find a size-k subset $\hat{S}$ that approximately minimizes $T(\cdot)$, and then to reject the null $H_{0,k} : |S| \leq k$ when $T(\hat{S})$ exceeds the same critical value $Q_{n,p,k}(1-\alpha)$ as before. We again output the smallest k for which we fail to reject. Our approach is analogous to how, when the likelihood is not convex, a statistician may use an approximately maximized likelihood to run a likelihood ratio test (e.g., when selecting the number of factors to use in a factor model [Beaujean, 2013]). Appendix A.3 describes how to modify Section 2.3.1's algorithms to search for this subset $\hat{S}$.

While our practical procedure does not guarantee error control, we still expect it to output $\hat{k}$ that will rarely be much larger than k^*. Again fix $k \geq k^*$. So long as our algorithms happen upon just one of the size-k subsets for which (3.2) is satisfied, we expect that $T(\hat{S})$ will be larger than $Q_{n,p,k}(1-\alpha)$ with probability at most α (and much less when k is larger than k^*). As mentioned before, the proportion of size-k subsets for which (3.2) is satisfied quickly approaches one as k increases, and it thus becomes correspondingly more likely that our approximation algorithms will happen upon such a subset.

Ultimately, if our procedure suggests selecting a size-$\hat{k}$ subset, we suggest selecting the size-$\hat{k}$ subset $\hat{S}$ that approximately minimizes $T(\cdot)$. This subset results in the most diagonal residual covariance we can find, and is thus the subset that best fits the data as per the subset factor model. Also, upon restricting the subset factor model to multivariate Gaussian distributions, $\hat{S}$ acts as an approximation for the MLE of S (see Appendix A.2.3 for details).

In what follows, we verify that our practical procedure suggests reasonable subset sizes in some difficult simulated settings where the number of samples n is not much larger than the number of variables p. Then, we demonstrate how it can be applied to dramatically reduce the length of a personality survey.

3.3.1 A simulated example

We consider a difficult simulated example that reflects the real data example to be presented in Section 3.3.2. For each trial, we draw $n = 200$ samples from a mean-zero distribution in the subset factor model (3.2) with $p = 50$ variables and a size $k = 20$ population factor set S. For context, there are over 47 trillion size-20 subsets of 50 variables. We consider two cases: one where the unique factors are Gaussian, and one where they are a collection of independent Rademacher, t-distributed, and centered exponential random variables. Within both cases we consider distributions with varying amounts of "signal", which we define to be the average population R^2 from the regression of the variables not in S on those in S. Appendix A.1.2 provides exact descriptions of the distributions we use, and Appendix A.1.3 documents additional results for other sample sizes n.

For each trial we apply our procedure at level $\alpha = 0.05$ and record (1) the size of the selected subset $\hat{S}$; (2) the size of the intersection of $\hat{S}$ and S; and (3) the population sum of square canonical correlations between the variables in $\hat{S}$ and S. The sum of square canonical correlations measures the degree to which the variables in S and $\hat{S}$ span similar subspaces (see Section 3.1 of Schneeweiss and Mathes [1995] for more detail), and it attains its maximum value of $k = 20$ when $S \subseteq \hat{S}$. To search for the subset that minimizes $T(\cdot)$ we use our swapping algorithm and try ten random initializations. Separately, we tried searching for the minimizing subset via a forward-backward algorithm. Although this approach is faster, the results were appreciably worse. These additional results as well as a description of our forward-backward algorithm can be found in Appendix A.1.4.

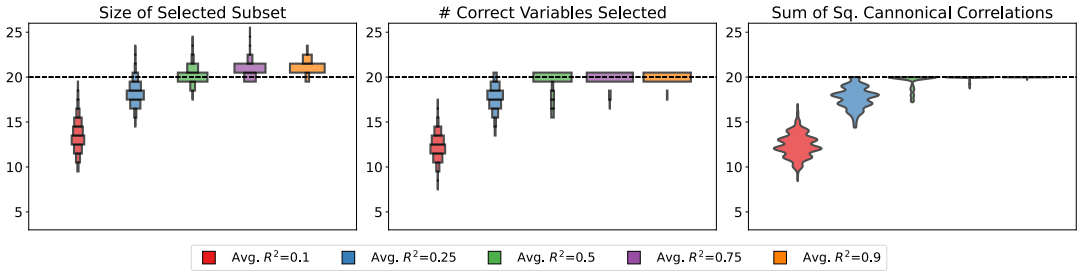
We provide the results from our simulations in Figure 3.1. As expected, our method under selects when the signal size is too small. In these cases, the covariance looks reasonably diagonal to begin with, so it is difficult to gauge whether more variables are needed to achieve a diagonal residual covariance. For reasonable to large signal sizes the procedure typically suggests selecting 20 or 21 variables, and it seldom suggests selecting more than 22. In these settings, the selected subset $\hat{S}$ is almost always a superset of the population subset S. The results when the unique factors are non-Gaussian versus Gaussian look nearly identical, suggesting that our procedure has reasonable finite sample performance even when the unique factors are highly non-Gaussian and n is not much larger than p.

3.3.2 Reducing survey length via our procedure

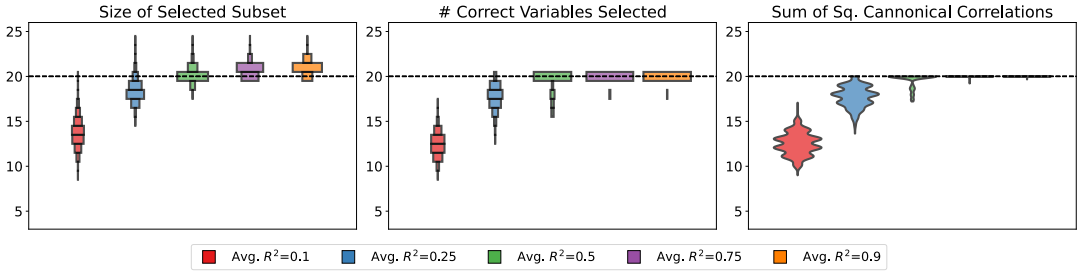
In recent years, psychologists have been increasingly interested in creating shorter versions of existing surveys. Kruyen et al. [2013] found that, among the 2273 articles that appeared in six peer-reviewed psychological journals from 2005 to 2010, 170 articles contained abbreviated surveys. Developing a novel shortened survey was a main contribution in 84 of these articles.

Despite the increasing demand for shortened surveys, there is no standardized way of shortening them. Psychologists default to a number of arbitrary heuristic approaches that require extensive human labor and are susceptible to personal biases. Furthermore, reliability, the one concrete and standard measure of a shortened survey's quality, is defined in the restrictive setting of classical test theory [Allen and Yen, 2001]. Estimates of reliability are highly sensitive to classical test theory's assumptions, and these assumptions are often violated in modern, multi-dimensional settings [McNeish, 2018].

We propose using our procedure to shorten surveys. Essentially, we suggest selecting a subset of questions that, once accounted for, render the information in the remaining questions independent. To demonstrate



(a) Gaussian unique factors



(b) Non-Gaussian unique factors

Figure 3.1: For one thousand trials, the size of the selected subset $\hat{S}$ (left), the size of the intersection of $\hat{S}$ and the population subset S (middle), and the population sum of squared canonical correlations between the variables in $\hat{S}$ and S (right) for the different simulation settings described in Section 3.3.1. Color indicates the average population R^2 from the regression of the variables not in S on those in S.

.

our procedure's effectiveness, we consider the Big Five Inventory (BFI) [John et al., 1991], a $p = 44$ question personality survey. There has been much interest in reducing the BFI survey, as evidenced by articles proposing a 10-question version [Rammstedt and John, 2007], a 15-question version [Gerlitz and Schupp, 2005], and multiple 20-question versions [Engvik and Clausen, 2011, Tucaković and Nedeljković, 2023, Veloso Gouveia et al., 2021]. Each question in the survey is attributed to one of the five personality factors [Digman, 1990]: extraversion, agreeableness, conscientiousness, neuroticism, and openness.

We consider the BFI dataset from Zhang et al. [2022], which has responses from $n = 228$ undergraduate students at a large US public university. It can be found on CRAN in the `EFAutilities` package. We apply our procedure at level $\alpha = 0.05$ and search for the minimizing subset using our swapping algorithm with just one random initialization (using more than one random initialization gives identical results). Within less than a tenth of a second, our procedure selects a size-19 subset of the original 44 questions, cutting the survey in more than half.

As is the case for all shortened surveys, we evaluate the quality of our survey by judging its validity and reliability [Kruyen et al., 2013]. We point readers who are unfamiliar with validity and reliability to the relevant introductory sections of Kruyen et al. [2013].

Validity: We evaluate the construct validity of our reduced survey, i.e., the degree to which it captures the relevant construct of interest. We cannot evaluate predictive validity (the degree to which our survey is predictive of an external criterion) because we are working with a stand-alone dataset. Examination of our selected questions suggest they indeed capture the relevant construct of interest. Table 3.1 provides a

Factor	# in Reduced Survey	# in Original Survey
Extraversion	4	8
Agreeableness	4	9
Conscientiousness	5	9
Neuroticism	3	8
Openness	3	10

Table 3.1: The number of questions attributed to each of the five personality factors for both the original and reduced surveys from Section 3.3.2.

breakdown of how many questions are attributed to each personality factor in both the original and the reduced survey. The selected questions are distributed fairly evenly across the five personality factors, and, in concordance with best practices [Costello and Osborne, 2005], our method selects at least three questions per factor. Furthermore, performing a 5-factor exploratory factor analysis reveals that the reduced survey maintains the original survey's factor structure, another standard indicator of construct validity [Kruyen et al., 2013]. Table 3.2 displays the loadings from this factor analysis.

Reliability: Reliability measures the consistency of a survey's responses across multiple re-administrations to the same individual. Although defining and estimating reliability relies heavily on the restrictive assumptions of classical test theory, it is still of much practical interest to pyschometricians. To measure reliability, we first split our reduced survey into five sub-surveys, each consisting of questions attributed to the same factor. Then we compute Cronbach's coefficient α [Cronbach, 1951] for each of these five sub-surveys. For the sake of comparison, we do the same for the three pre-existing 20-question versions of the BFI survey. The results are given in Figure 3.2. Along with being similar to the reliabilities of the three pre-existing surveys, our reliabilities are both (1) within the range of reliabilities one typically sees from shortened surveys [Kruyen et al., 2013, Figure 3] and (2) within the range of reliabilities contemporary pyschometricians consider adequate [Clark and Watson, 1995]. Admittedly, our method aims to find the most parsimonious subset, and parsimony can be at odds with reliability. If the practitioner is particularly concerned about reliability, they can always add additional questions.

	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Extraversion 1	0.11	**0.77**	0.10	-0.06	-0.03
Extraversion 2	-0.14	**0.54**	0.10	0.25	0.13
Extraversion 3	0.23	**-0.54**	0.07	0.19	-0.05
Extraversion 4	-0.02	**0.81**	-0.01	0.10	0.02
Agreeableness 1	0.03	0.17	-0.06	**0.47**	0.15
Agreeableness 2	0.14	-0.16	0.16	**-0.66**	0.02
Agreeableness 3	0.12	-0.08	0.21	**0.67**	0.07
Agreeableness 4	**0.35**	0.09	-0.06	**-0.52**	-0.07
Conscientiousness 1	0.10	0.11	-0.01	0.23	**0.48**
Conscientiousness 2	0.09	0.11	0.14	-0.01	**-0.60**
Conscientiousness 3	-0.03	-0.09	0.24	-0.03	**0.55**
Conscientiousness 4	0.06	0.11	-0.02	0.05	**0.66**
Conscientiousness 5	**0.50**	-0.03	0.01	0.12	**-0.36**
Neuroticism 1	**-0.57**	0.02	0.03	0.05	0.10
Neuroticism 2	**0.63**	-0.07	-0.15	0.07	-0.08
Neuroticism 3	**0.79**	0.03	-0.02	-0.13	0.09
Openness 1	0.01	0.11	**0.68**	0.01	-0.06
Openness 2	-0.10	0.04	**0.76**	-0.01	0.03
Openness 3	0.05	-0.14	**0.50**	0.07	0.00

Table 3.2: Factor loadings (after rotation) from a exploratory factor analysis on the reduced 19-question survey from Section 3.3.2. As is common in the psychometrics literature, loadings with magnitudes larger than 0.30-0.32 are bolded [Costello and Osborne, 2005].

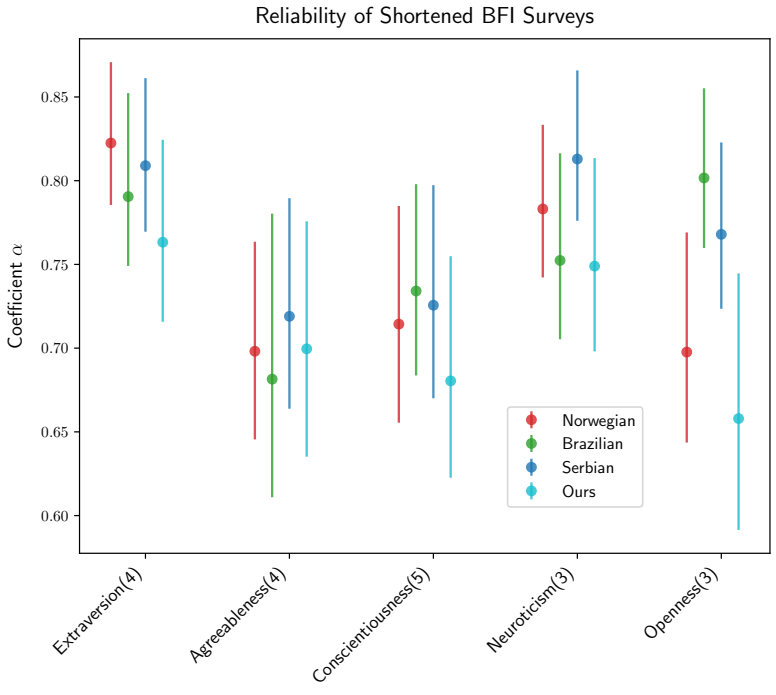


Figure 3.2: For the five personality sub-surveys, we give Cronbach's coefficient α with 95% bootstrap CIs ($B = 1000$ bootstrap samples) for our survey and the Norwegian [Engvik and Clausen, 2011], Brazilian [Veloso Gouveia et al., 2021], and Serbian [Tucaković and Nedeljković, 2023] 20-question surveys. The x-axis gives our number of selected questions per sub-survey. All other surveys chose four questions per sub-survey.

Chapter 4

A simple and unifying framework for selective inference

In this chapter we develop a simple and unified framework for selective inference centered on p-values. We first define selectively dominant p-values, a class of p-values that enable us to easily do inference after selection. We then give a precise characterization of when p-values are selectively dominant and illustrate that the most commonly used p-values are all selectively dominant. Following this discussion, we proceed by providing a few examples of how one may apply our selective dominance machinery to easily derive selective inference methods. As a closing example, we derive a novel, simple, and general data-carving procedure. To demonstrate this procedure's efficacy, we apply it to perform publication-bias aware meta-analysis on a real dataset. Finally, we provide a brief discussion on the benefits of our new framing of selective inference, and some commentary on how it fits into the current literature. A more involved discussion of how to apply our framework to the inference on winners and rank verification problems is deferred to Chapter 5 and Chapter 6 respectively.

For now, we focus on doing inference using one p-value. This allows for easier exposition and is sufficient for most problems. In Appendix B.16 we generalize our framework to accommodate inferences involving multiple p-values (provided they obey a specific dependency structure). This enables us to come up with new adaptive variants of methods that combine p-values, like Fisher's combination test, and we discuss these variants in Appendix B.16.1.

4.1 Notation

Throughout this and the remaining chapters, we will often encounter expressions like $P(A|X, B)$, where P is a probability measure, A and B are events, and X is a random variable. If B is a probability zero event, i.e., $P(B) = 0$, we set $P(A|X, B)$ to be identically zero. If not, we can define the conditional probability measure $Q(\cdot) = P(\cdot|B)$ given the event B, and we consider $P(A|X, B)$ to be the conditional expectation $E_Q[I_A|X]$ of the indicator I_A under Q. Also, for a null hypothesis H_0, the notation P_{H_0} refers to an arbitrary distribution under the null.

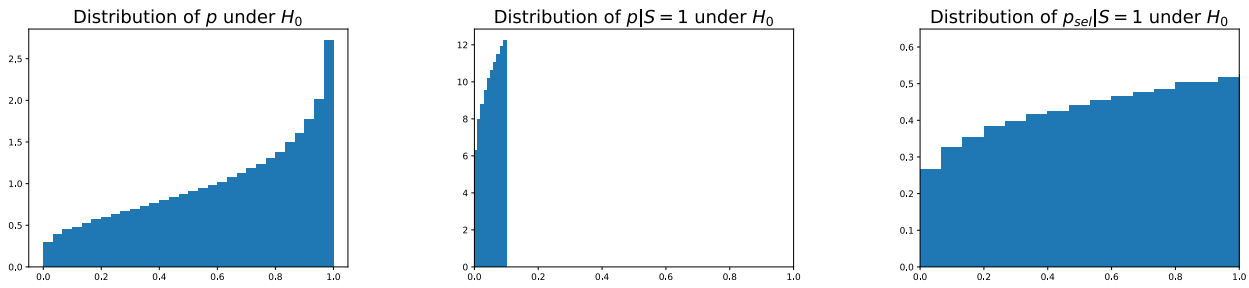


Figure 4.1: The first panel (left) depicts an example distribution of a p-value p under the null. The distribution stochastically dominates the uniform distribution. The second panel (middle) depicts the distribution $p|S=1$ of the same null p-value p given that it was selected for being most $\alpha = 0.1$. This distribution does not stochastically dominate the uniform distribution. The third panel (right) depicts the null distribution $p_{sel}|S=1$ of Theorem 5's selective p-value p/α given selection. Thanks to Theorem 5's correction, this distribution again is stochastically dominates the uniform distribution.

4.2 Selective dominance

In classical statistics, a p-value is a random variable $p \in [0,1]$ that stochastically dominates the uniform distribution $U \sim \text{Unif}([0,1])$ under the null H_0:

$$P_{H_0}(p \leq t) \leq P(U \leq t) = t \text{ for all } t \in [0,1].$$

The left-most panel of Figure 4.1 illustrates what the distribution of a null p-value may look like. When working with p-values, we maintain Type I error control if we reject H_0 when p is small:

$$P_{H_0}(\text{reject null}) = P_{H_0}(p \leq \alpha) \leq P(U \leq \alpha) = \alpha.$$

In essence, stochastic dominance allows us to use the uniform distribution as a reference distribution. We control the probability of p being small under the null by comparing it to the probability of a uniform being small.

In the problems we consider, we use p to test the null H_0 only after it has been "selected". In full generality, we consider a p-value p for testing the null H_0 that is conditionally valid given some random vector Z:

$$P_{H_0}(p \leq t|Z) \leq P(U \leq t) = t \text{ for all } t \in [0,1], \text{ a.e. under } P_{H_0}. \tag{4.1}$$

The random vector Z plays an integral role in determining the relationship between p and a binary selection random variable $S \in \{0,1\}$, which takes value one when p is selected and zero otherwise. This relationship is governed by a **selection function** [1],

$$s(x,z) = P(S = 1 \mid p = x, Z = z).$$

Intuitively, we imagine observing $p = x$ and $Z = z$ and then flipping a biased coin that comes up heads with probability $s(x,z)$. We only use p to test the null H_0 when the coin comes up heads, and otherwise do not perform inference. This process turns out to capture what happens in a wide range of selective inference

[1] Formally, the selection function $s(\cdot,\cdot)$ is an instantiation of the conditional expectation $s(p,Z) = E[S|p,Z]$.

problems. In the problems we consider, we decide the selection process, so $s(x, z)$ is known. In cases where no Z is present, we can imagine $Z = 0$ everywhere and write the selection function $s(x)$ purely in terms of x.

Because we only perform inference when the coin comes up heads, our goal should be to design a procedure that controls Type I error conditional on this selection event:

$$P_{H_0}(\text{reject } H_0 \mid S = 1) \leq \alpha \tag{4.2}$$

As illustrated by our next example, the classical approach of rejecting when $p \leq \alpha$ does not maintain selective Type I error control as in (4.2).

Example 1 (Publication bias and the failure of classical inference)**.** *Consider a p-value p that is uniform on $[0, 1]$ under the null H_0. If we use the selection function $s(x) = I(p \leq \alpha)$, i.e., we select p when it is at most α, then $p \mid S = 1 \sim$ Unif$([0, \alpha])$. Our classical procedure will clearly fail to control Type I error conditional on selection:*

$$P_{H_0}(\textit{reject } H_0 \mid S = 1) = P_{H_0}(p \leq \alpha \mid S = 1) = 1 > \alpha.$$

Example 1 is a standard example in the literature on publication bias. If researchers publish studies only when they observe a p-value below α, readers always see significant findings regardless of the true data generating distribution. As a consequence, the reader's observed Type I error rate can be as high as one. Figure 4.1's middle panel displays the distribution of the left panel's null p-value, but after it has been selected for inference via Example 1's selection process. It is clear from the picture that the null p-value distribution no longer stochastically dominates the uniform distribution after selection.

Essentially, after selection, the uniform distribution no longer suffices as a reference distribution. Naturally, we may instead try and use the distribution of a uniform after it has been selected by the same selection process. Formally, let $U \sim \text{Unif}([0, 1])$ be a uniform random variable that exists on the same probability space as p and is independent of Z, and let $S' \in \{0, 1\}$ be a different binary selection random variable whose joint distribution with U and Z is governed by the same selection function [2]

$$P(S' = 1 \mid U = x, Z = z) = s(x, z).$$

Instead of the uniform distribution, we can use the the conditional distribution $U \mid Z, S' = 1$ of U given selection as our reference distribution. This approach is valid exactly when our p-value is selectively dominant, as we define below.

Definition 1 (Selective dominance)**.** *Consider a p-value p for the null H_0 that is valid given Z as in (4.1). We say that p is **selectively dominant given** Z if, under the null H_0, it has a conditional probability density function (PDF) given Z, and it satisfies*

$$P_{H_0}(p \leq t | Z, S = 1) \leq P(U \leq t | Z, S' = 1) \textit{ for all } t \in [0, 1], \textit{ a.e. under } P_{H_0}(\cdot | S = 1) \tag{4.3}$$

for every selection function $s(x, z)$.

Essentially selective dominance requires that, under the null, the conditional distribution of p given

[2] Again, formally $E[S'|U, Z] = s(U, Z)$.

selection should stochastically dominate the conditional distribution of U given the same selection. Moreover, this should be true for any selection process. Again, we mention that if $P(S = 1) = 0$ (resp. $P(S' = 1) = 0$) for some selection function, then we define the left-hand side (resp. right-hand side) of (4.3) to be zero. We show in Appendix B.19.1 that $P(S' = 1) = 0 \implies P(S = 1) = 0$, so this convention does not cause any issues.

As we will soon see, all the p-values that practitioners commonly use are selectively dominant as described in (4.3). In Definition 1, we restrict to p-values with conditional PDFs under the null because it makes our theory and methods simpler to state. Because we can always make a p-value both have a conditional PDF and be more powerful via randomization, this restriction is never a practical issue. Also, after applying our machinery with randomized p-values, the user can always de-randomize the resulting method if they would like.

To perform valid post-selection inference using a selectively dominant p-value, we can transform it so that it remains a p-value after selection. As Theorem 5 displays, we can "undo" the effects of selection by applying the conditional cumulative distribution function (CDF) $F_{U|Z,S'=1}(\cdot)$ of U given Z and selection to p. In line with prior literature, we refer to this transformed p-value as a **selective p-value**. For simple selection functions, this selective p-value is often computable in closed form.

Theorem 5 (Selective dominance and error control)**.** *Consider a p-value p for the null H_0 that is selectively dominant given Z as in Definition 1. Then, for any selection function $s(x, z)$, the selective p-value*

$$p_{sel} = F_{U|Z,S'=1}(p) = \frac{\int_0^p s(x, Z)dx}{\int_0^1 s(x, Z)dx} \tag{4.4}$$

has a null distribution that stochastically dominates the uniform conditional on Z and selection:

$$P_{H_0}(p_{sel} \leq t \mid Z, S = 1) \leq t \text{ for all } t \in [0, 1], \text{ a.e. under } P_{H_0}(\cdot|S = 1), \tag{4.5}$$

$$P_{H_0}(p_{sel} \leq t \mid S = 1) \leq t \text{ for all } t \in [0, 1]. \tag{4.6}$$

Further, if P is a distribution in H_0 under which $P(S = 1) > 0$ and p has an exact uniform distribution given Z, i.e.,

$$P(p \leq t|Z) = t \text{ for all } t \in [0, 1], \text{ a.e. under } P(\cdot|S = 1), \tag{4.7}$$

then (4.5) and (4.6) hold with equality.

Essentially, Theorem 5 tells us that if we want selective Type I error control as in (4.2), then we should reject H_0 when p is less than the α quantile of $U \mid Z, S' = 1$ rather than the α quantile of U. We show in Appendix B.19.1 that the denominator of (4.4) is positive a.e. under $P_{H_0}(\cdot|S = 1)$, so p_{sel} is well defined whenever we actually use it to make an inferential statement.

The right-most panel of Figure 4.1 depicts what happens when we apply Theorem 5's correction to the left panel's p-value (we will derive this correction later). Unlike the null distribution of p given selection (middle panel), the null distribution of p_{sel} given selection again stochastically dominates the uniform distribution.

4.3 Characterizing selectively dominant p-values and examples

Theorem 6 tells us that p-values are selectively dominant precisely when their conditional PDF is non-decreasing under the null.

Theorem 6 (Selective dominance and increasing density)**.** *If, under the null H_0, the conditional PDF of the p-value p given $Z = z$ is non-decreasing on $[0, 1]$ for every z, then p is selectively dominant given Z as described in Definition 1. Conversely, if the conditional PDF of p given $Z = z$ is everywhere continuous and not non-decreasing on $[0, 1]$ for a set of z that has positive probability under some distribution in H_0, then p is not selectively dominant given Z.*

In what follows, we give a number of examples of selectively dominant p-values. Our examples include all the common p-values that practitioners use.

Example 2 (Two-sided testing in parametric families)**.** *Consider observing data from a parametric family P_θ and testing the null $H_0 : \theta = \theta_0$. Because the null is a point null, most p-values we construct will have an exact Unif$([0, 1])$ distribution under the null and are therefore trivially selectively dominant.*

Example 3 (One-sided testing in monotone likelihood ratio families)**.** *Consider observing one-dimensional data from a parametric family $X \sim P_\theta$ that admits density $p_\theta(x)$ with respect to some base measure μ. We say that P_θ has a monotone likelihood ratio (MLR) in the real valued function $T(x)$ if, the densities $p_\theta(x)$ share a common support and, for any $\theta \leq \theta'$, the ratio $p_{\theta'}(x)/p_\theta(x)$ is a non-decreasing function of $T(x)$. In this case, the UMP test for the null $H_0 : \theta \leq \theta_0$ rejects when $T(X)$ is large. The associated randomized p-value for this test (see Appendix B.15) is selectively dominant.*

Example 4 (Testing in in exponential families)**.** *Suppose we observe data $X \in \mathbb{R}^m$ from an exponential family P_θ parameterized by $\theta \in \mathbb{R}^n$ i.e., under P_θ the data X has density*

$$g_\theta(x) = \exp(\theta_1 T_1(X) + \cdots + \theta_n T_n(X) - \psi(\theta))g(x)$$

with respect to some base measure μ. In both the case of testing the two-sided null $H_0 : \theta_1 \neq \theta_{0,i}$ or one-sided null $H_0 : \theta_i \leq \theta_{0,i}$, the UMPU test conditions on the nuisance statistics $T_{-i}(X)$. The p-value associated with the UMPU test for $H_0 : \theta_1 \neq \theta_{0,i}$ has an exact Unif$([0, 1])$ distribution conditional on $T_{-i}(X)$, so it is trivially selectively dominant given $Z = T_{-i}(X)$. For testing $H_0 : \theta_1 \leq \theta_{0,i}$, we are in the setting of an MLR family once we condition on $T_{-i}(X)$, so Example 3 implies that the p-value associated with the UMPU test is also selectively dominant given $Z = T_{-i}(X)$.

Example 5 (Permutation testing)**.** *In a permutation test we observe data $X \in \mathcal{X}$ and compute a test statistic $T(X)$ that, under the null H_0, has a distribution that is invariant under a finite group of transformations $G : \mathcal{X} \to \mathcal{X}$. That is, $T(X) \overset{d}{=}_{H_0} T(g(X))$ for all $g \in G$. To run the test, we consider a collection of group elements $g_1, g_2, \ldots, g_w$ where $g_1 = id$ is fixed to be the identity transformation and $g_2, \ldots, g_w$ are either a random sample from G with replacement or a random sample from $G \setminus \{id\}$ without replacement. The test then rejects when $T(X)$ is large compared to the $T(g_j(X))$. In particular, the randomized permutation test from Hemerik and Goeman [2018] uses a p-value that always has an exact Unif$([0, 1])$ distribution under H_0 and is therefore trivially selectively dominant. Details about this p-value can be found in Appendix B.1*

Example 6 (F-tests for regression)**.** *Suppose we observe data $Y \in \mathbb{R}^n$ from a linear model $Y = \boldsymbol{X}\beta + \epsilon$ parameterized by $\beta \in \mathbb{R}^d$, where $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ has full column rank with probability one and is independent of $\epsilon \sim N(0, \sigma^2 I_n)$. We can run an F-test to try to reject the null $H_0 : \beta_{q+1} = \cdots = \beta_d = 0$ that the last $d - q$ coefficients are zero. The p-value associated with the F-test has an exact Unif$([0, 1])$ distribution given $\boldsymbol{X}$, so it is trivially selectively dominant given $Z = \boldsymbol{X}$. We describe this test and its associated p-value more explicitly in Appendix B.2.*

Establishing Example 3 and Example 4 is non-trivial, and the bulk of Appendix B.15 is spent doing so. The majority of the article focuses on examples from these two non-trivial settings.

4.4 Example applications of selective dominance

Having developed our machinery, we provide a few examples that illustrate how to use it.

As an introductory example, we show how to correct for Example 1's publication bias. Using our selective dominance machinery, we can provide a one-line derivation of the p-value adjustment from Hung and Fithian [2020]. Hung and Fithian [2020] derive this correction specifically for p-values coming from z- and t-tests, but our machinery applies for all selectively dominant p-values.

Example 7 (Correcting for publication bias)**.** *Suppose we have a selectively dominant p-value p for the null hypothesis H_0, and we choose to test H_0 only after observing that $p \leq \alpha$. We can apply our framework with $s(x) = I(x \leq \alpha)$. The selective p-value from (4.4) is p/α, so Theorem 5 tells us that rejecting when $p \leq \alpha^2$ controls selective Type I error:*

$$P_{H_0}(p \leq \alpha^2 | S = 1) = P_{H_0}(p/\alpha \leq \alpha | S = 1) \leq \alpha$$

In Example 1, if one computes the selective p-value p_{sel} from (4.4) carefully, they will find it equals $\min(p/\alpha, 1)$, rather than p/α as we have claimed. However, these two expressions are almost surely equal under $P(\cdot|S = 1)$, the conditional probability measure given selection. Theorem 5's result therefore applies regardless of which one we choose to use. Throughout the remainder of our examples, we will continue to give expressions for p_{sel} that, at the very least, are equal to (4.4) a.e. under $P(\cdot|S = 1)$.

Since essentially all the p-values researchers use are selectively dominant, Example 7 gives a simple way for readers to make valid inferences in the presence of publication bias: declare a studies' result significant when the associated p-value is at most α^2.

Our rule of thumb of rejecting when $p \leq \alpha^2$ should also deliver valid inferences in the presence of p-hacking. Rather than discarding an experiment after observing a p-value larger than α, researches more often tweak their analysis until the p-value crosses the significance threshold. This process, known as p-hacking, is difficult to study theoretically (which is why Hung and Fithian [2020] do not study it). But it has been empirically well-established that, under the null, p-values resulting from p-hacking have left-skewed distributions, i.e., null p-hacked p-values can be reasonably modeled as having an increasing density on $[0, \alpha]$ [Simonsohn et al., 2013]. The transformed quantity p/α then has a null density that is increasing on $[0, 1]$, so Theorem 6 guarantees that it is a valid p-value. Thus, we again find it safe to reject when $p/\alpha \leq \alpha \iff p \leq \alpha^2$.

Our second example shows how to use Theorem 5 to perform inference using the "winning" p-value. It illustrates how our selective dominance machinery enables us to easily test data-dependent hypotheses. Example 8 uses the fact that, if p is selectively dominant and Z is independent of p, then p is selectively dominant given Z. Although this is intuitively obvious, we show it carefully in Appendix B.3.

Example 8 (Inference on the winning p-value)**.** *Suppose we have n independent and selectively dominant p-values p_i for the null hypotheses $H_{0,i}$, and we choose to test only the jth null $H_{0,j}$ after observing that p_j is the smallest of the bunch. Applying our framework with $p = p_j$, $Z = p_{-j}$, and the selection function $s(x, z) = I(x < \min_k z_k)$, it is straightforward to compute that the selective p-value p_{sel} from (4.4) is $p_j / \min_{i \neq j} p_i$, so Theorem 5 tells us that rejecting when $p_j \leq \alpha \min_{i \neq j} p_i$ controls selective Type I error:*

$$P_{H_{0,j}}(p_j \leq \alpha \min_{i \neq j} p_i \mid S = 1) = P_{H_{0,j}}(p_j / \min_{i \neq j} p_i \leq \alpha \mid S = 1) \leq \alpha. \tag{4.8}$$

If we let W be the index of the smallest p-value (with ties broken randomly), it is now easy to see that rejecting the data-dependent "winning" null $H_{0,W}$ when $p_{(1)} \leq \alpha p_{(2)}$ controls Type I error both conditionally on W and marginally. Conditional error control is immediate: If $H_{0,j}$ is not true, then trivially $P(\text{falsely reject } H_{0,W} \mid W = j) = 0 \leq \alpha$. For the case that $H_{0,j}$ is true, the event $W = j$ is the same event as selecting p_j for inference in (4.8), so

$$\begin{aligned} P(\text{falsely reject } H_{0,W} \mid W = j) &= P(p_{(1)} \leq \alpha p_{(2)} \mid W = j) \\ &= P(p_j \leq \alpha \min_{i \neq j} p_i \mid W = j) \\ &\leq \alpha. \end{aligned}$$

Marginal error control then follows from the law of total probability.

$$\begin{aligned} P(\text{falsely reject } H_{0,W}) &= \sum_{j=1}^{n} P(W = j) P(\text{falsely reject } H_{0,j} \mid W = j) \\ &\leq \alpha \sum_{j=1}^{n} P(W = j) \\ &\leq \alpha. \end{aligned}$$

If the nulls are all true and the p_i are exactly uniform, then the inequalities become equalities and our error control is tight.

Rejecting the null $H_{0,W}$ when $p_{(1)} \leq \alpha p_{(2)}$ may seem like a strange procedure, but we will see later that doing so is central to the conditional inference for winners method that arises from Fithian et al. [2017] and appears often in the contemporary inference on winners literature. We discuss inference on winners in much more depth in Chapter 5.

Next, give a simple warm-up example that illustrates how our framework can be applied to rank verification problems. We discuss rank verification in much more depth in Chapter 6.

Consider the following classic hypothetical. A researcher wants to test if the unknown means of two univariate Gaussian samples, $X_1 \sim N(\mu_1, 1/\sqrt{2})$ and $X_2 \sim N(\mu_2, 1/\sqrt{2})$, are different. They end up rejecting the null hypothesis $H_0 : \mu_1 = \mu_2$ because they observe that the two-sided p-value $2(1 - \Phi(|X_1 - X_2|))$ is

below α. After rejecting, they note $X_1 > X_2$, and claim "not only are the two means are different, but they must be different because μ_1 is bigger than μ_2". The researcher, however, only rejected the null that the means are equal. Can they make a claim about the direction of inequality?

It seems that what the researcher really wants is a rank verification procedure. After observing that X_1 is larger than X_2, they want to verify that the mean μ_1 is actually larger than the mean μ_2, i.e., that they can rank the means $\mu_1 > \mu_2$. It is well known (and quite simple to show) that the researcher can indeed making their desired claim while maintaining a sensible notion of error control (see Jones and Tukey [2000]). Our framework provides new neat way of understanding why that is the case.

Example 9 (Rank verification in a simple case)**.** *Suppose that p is a selectively dominant p-value for testing the null H_0, but we only choose to test H_0 when $p < 1/2$. Applying our framework with the p-value p and selection function $s(x) = I(x < 1/2)$, Theorem 5 tells us that we control selective Type I error if we reject according to the selective p-value $p_{sel} = 2p$:*

$$P_{H_0}\left(p \leq \frac{\alpha}{2} | S = 1\right) = P_{H_0}\left(2p \leq \alpha | S = 1\right) \leq \alpha. \tag{4.9}$$

Now, consider data $X_1 \sim N(\mu_1, 1/\sqrt{2})$ and $X_2 \sim N(\mu_2, 1/\sqrt{2})$. For the one-sided nulls $H_{0,ij} : \mu_i \leq \mu_j$ the UMP p-values $p_{ij} = 1 - \Phi(X_i - X_j)$ are selectively dominant by Example 3. Denoting the winner $W = \text{argmax}_{j=1,2} X_j$ and runner-up $R = \text{argmin}_{j=1,2} X_j$, it is now clear rejecting the data-dependent null $H_{0,WR} : \mu_W \leq \mu_R$ when $p_{WR} \leq \alpha/2$ maintains Type I error control both conditionally on W and marginally. If $H_{0,ij}$ is not true, then trivially $P(\text{falsely reject } H_{0,WR} \mid W = i) = 0 \leq \alpha$. For the case that $H_{0,ij}$ is true, the event $W = i$ is identical to the event $p_{ij} < 1/2$, and hence is the same event as selecting p_{ij} for inference in (4.9). Therefore, in this case,

$$\begin{aligned} P(\text{falsely reject } H_{0,WR} \mid W = i) &= P\left(p_{WR} \leq \frac{\alpha}{2} \mid W = i\right) \\ &= P\left(p_{ij} \leq \frac{\alpha}{2} \mid W = i\right) \\ &\leq \alpha. \end{aligned}$$

This establishes error control conditional on W. Marginal error control then follows from the law of total probability:

$$\begin{aligned} P(\text{falsely reject } H_{0,WR}) &= \sum_{i=1,2} P(W = i) P(\text{falsely reject } H_{0,ij} \mid W = i) \\ &\leq \alpha \sum_{i=1,2} P(W = i) \\ &= \alpha. \end{aligned}$$

If $\mu_1 = \mu_2$ then the inequalities become equalities and our error control is tight.

Finally, we show how our framework also applies to data-carving. Specifically, we consider Fithian et al. [2017]'s variant of the file-drawer problem. Fithian et al. [2017] argue that data-splitting, which involves using one chunk of the data for selection and the remaining independent chunk for inference, is often an inadmissible approach in selective inference problems. In such settings, data-carving, as we describe below,

results in strictly more powerful procedures. Although it initially appears that data-carving does not involve selecting a p-value via the mechanism we described earlier, we use a coupling argument to show that it can be viewed in this way. This both serves to illustrate the breadth of our framework's applicability, as well as provide a new perspective on data carving.

Example 10 (Data carving and the file-drawer problem)**.** *In the file-drawer problem, we observe two independent samples $X_1, X_2 \sim N(\mu, 2)$ (e.g., X_1 comes from the first half of the data and X_2 from the second). We test the null $H_0 : \mu \leq 0$, but only when we observe that $X_1 > t$ for some $t \in \mathbb{R}$. Appendix B.4 provides details for the computations done in this example.*

Data splitting ignores the first observation, which was used for selection, and simply tests the H_0 using the p-value $p_{split} = 1 - \Phi(X_2/\sqrt{2})$. Intuitively, because this p-value is independent of the selection process, we should maintain Type I error control without any correction. Applying our framework with $p = p_{split}$, $Z = X_1$, and the selection function $s(x, z) = I(z > t)$, we indeed find that the selective p-value is the same as p_{split}.

Data-carving attempts to use the more powerful p-value $p_{full} = 1 - \Phi((X_1 + X_2)/2)$. This p-value leverages information from both samples, despite one of them being used for selection. How can we apply our framework to this data-carving problem? In how we have stated the problem, it is <u>not</u> the case that we observe p_{full} and then decide whether or not to use it for inference. Instead, we decide based on X_1, and unlike in data-splitting, p_{full} is not a valid p-value given X_1. We can compute, however, the probability that selection happened given that p_{full} took a particular value:

$$P(X_1 > t | p_{full} = x) = 1 - \Phi(t - \Phi^{-1}(1 - x)).$$

We may as well imagine that we observed that $p_{full} = x$, and then decided to use it to test the null H_0 with probability $1 - \Phi(t - \Phi^{-1}(1 - x))$. Although this is not what happens in the original problem (in our new characterization, we may test H_0 even when $X_1 \leq 3$), the conditional distribution of p_{full} given selection is the same in both cases. Hence, we can apply our framework with $p = p_{full}$ and $s(x) = 1 - \Phi(t - \Phi^{-1}(1 - x))$. Theorem 9 of Fithian et al. [2017] tells us that Theorem 5's resulting selective p-value,

$$p_{carve} = \frac{\int_0^{p_{full}} 1 - \Phi(t - \Phi^{-1}(1 - x))dx}{\int_0^1 1 - \Phi(t - \Phi^{-1}(1 - x))dx} = \frac{\int_0^{p_{full}} 1 - \Phi(t - \Phi^{-1}(1 - x))dx}{1 - \Phi(t/\sqrt{2})},$$

will result in strictly more rejections than p_{split}, while still maintaining selective Type I error control.

For the interested reader, we discuss more connections between our framework, data splitting, data carving, data fission [Leiner et al., 2023], and data thinning [Dharamshi et al., 2024, Neufeld et al., 2024] in Appendix B.5.

Our framework also applies to regression problems, including Lee et al. [2016b]'s foundational problem of doing inference on LASSO selected regression coefficients. This example is a bit long, so we have deferred it to Appendix B.6. It demonstrates how for some selective inference problems, the difficulty lies in characterizing what the selection function is.

Examples 7 to 10 all share a common theme. In all three examples, the practitioner cheats. They peek at the p-value and, to varying degrees, they only test the null when the p-value looks promising. The purpose of selective procedures is to adjust the p-value in a way that accounts for this cheating. The harsher the

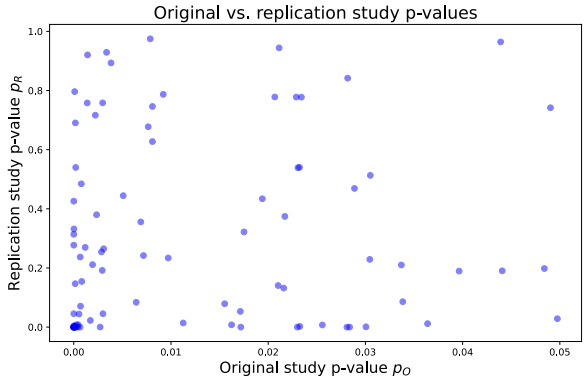


Figure 4.2: Scatter plot of the original p_O and replication p_R p-values for 92 psychology studies from the open science collaboration's replication analysis Collaboration [2015]. Note that the x-axis, which ranges from $[0, 0.05]$ is on a different scale than the y-axis, which ranges from $[0, 1]$.

cheating is, the more this adjustment inflates the original p-value.

4.5 Publication-bias aware meta-analysis

Performing replication studies is a crucial part of the scientific process, especially when one is wary that the original study may suffer from publication bias. To judge the prevalence of publication bias in psychology, the open science collaboration conducted a mass replication analysis of psychology studies [Collaboration, 2015]. Via their efforts, we have access to p-values from 92 pairs of original and replication psychology studies [3], depicted in Figure 4.2. We refer to p-values from the original study as p_O and p-values from the replication study as p_R. The p-values p_O from the original studies are significant at the $\alpha = 0.05$ level, while only 34 of the replication p-values p_R are significant.

Although the original study p-values suffer from publication bias, they still contain valuable and usable information. By Example 7 and its subsequent discussion, p_O/α should be a valid p-value even in the presence of publication bias or p-hacking. Then, via Fisher's combination test, we can use both the corrected original p-value p_O/α and uncorrected replication p-value p_R for inference. As a refresher, Fisher's test considers n independent p-values p_i for the nulls $H_{0,i}$ and rejects the global null $\cap_{i=1}^n H_{0,i}$ when the test statistic $-2\sum_{i=1}^n \log(p_i)$ is at least as large as the $1-\alpha$ quantile of the χ^2_{2n} distribution. In our case, we have two independent p-values p_O/α and p_R that test the same null hypothesis, and we can reject this null hypothesis when

$$-2(\log(p_O/\alpha) + \log(p_R)) \geq \text{Quantile}(1-\alpha, \chi^2_4).$$

This approach is analogous to data-carving. After using part of our data for selection (the original study) and part for inference (the replication study), we still make use of the information remaining in the first part after selection for inference as well. Unlike existing approaches to data-carving, which are often complex and problem specific, using Fisher's combination test along with our selective dominance framework provides a general and simple way to data-carve.

Our combination approach allows us to make more powerful inferences on the open science dataset. Our

[3] We exclude seven studies with original p-value p_O larger than $\alpha = 0.05$.

approach finds that 47 study pairs have significant findings, whereas using solely the original study p-value p_O/α or the replication study p-value p_R results in only 39 or 34 significant findings respectively. It is surprising that, even after a harsh adjustment for publication bias, the corrected original study p-values result in more discoveries than replication study p-values. It is hard to gauge if this is due to chance, differences between the original and replication studies (e.g., minor differences in population demographics, devices used for measurement, sample size), or because there is somehow even harsher selection bias in the original studies than what we have accounted for. It may even be the case that some replicators felt incentivized to induce bias in the opposite direction, and tried to ensure that the replication studies were not significant.

4.6 Some discussion

First, we discuss why we believe that our treatment of selective inference is easier to grasp than what is presented in the current literature. Plainly stated, when statisticians and practitioners alike are presented with a null hypothesis H_0, the first thing they do is come up with a marginally valid p-value p. The standard selective inference approach is to come up with a quantity that is pivotal under the conditional distribution of the test statistic given selection. This can be a daunting task, and our framework (specifically Theorem 5) provides an automatic way of finding such a pivot via the marginal p-value p, an object that people are familiar with and is common to all inferential problems. Practitioners in particular are very accustomed to methods that take a marginal p-value p and perform an adjustment to it (e.g. a Bonferonni correction multiplies each p-value by n, a Benjamini-Hochberg adjustment multiplies the ith smallest p-value by n/i). Hopefully, this simplified view changes selective inference from a subject that is taught in a topics statistics course (graduate level) to one that can be covered immediately after (or even during) the lecture on p-values in an introductory statistics course (graduate level).

Second, we discuss a technical contribution of our framework. The current selective inference literature is also mostly focused on making selective confidence intervals, i.e., coming up with a selective test for the point null $H_0 : \theta = \theta_0$ and inverting it to get an interval for θ_0. As such, there is very little literature on one-sided selective testing. The only examples of one-sided selective tests that we are aware of establish validity via nice properties of exponential families (e.g., Fithian et al. [2017], Hung and Fithian [2019b]). As we have seen, however, exponential family data is not necessary for us to make one-sided selective tests. In reality, the real driving force behind our ability to do selective inference is that the data came from an MLR family, or more specifically that the marginal p-value has a non-decreasing density under the null. This sort of shape-constraint is the important property at the crux of many existing procedures (e.g., Hung and Fithian [2020], Lei and Fithian [2018], Zhao et al. [2019]). In fact, both Hung and Fithian [2020], Zhao et al. [2019] leverage this shape-constraint to perform selective inference identical to us, but only in cases where the data is continuous, there is no random variable Z that selection is allowed depend on, and the selection function is $s(x) = I(x \leq \tau)$ for some fixed threshold τ. Hung and Fithian [2020], Zhao et al. [2019] only consider our p-value approach for one-off use cases, and do not provide a full framework as a consequence. In contrast, we fully flesh out this approach so that it can accommodate essentially all modern selective inference techniques.

Chapter 5

Inference on winners via selective inference

In this chapter, we use Chapter 4's framework to study the inference on winners problem. First, we characterize our motivating selective inference problem from Chapter 1 in terms of p-values, and demonstrate that it boils down to Example 8 from Chapter 4. Written in terms of p-values, the method becomes easier to derive, interpret, implement, and generalize beyond Gaussianity. Particularly, we are able to gain insight on (1) when the conditional lower confidence bound (LCB) from (1.3) could be useful and (2) the extent to which the conditional LCB explodes to $-\infty$. Also, the p-value viewpoint enables us to apply the closure principle, yielding a new conditional sequential hypothesis testing procedure that makes strictly more discoveries than the original method.

Next, we provide theoretical results that suggest that, even though the conditional confidence region from Chapter 1 suffers from the exploding interval problem, this conditional approach can still be quite powerful in certain settings. Particularly, we show that it attains the maximum possible power in a generalized variant of the needle-in-a-haystack problem (when the needle is sufficiently well separated from the haystack). For sake of comparison, we also study locally simultaneous inference in this setting [Zrnic and Fithian, 2024]. Locally simultaneous inference is a competing method that is (1) strictly more powerful than the classical simultaneous inference approach and (2) designed to maintain the benefits of conditional inference while avoiding the exploding interval problem. We show that locally simultaneous inference suffers from a curse of dimensionality that renders it powerless, even when the conditional approach achieves the maximum possible power.

Finally, we use our p-value characterization to derive a new variant of hybrid inference [Andrews et al., 2023], an inference on winners procedure that interpolates between the conditional and simultaneous inference approaches. Hybrid inference manages to avoid both the exploding intervals of conditional inference and the curse of dimensionality of locally simultaneous inference. Again, writing hybrid inference in terms of p-values makes it easier to derive, interpret, implement, and generalize beyond Gaussianity. It also again enables us to apply the closure principle to find a new hybrid sequential hypothesis testing procedure that makes strictly more discoveries than the original approach. Our p-value viewpoint also illustrates that hybrid inference strictly dominates its naive union bound competitor, which splits the Type I error budget

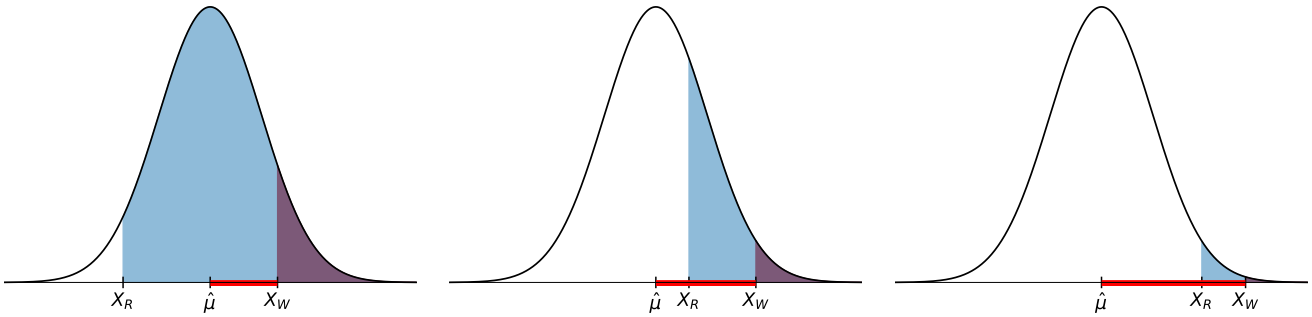


Figure 5.1: We plot the level $\alpha = 0.1$ conditional LCB $\hat{\mu}$ for different gaps between the winning value X_W and the runner up value X_R and highlight the distance between $\hat{\mu}$ and X_W in red. The LCB $\hat{\mu}$ is chosen exactly so that the tail probability $P(N(\hat{\mu}, 1) > X_R)$, shaded in blue, is $1/\alpha = 10$ times the tail probability $P(N(\hat{\mu}, 1) > X_W)$, shaded in red (the overlap appears purple). As X_W and X_R get closer, we need to take $\hat{\mu}$ further back for this condition to be satisfied.

between the conditional and simultaneous approaches and then takes the better of the two. Unfortunately, our viewpoint also suggests that, while it results in a strict improvement, hybrid inference cannot do too much better than this naive alternative.

This chapter focuses on the independent data setting. The problems we discuss in Chapter 6, however, illustrate how one can perform inference on winners when the data comes from a multi-parameter exponential family or multivariate Gaussian with known covariance. This encompasses many correlated data settings.

5.1 Conditional inference

In this sub-section we discuss Fithian et al. [2017]'s conditional approach for performing inference on winners. This conditional approach turns out to be highly related to the testing procedure that we derived in Example 8. To be concrete, we will imagine performing inference at level $\alpha = 0.1$.

Corollary 1 (Testing the winner)**.** *Suppose that p_i are n independent and selectively dominant p-values for the nulls $H_{0,i}$, and let W be the index of the smallest p-value (with ties broken randomly). Rejecting $H_{0,W}$ when $p_{(1)} \leq \alpha p_{(2)}$ controls Type I error at level α conditionally on W, and therefore also marginally.*

Unlike Sidak's simultaneous approach, which rejects the winning null when the smallest p-value is small in absolute terms, the conditional approach rejects the winning null when the smallest p-value is small relative to the second smallest p-value. This procedure is strange, but fairly easy to interpret: we reject the winning null when the most extreme observation is $1/\alpha = 10$ times more extreme under its null than the second most extreme observation is under its null. This can be quite a stringent requirement!

Once written in terms of p-values, its easy to mathematically see the merits and pitfalls of the conditional approach. If all the p-values except the smallest provide essentially no evidence against the null, then $p_{(2)} \approx 1$ and we reject when $p_{(1)} < \alpha$, the same p-value cutoff as a one-dimensional problem. On the other hand, if even just one other p-value provides a similar amount of evidence against the null as the smallest p-value, then then $p_{(2)} \approx p_{(1)}$ and we will never reject because $p_{(1)} \not\leq \alpha p_{(1)} \approx \alpha p_{(2)}$.

The p-value viewpoint also makes it clear that the conditional approach can only outperform the simultaneous approach when some null p-values are conservative (i.e., they have super-uniform distributions). We

give a heuristic argument here, and a more formal statement in Appendix B.7. Suppose one of our p-values p_1 is a very strong signal (so it is very small with high probability) but the remaining p-values $p_2, \ldots, p_n$ are null p-values that are uniform (i.e., they are not conservative). Our conditional procedure will reject when our smallest p-value, likely p_1, is less than α times the smallest of $p_2, \ldots, p_n$. The minimum of these $n-1$ uniform p-values is $1/n$ on average. Hence, roughly speaking, the conditional approach also rejects when $p_{(1)}$ is less than α/n, which is essentially the same as Sidak's simultaneous approach for moderately sized n or small α.

5.1.1 Confidence regions for the winner

In parametric settings, we can invert Corollary 1's test to get selective confidence regions for the winning parameter. Consider observing independent data $X_i \sim P_{\theta_i}$ from an MLR family P_θ parametrized by $\theta \in \mathbb{R}$. Let $p_i^{\theta_0}$ (which is a function of solely X_i) be the UMP p-value for testing the null $H_{0,i}^{\theta_0} : \theta_i \leq \theta_0$. Details regarding these p-values can be found in Appendix B.15.1. We can define the winner $W = \operatorname{argmin}_{j \in [n]} p_j^{\theta_0}$ to be the index of the smallest and therefore most promising p-value. This winning index will be the same irrespective of θ_0 [1]. By inverting Corollary 1's test, we get an LCB

$$\{\theta_0 : p_{(1)}^{\theta_0} / p_{(2)}^{\theta_0} > \alpha\} \tag{5.1}$$

for the winning parameter θ_W that holds conditionally on W with probability exactly $1-\alpha$:

$$P(\theta_W \in \{\theta_0 : p_{(1)}^{\theta_0} / p_{(2)}^{\theta_0} > \alpha\} | W) = 1 - \alpha.$$

The fact that the confidence region (5.1) is actually an LCB is a consequence of the selective p-value $p_{(1)}^{\theta_0} / p_{(2)}^{\theta^0}$ being monotone non-decreasing in null parameter θ_0. Appendix B.15.3 provides general conditions (which will hold for essentially every selective problem) under which selective p-values like $p_{(1)}^{\theta_0} / p_{(2)}^{\theta^0}$ are monotone in the null parameter. We show these conditions apply here, and also argue that (5.1) has exact $1-\alpha$ coverage in Appendix B.8. We also show in Appendix B.8 how to invert Corollary 1's test to get a confidence interval (CI) instead of LCB, and that both our CI and LCB match Fithian et al. [2017]'s approach in the Gaussian case.

Writing the conditional LCB terms of p-values helps us better understand its behavior. In particular, recalling our motivating Gaussian example (1.3), we learn that Figure 1.1's LCB stretches back exactly to the $\hat{\mu}$ under which it is $1/\alpha = 10$ times less likely to see something as extreme as the winner than something as extreme as the runner-up. Noting this fact, Figure 5.1 illustrates why the LCB (1.3) diverges to $-\infty$ as the winner and runner-up get closer. If the winner and runner-up are very close, we need the LCB $\hat{\mu}$ to be very far back for the winner to be ten times more extreme than the runner-up. Thanks to the decay of the Gaussian tail, however, we can always find such a mean if we go far back enough.

For non-Gaussian data, the amount the conditional LCB (5.1) stretches back depends on the right tail decay of P_θ. The faster the tail decays as θ decreases, the larger the first $\hat{\theta}$ for which the winner is $1/\alpha = 10$ times as extreme as the runner-up. Seeing as the Gaussian distribution, which has a rapidly decaying tail, still results in exploding lower bounds, we should expect that the distance from the winning observation to

[1] If we use the same auxiliary randomness to compute $p_i^{\theta_0}$ for every θ_0, then one index will result in the smallest p-value for every θ_0 and W. The way we construct our UMP p-values in Appendix B.15.1 ensures that this is the case.

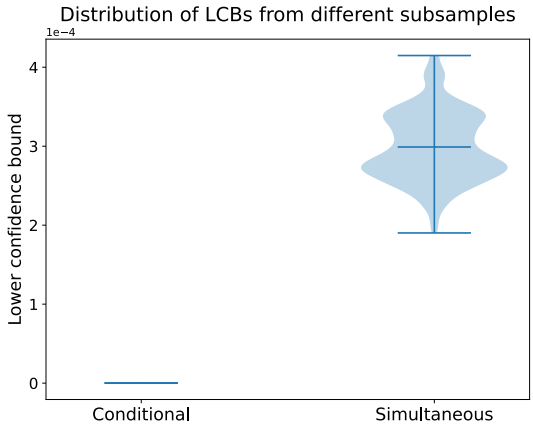


Figure 5.2: Over $B = 1000$ different sub-samples of $n = 2$ failure times from the dataset Molotaliev [2024], the distribution of the conditional and simultaneous LCB for the "winning" parameter. The conditional LCB is always vacuous.

the lower bound can often be quite large in many settings.

As an example, consider observing independent exponential random variables $X_i \sim \text{Exp}(\lambda_i)$. Crucially, the parameter space $\lambda \in (0, \infty)$ is bounded below. The exponential distribution has an MLR in $T(x) = 1/x$, so the UMP test for $H_{0,i}^{\lambda_0} : \lambda_i \leq \lambda_0$ uses a p-value $p_i^{\lambda_0}$ that rejects when X_i is small. It turns out that

$$\lim_{\lambda^0 \downarrow 0} p_{(1)}^{\lambda_0} / p_{(2)}^{\lambda_0} = X_{(1)}/X_{(2)},$$

so the conditional LCB for the winning parameter λ_W,

$$\{\lambda_0 : p_{(1)}^{\lambda_0} / p_{(2)}^{\lambda_0} > \alpha\}, \tag{5.2}$$

is vacuous whenever $X_{(1)}/X_{(2)} > \alpha$, i.e., with positive probability the confidence region (5.1) spans the whole parameter space $(0, \infty)$. A careful derivation of this test and result can be found in Appendix B.9.

The failure of the exponential conditional LCB (5.2) manifests in real data examples. On a dataset of car engine failure times [Molotaliev, 2024], we find that the conditional LCB (5.2) is always vacuous. The dataset has the failure times of one-hundred car engines, which we model as independent exponential random variables. Over many sub-samples of just $n = 2$ failure times, the LCB (5.1), which does inference on the worse of the two engines, always gives a vacuous lower bound of zero. In contrast, the simultaneous approach always gives a non-vacuous lower bound. Figure 5.2 depicts the results. The result is concerning. Despite conditional LCB having exact $1 - \alpha = 0.9$ coverage, our empirical coverage is one (the vacuous LCB must always cover the parameter). This suggests that the exponential distribution is likely not an appropriate model for this dataset (the failure times are not spread out enough), even though it is often a natural choice for modeling failure times.

5.1.2 More discoveries via the closure principle

Once written in terms of p-values, it is natural to treat Corollary 1's test as a test of the global null and try to close it (as in Marcus et al. [1976]). Closing a global null test precludes us from making confidence regions, but it allows us to make more individual discoveries. Closed global null testing procedures are often computationally intractable to implement, so it is interesting that Corollary 1's global null test admits a tractable closure.

Corollary 2 (Closed testing for winners)**.** *Suppose that p_i are n independent and selectively dominant p-values for the nulls $H_{0,i}$. Let $H_{0,(j)}$ denote the null corresponding to the jth smallest p-value (with ties broken randomly) and define $p_{(n+1)} = 1$. Rejecting the null hypotheses $H_{0,(k)}$ for which $p_{(j)} \leq \alpha p_{(j+1)}$ for all $j \leq k$ controls the family-wise error rate at level α.*

As is often the case for closed procedures, Corollary 2 procedure is best understood sequentially. We reject $H_{0,(1)}$ when $p_{(1)} \leq \alpha p_{(2)}$. Then, if we rejected $H_{0,(1)}$, we reject $H_{0,(2)}$ when $p_{(2)} \leq \alpha p_{(3)}$, so on and so forth until we fail to reject.

5.2 Conditioning can gain power

In this section we will study the power of the conditional inference on winners procedure from Corollary 1. To do so, we will consider a sequence of problems with increasing dimension. Particularly, we imagine observing isotropic Gaussian data $X^{(n)} \sim N(\mu^{(n)}, I_n)$ centered at a mean vector $\mu^{(n)}$. Defining the individual nulls $H_{0,i}^{(n)} : \mu_i^{(n)}$ our goal is to test the global null $\cap_{i=1}^n H_{0,i}^{(n)} : \mu_i^{(n)} \leq 0$ that all the means are less than or equal to zero. Because we now have a sequence of problems, we also have a sequence of winners and runner-ups. We will let $W(n)$ and $R(n)$ denote the indices of the largest and second largest of the $X_j^{(n)}, 1 \leq j \leq n$ respectively. To apply Corollary 1, we consider the p-values $p_i^{(n)} = 1 - \Phi(X_i^{(n)})$ which are valid selectively dominant p-values for the nulls $H_{0,i}^{(n)}$ be Example 3.

It will be helpful to recall that if Z_i are independent standard Gaussian random variables, then

$$\max_{i=1}^{n} Z_i = \sqrt{2 \log n} + o_p(1),$$

Meaning roughly that we can treat $\max_{i=1}^n Z_i$ as non-random and equal to $\sqrt{2 \log n}$ when n is large. We will be somewhat informal in our discussion, but rigorous proofs of all the results can be found in the appendix.

5.2.1 Power of the conditional approach

First, we characterize a regime where Corollary 1's procedure is powerful. To do so, we focus our attention on a generalization of the needle-in-the-haystack problem.

In the typical needle-in-the-haystack problem, we imagine testing the null that all the means are exactly equal to zero, and consider alternatives where one mean is positive and the remaining means are zero, e.g., $\mu_1^{(n)} > 0$ and $\mu_j^{(n)} = 0$ for $j > 1$. It is well known that for such alternatives, Sidak's procedure (and also Bonferonni), which rejects when

$$1 - \Phi(X_j^{(n)}) \leq \alpha_n = 1 - (1 - \alpha)^{1/n} \approx \alpha/n \quad \text{for any } 1 \leq j \leq n,$$

has limiting power equal to one whenever $\mu_1^{(n)} \gg \sqrt{2\log n}$. It is quite intuitive that the test is powerful in this regime. Under the null, the maximum observation we see should be around $\sqrt{2\log n}$. If $\mu_1^{(n)} \gg \sqrt{2\log n}$., then $X_1^{(n)}$ will be noticeably larger than than this threshold, and we can be sure the data did not come from the null. On the other hand, if $\mu_1^{(n)} \ll \sqrt{2\log n}$, then no observation will be larger than $\sqrt{2\log n}$. Sidak's procedure will be unable to distinguish between the alternative and null, and it will be powerless in the limit.

In our generalized needle-in-the-haystack problem, we imagine testing the composite global null $H_0^{(n)}$ that all the means are at most zero, and consider alternatives where one mean is positive and the remaining means are non-positive, e.g., $\mu_1^{(n)} > 0$ and $\mu_j^{(n)} \leq 0$ for $j \leq 1$. Our result tells that whenever the gap between the winner $X_{W(n)}^{(n)}$ and runner-up $X_{R(n)}^{(n)}$ diverges to infinity in probability (i.e., $X_{W(n)}^{(n)} - X_{R(n)}^{(n)} \xrightarrow{p} \infty$), Corollary 1's procedure, which rejects when

$$\frac{1 - \Phi(X_{W(n)}^{(n)})}{1 - \Phi(X_{R(n)}^{(n)})} \leq \alpha, \tag{5.3}$$

has the maximal possible limiting power. To be concrete, $X_{W(n)}^{(n)} - X_{R(n)}^{(n)} \xrightarrow{p} \infty$ means for all $B \in \mathbb{R}$ that

$$\liminf_{n\to\infty} P(X_{W(n)}^{(n)} - X_{R(n)}^{(n)} > B) = 1.$$

To provide some clarity on this condition, we give some sequences of mean vectors $\mu_n^{(n)}$ for which it holds. For example, $X_{W(n)}^{(n)} - X_{R(n)}^{(n)} \xrightarrow{p} \infty$ under the sequence of alternatives

$$\mu_j^{(n)} = \begin{cases} \sqrt{3\log n} & \text{if } j = 1, \\ 0 & \text{if } j > 1. \end{cases}$$

For this sequence, $X_1^{(n)} \approx \sqrt{3\log n}$ will be the clear winner and will have an increasing gap to the runner-up, which is the maximum $\max_{j>1} X_j^{(n)} \approx \sqrt{2\log n}$. Identical reasoning tells us that $X_{W(n)}^{(n)} - X_{R(n)}^{(n)} \xrightarrow{p} \infty$ under the sequence

$$\mu_j^{(n)} = \begin{cases} 1 & \text{if } j = 1, \\ 1 - \sqrt{3\log n} & \text{if } j > 1. \end{cases}$$

Interestingly, for this later sequence, Sidak's test (which is also valid for the composite null $H_0^{(n)}$) has limiting power zero! This is not the case for Corollary 1's procedure, which will have the highest possible limiting power of $1 - \Phi(\Phi^{-1}(1-\alpha) - 1) > 0$.

How is it possible that Corollary 1's procedure can be powerful while Sidak's procedure is powerless? Intuitively, it is possible for a procedure to be powerful because the observations $X_j^{(n)}$ for $j > 1$ are so negative that it is obvious upon observing them that they come from null means. Corollary 1 knows to ignore these observations, and simply draw an inference using $X_1^{(n)}$, while Sidak's procedure does not. More generally, when the gap between the winner and runner-up is big, it is clear that the winner was the largest not because of selection bias, but because it came from a large mean. In the composite null problem, there are many more interesting ways for this gap to be big. Unlike Sidak's procedure, Corollary 1's procedure is able to adapt to the compositeness of $H_0^{(n)}$.

Ideally, we would characterize when $X_{W(n)}^{(n)} - X_{R(n)}^{(n)} \xrightarrow{p} \infty$ in terms of some conditions on the mean vector

$\mu^{(n)}$. This is difficult, however, as there are many different sequences of needle-in-the-haystack mean vectors that elicit this behavior. For example, we later justify that $X^{(n)}_{W(n)} - X^{(n)}_{R(n)} \xrightarrow{p} \infty$ for the sequence

$$\mu^{(n)}_j = \begin{cases} \log\log n & \text{if } j = 1, \\ \log\log n - 2\sqrt{\log n} & \text{if } 1 < j \leq \lfloor \sqrt{n} \rfloor, \\ -n & \text{if } \lfloor \sqrt{n} \rfloor + 1 \leq j \leq n. \end{cases} \tag{5.4}$$

If we wanted, we could keep coming up with different sequences $\mu^{(n)}$ that satisfy this condition, and it is not clear how to characterize all such sequences cleanly. Given a sequence $\mu^{(n)}$, however, it is typically easy to argue whether $X^{(n)}_{W(n)} - X^{(n)}_{R(n)} \xrightarrow{p} \infty$ or not.

Theorem 7 below is a summary of our discussion, and it gives a formal statement of our result.

Theorem 7. *Consider data $X^{(n)}_i \sim N(\mu^{(n)}_i, 1)$ and the global null $H^{(n)}_0 = \cap_{i=1}^n H^{(n)}_{0,i}$ with $H_{0,i} : \mu^{(n)}_i \leq 0$. If $\mu^{(n)} \in \mathbb{R}^n$ is a sequence of alternatives such that*

1. *$\mu^{(n)}_1$ is positive,*
2. *$\mu^{(n)}_j$ is non-positive for $j > 1$,*
3. *$X^{(n)}_{W(n)} - X^{(n)}_{R(n)} \xrightarrow{p} \infty$, where $W(n)$ and $R(n)$ (resp.) are indices of the largest and second largest (resp.) $X^{(n)}_i$ for $1 \leq i \leq n$,*

then Corollary 1's test for $H^{(n)}_0$ attains the maximal possible limiting power. That is,

$$\liminf_{n\to\infty} \inf_{\phi_n \in \mathcal{H}_n} P(\phi^{cond}_n \text{ rejects }) - P(\phi_n \text{ rejects }) \geq 0$$

where $\mathcal{H}_n$ is the set of all tests that maintain level α Type I error control under $H^{(n)}_0$, and ϕ^{cond}_n is the test that rejects when (5.3) holds.

In wrapping up this part of the discussion, we emphasize that the regime we have studied is a regime where we should expect methods to be powerful. There is only one possible signal, and the assumption $X^{(n)}_{W(n)} - X^{(n)}_{R(n)} \xrightarrow{p} \infty$ essentially ensures that this signal is well-separated and easy to identify. It is believable that a procedure should be able to adapt to the data generating distribution and reduce the problem to classical one-dimensional Gaussian mean testing. If anything, it is less surprising that Corollary 1's procedure is powerful, but rather more concerning that Sidak's simultaneous inference procedure can be powerless. Even more surprising and concerning is what we discuss next. Locally simultaneous inference, a simultaneous inference procedure designed specifically to be adaptive to the compositeness of $H^{(n)}_0$, can also be powerless in this easy regime.

5.2.2 Comparison to locally simultaneous inference

In this sub-section we discuss applying locally simultaneous inference (LSI) to the previous sub-section's problem. LSI is an adaptive simultaneous inference technique that is strictly more powerful than vanilla simultaneous inference. In full generality, LSI can be applied to a wide variety of multi-dimensional inferential problems. We only give a high-level overview of the method, and discuss only how it can be used to perform

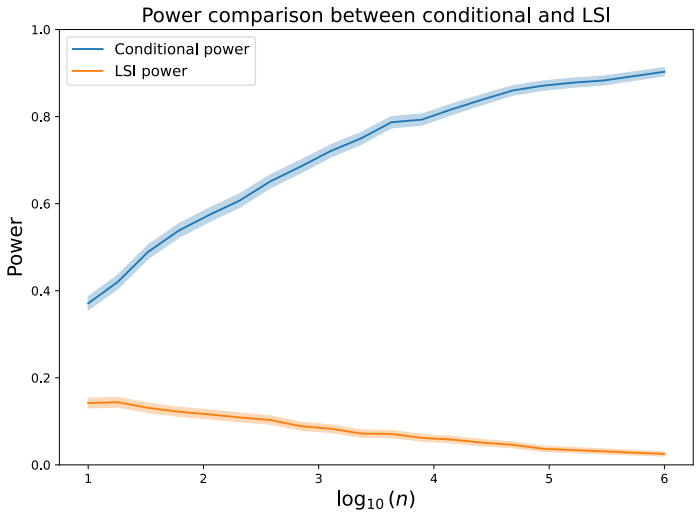


Figure 5.3: Power of Corollary 1's procedure versus that of locally simultaneous inference ($\alpha = 0.1$, $\beta = \alpha/10$) for the sequence of needle-in-the-haystack alternatives $\mu^{(n)} \in \mathbb{R}^n$ given by $\mu_1^{(n)} = \sqrt{0.8 \log n^\zeta}$, $\mu_i^{(n)} = \sqrt{0.8 \log n^\zeta} - 2\sqrt{\log n}$ for $i = 2, \ldots, \lfloor n^\zeta \rfloor$, and $\mu_i^{(n)} = -n$ for $i = \lfloor n^\zeta \rfloor + 1, \ldots, n$, with $\zeta = 0.9$. Plot shows that LSI becomes powerless as $n \to \infty$, while the conditional method has power approaching one. Shaded region denotes one standard error bands.

inference on winners in the isotropic Gaussian data setting. A full treatment of LSI can be found in the original article [Zrnic and Fithian, 2024].

It is worth mentioning that our Gaussian inference on winners is the motivating problem for LSI. One express goal of LSI is to retain the benefits of Corollary 1's conditional procedure while avoiding the exploding interval problem discussed in Chapter 1. In an attempt to achieve this, LSI avoids conditioning and instead leverages solely marginal simultaneous inference techniques.

At a high level, the goal of LSI is to avoid performing simultaneous inference on all n means. Instead, LSI uses the data to determine a data-adaptive subset of means that would have been likely to produce a winner [2], and applies a simultaneous inference correction as if we only ever considered doing inference on these means. Despite this simultaneous correction being data-dependent, LSI still manages to return a provably valid inference.

Getting more into the weeds, we explicitly state when LSI rejects the global null $H_0^{(n)}$. LSI splits the α Type I error budget into two parts, $0 < \beta^{(n)} < \alpha$ and $\gamma^{(n)} = \alpha - \beta^{(n)}$. The $\beta^{(n)}$ part of the budget is used to determine which observations could have potentially been the winner. Define

$$N = |\{j : X_j^{(n)} \geq X_{W(n)}^{(n)} - 4z_{1-\beta_n^{(n)}/2}\}| \tag{5.5}$$

to be the random number of observations that are within $4z_{1-\beta_n^{(n)}/2}$ of the winner. The remaining $\gamma^{(n)}$ of the budget is used to perform simultaneous inference, but the simultaneous correction pretends that only these N observations were ever present. Explicitly, LSI rejects $H_0^{(n)}$ when

$$1 - \Phi(X_{W(n)}^{(n)}) \leq \max\{\gamma_N^{(n)}, \alpha_n\}, \qquad \gamma_N^{(n)} = 1 - (1 - \gamma^{(n)})^{\frac{1}{N}} \tag{5.6}$$

[2] Technically, it finds a subset that, with high probability, contains a subset of means that would have been likely to produce the winner

Note that, because of the maximum in (5.6), LSI will never do worse than Sidak's procedure. The fact that rejecting according to the condition (5.6) is a Type I error controlling procedure is a consequence of Theorem 3 in Zrnic and Fithian [2024].

Intuitively, LSI should perform well whenever Corollary 1's conditional procedure does. When the gap between the winner and another observation is sufficiently large, LSI realizes that this other observation had minimal chance of winning, and simply ignores it (i.e., it recognizes that the winner is large because it came from a larger mean and not because of selection bias). If the gap between the winner and every observation is large enough, then we will have $N = 1$ and LSI totally avoids doing any simultaneous correction (i.e., LSI reduces to a one-dimensional Gaussian mean testing problem). In this case, LSI should be as powerful as is possible. But what is a large enough gap for LSI?

Theorem 8 gives a surprising result. LSI can be powerless in needle-in-the-haystack settings even when Corollary 1's conditional test attains maximal power. The proof sketch below gives an idea of why.

Theorem 8. *Consider data $X_i^{(n)} \sim N(\mu_i^{(n)}, 1)$ and the global null $H_0^{(n)} = \cap_{i=1}^n H_{0,i}^{(n)}$ with $H_{0,i} : \mu_i^{(n)} \leq 0$. For every $\delta \in [\alpha, 1]$, there exists a sequence of alternatives $\mu^{(n)} \in \mathbb{R}^n$ satisfying the assumptions of Theorem 7 such that, for all Type I error budgeting sequences $0 < \beta_n^{(n)} < \alpha$, LSI's test has zero limiting power despite the maximal limiting power being δ. That is,*

$$\limsup_{n\to\infty} P(\phi_n^{LSI} \text{ rejects }) = 0.$$

where ϕ_n^{LSI} is the test that rejects when (5.6) holds, despite the fact that

$$\liminf_{n\to\infty} \sup_{\phi_n \in \mathcal{H}_n} P(\phi_n \text{ rejects }) = \delta,$$

where $\mathcal{H}_n$ is the set of all tests that maintain level α Type I error control under the global null $H_0^{(n)}$.

Proof sketch. We show the case where $\delta = 1$. Consider the sequence of alternatives given by (5.4). The test that rejects when $1 - \Phi(X_1^{(n)}) \leq \alpha$ is in $\mathcal{H}_n$ and obviously has limiting power one. Based on how we have set our means, we will have

$$\begin{aligned}
X_1^{(n)} &\approx \log\log n \\
\max_{2\leq i\leq \lfloor\sqrt{n}\rfloor} X_i^{(n)} &\approx \log\log n - 2\sqrt{\log n} + \sqrt{2\log\sqrt{n}} = \log\log n - \sqrt{\log n} \\
\min_{2\leq i\leq \lfloor\sqrt{n}\rfloor} X_i^{(n)} &\approx \log\log n - 3\sqrt{\log n} \\
\max_{\lfloor\sqrt{n}\rfloor+1\leq i\leq n} X_i^{(n)} &\approx \sqrt{2\log n} - n
\end{aligned}$$

Therefore we see that $X_1^{(n)}$ will be a clear winner and also have a gap to the runner-up which diverges to ∞ in probability. Further, there are $\sqrt{n}$ observations within $4z_{1-\beta_n^{(n)}/2} \succeq 4\sqrt{2\log n}$ of the winner, so $N \approx \sqrt{n}$ (where N is from (5.5)). Because $\mu_1^{(n)}$ is growing very slowly, the simultaneous correction over $\sqrt{n}$ observations will be too harsh and LSI will be powerless. □

In Figure 5.3, we provide simulation results that verify Theorem 8's conclusion. More specifically, we consider a particular sequence of needle-in-the-haystack alternatives (similar to the one in the proof sketch),

and show empirically that LSI is powerless as $n \to \infty$, whereas Corollary 1's test has a limiting power of one.

Briefly, we mention that because LSI is at least as powerful as Sidak's procedure, it is easy to come up with non needle-in-the-haystack alternatives where LSI attains maximal power but Corollary 1's procedure is powerless. For example, taking $\mu_i^{(n)} = 1$ for all $1 \leq i \leq n$ would suffice. It remains concerning, however, that LSI is designed to be powerful for needle-in-the-haystack alternatives, but it sometimes fails to be.

Theorem 8 suggests that LSI suffers from a curse of dimensionality that the conditional procedure avoids. We discuss why this may be the case in the next sub-section.

5.2.3 The (seeming) necessity of conditioning

Finally, we provide some brief discussion on why conditioning may be the crucial tool that allowed us to gain power in this needle-in-the-haystack setting. Initially, this may seem surprising and counterintuitive. General wisdom dictates that conditionally valid methods should be less powerful than their marginal counterparts. This is because requiring conditional validity imposes an additional constraint on the Type I error control (conditional validity implies marginal validity, but not vice-versa). Hence, we may expect reasonable conditional tests to be less powerful than reasonable marginal tests, which are derived without this additional constraint in place. In actuality, conditioning (often times on nuisance statistics) allows us to effectively reduce the dimension of the problem. We will see how, in our Gaussian inference on winners problem, conditioning enables us to avoid a curse of dimensionality that standard marginally valid procedures suffer from. Prior to proceeding, we mention that many of these ideas are similar to those discussed in Fithian et al. [2017].

For simplicity, we will now return to having a fixed dimension n, so $X \sim N(\mu, I_n)$ is a random vector in $\mathbb{R}^n$, and W and R (respectively) are the random indices of the largest and second largest entries of X (respectively).

Essentially, all the procedures we have discussed up to this point amount to doing inference on the random estimand μ_W, i.e., the mean of the winning observation. A standard way to do inference on an estimand is to come up with an estimator for it, and then study the deviations of the estimator. In our case, we use the estimator X_W and we study the deviations $X_W - \mu_W$.

In conditional inference, we opt to study the deviations $X_W - \mu_W$ conditional on W. Once we have conditioned on W, our estimand μ_W goes from being random to fixed. Along with our fixed estimand, there $n-1$ fixed nuisance parameters. This is now reminiscent of a classical statistics problem (like in Section 4.4 of Lehmann et al. [1986]). By conditioning further on nuisance statistics, we can remove the influence of these additional nuisance parameters and truly reduce to a one-dimensional inference problem. To be concrete, let $X_{-W} \in \mathbb{R}^{n-1}$ be a vector containing all the entries of X aside from X_W. The conditional $1-\alpha$ quantile of this deviation,

$$q_{1-\alpha}^{cond}(\mu_w, x_{-w}) = \text{Quantile}(X_W - \mu_W | W = w, X_{-W} = x_{-w}),$$

depends on μ only through the value of the winning mean. Therefore, when we invert the conditional $1-\alpha$

high probability region

$$R^{cond}_{1-\alpha}(\mu) = \{x : x_w - \mu_w \leq q^{cond}_{1-\alpha}(\mu_w, x_{-w}) \text{ for } w = \underset{1\leq i\leq n}{\operatorname{argmax}} x_i\}$$

to get an exact $1-\alpha$ confidence region

$$CR^{cond}_{1-\alpha}(X) = \{\mu \in \mathbb{R}^n : X \in R^{cond}_{1-\alpha}(\mu)\} = \{\mu : \mu_W \geq X_W - q^{cond}_{1-\alpha}(\mu_W, x_{-W})\},$$

the resulting confidence region provides non-trivial inferences only for the one-dimensional mean μ_W (which, again, is a non-random parameter once we condition on W). This confidence region is the same as (1.3) from Chapter 1.

In contrast, if we study the deviations $X_W - \mu_W$ marginally, then the $1-\alpha$ quantile

$$q^{marg}_{1-\alpha}(\mu) = \text{Quantile}(1-\alpha, X_W - \mu_W), \tag{5.7}$$

depends on the entire μ vector. Inverting the $1-\alpha$ high probability region

$$R^{marg}_{1-\alpha}(\mu) = \{x : x_w - \mu_w \leq q^{marg}_{1-\alpha}(\mu_w, x_{-w}) \text{ for } w = \underset{1\leq i\leq n}{\operatorname{argmax}} x_i\}$$

to get an exact $1-\alpha$ confidence region

$$CR^{marg}_{1-\alpha}(X) = \{\mu \in \mathbb{R}^n : X \in R^{marg}_{1-\alpha}(\mu)\} = \{\mu : \mu_W \geq X_W - q^{marg}_{1-\alpha}(\mu_W, x_{-W})\}, \tag{5.8}$$

is not similar to what we do in a one-dimensional inferential problem. In fact, due to the high probability region (5.7)'s multi-dimensional dependence on μ, it is unclear if finding the confidence region (5.8) is even computationally feasible. Despite targeting a one-dimensional "estimand" μ_W, this confidence region inadvertently provides non-trivial inferences for the entire mean vector, in part due to the estimand being random.

Even though the confidence region (5.8) may be computationally difficult to find, we can still verify that the inferences it results in are, at best, as good as those from Sidak's procedure. This is the subject of Proposition 3 below.

Proposition 3. *For n-dimensional Gaussian data $X \sim N(\mu, I_n)$, the confidence region $CR^{marg}_{1-\alpha}(X)$ is an exact $1-\alpha$ confidence region for μ, i.e.,*

$$P(\mu \in CR^{marg}_{1-\alpha}) = 1-\alpha.$$

Furthermore, if Sidak's procedure cannot affirm that there is a positive mean, i.e., if

$$1 - \Phi(X_W) > \alpha_n, \qquad W = \underset{1\leq i\leq n}{\operatorname{argmax}} X_i,$$

then the projection of $CR^{marg}_{1-\alpha}$ onto any coordinate axis contains a non-positive value.

The same issue we find with the marginal confidence region (5.7) is present with LSI. Although LSI attempts to apply a smaller simultaneous correction than Sidak's approach, the first step of any reasonable

LSI procedure involves making simultaneously valid confidence intervals for each coordinate of μ. Doing so appears to necessarily expose LSI to a curse of dimensionality that Corollary 1's procedure avoids.

In summary, it appears that targeting random estimands but performing unconditional inference will necessarily expose us to a curse of dimensionality. So how can we deliver powerful inferences while also avoiding the exploding confidence regions we saw in Chapter 1? One idea would be to simply target a non-random estimand. For example, we could try and target $\mu_{(1)}$ instead of μ_W, where $\mu_{(1)} \geq \mu_{(2)} \geq \cdots \geq \mu_{(n)}$ are the sorted entries of μ. If we observe X only after the entries were randomly shuffled, these sorted parameters still identify our data generating distribution. Perhaps it is possible to come up with an estimator for $\mu_{(1)}$ and nuisance statistics that, once conditioned on, get rid of the nuisance parameters $\mu_{(2)}, \ldots, \mu_{(n)}$. We leave this as an area for further exploration. Another approach is hybrid inference, which we discuss next.

5.3 Hybrid inference

Hybrid inference, originally proposed by Andrews et al. [2023], is an inference on winners procedure that attempts to balance the benefits of the simultaneous and conditional approaches. It is a very elegant idea, but it currently only applies to Gaussian data and can be difficult to parse and implement. Using our selective dominance framework, we give a simpler exposition of hybrid inference that enables its application in more general settings, provided that the data is independent. As a bonus, our new procedure is very easy to understand and implement.

Corollary 3 presents our hybrid testing procedure. We give the sketch of a proof and defer a detailed proof to Appendix B.19.3.

Corollary 3 (Hybrid test for the winner)**.** *Suppose that p_i are n independent and selectively dominant p-values for the nulls $H_{0,i}$, and let W be the index of the smallest p-value (with ties broken randomly). Fix some $\beta \leq \alpha$ and define $\beta_n = 1 - (1-\beta)^{1/n}$. Rejecting $H_{0,W}$ when*

$$p_{(1)} \leq \frac{\alpha - \beta}{1 - \beta} p_{(2)} + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_n \tag{5.9}$$

controls Type I error at level α.

Proof sketch. Let B be the event that the smallest p-value comes from a null and is at most β_n. We know from Sidak's procedure that $P(B) \leq \beta$. Hence, on the complementary event B^c, which has probability $\geq 1 - \beta$, it suffices to ensure that we fail to falsely reject $H_{0,W}$ with probability at least $(1-\alpha)/(1-\beta)$. Supposing $H_{0,j}$ is true, imagine testing $H_{0,j}$ using p_j only when B^c happens and $W = j$. This is exactly like selecting p_j to use for inference when it is between β_n and $\max_{i \neq j} p_i$. For this selection, Theorem 5's selective p-value is given by $(p_j - \beta_n)/(\max_{i \neq j} p_i - \beta_n)$. Thus, we can ensure that we fail to reject $H_{0,j}$ when B^c and $W = j$ happen with probability at least $(1-\alpha)/(1-\beta)$ if we fail to reject whenever

$$\frac{p_j - \beta_n}{\max_{i \neq j} p_i - \beta_n} > 1 - \frac{1-\alpha}{1-\beta} \iff p_j > \frac{\alpha - \beta}{1 - \beta} \max_{i \neq j} p_i + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_n.$$

Our hybrid inference procedure fails to reject in this case. □

Written in terms of p-values, it is easy to see how the hybrid approach balances the benefits of the simultaneous and conditional approaches. It will reject both when the smallest p-value is small in absolute terms or when it small relative to the second smallest p-value. When the other p-values provide essentially no evidence against the null (i.e., $p_{(2)} \approx 1$), hybrid rejects when $p_{(1)}$ is at most $(\alpha - \beta)/(1 - \beta)$, a cutoff that has no dependence on the problem dimension n. Thus, hybrid performs at least on par with the conditional procedure run at level $(1 - \alpha)/(1 - \beta)$. On the other hand, even if some other p-value provides as much evidence against the null as the smallest, hybrid still rejects whenever the level β simultaneous approach does. This is because when $p_{(1)} \leq \beta_n$, the hybrid cutoff is a mixture of two things that are at least $p_{(1)}$, so we reject.

The parameter β allows hybrid inference to interpolate between the simultaneous and conditional approaches. When we set $\beta = 0$ then the hybrid cutoff (5.9) becomes $\alpha p_{(2)}$ and we recover the conditional method, and if we set $\beta = \alpha$ it becomes $\alpha_n = 1 - (1 - \alpha)^{1/n}$ and we recover the simultaneous method.

5.3.1 Confidence regions

In parametric settings, we can get hybrid confidence regions for the winning parameter by inverting Corollary 3's test. Again suppose we have independent data $X_i \sim P_{\theta_i}$ from an MLR family P_θ parametrized by $\theta \in \mathbb{R}$, and let $p_i^{\theta_0}$ be the UMP p-value for testing the null $H_{0,i}^{\theta_0} : \theta_i \leq \theta_0$. By inverting Corollary 3's test we get a hybrid LCB for the winning parameter θ_W:

$$\left\{\theta_0 \in \mathbb{R} : p_{(1)}^{\theta_0} > \frac{\alpha - \beta}{1 - \beta} p_{(2)}^{\theta_0} + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_n \right\}. \tag{5.10}$$

We argue in Appendix B.10 that the confidence region (5.10) indeed gives a LCB. We also show in Appendix B.10 how to invert Corollary 3's test to get a CI (rather than an LCB), and that both our CI and LCB match the original construction from Andrews et al. [2023] in the Gaussian case.

5.3.2 Comparison to the union bound

Another way to balance the benefits of the conditional and simultaneous approaches is to apply a union bound. Naively, we can reject the winning null whenever the level β simultaneous approach rejects or the level $\alpha - \beta$ conditional approach rejects, i.e., whenever

$$p_{(1)} \leq \max\{(\alpha - \beta) p_{(2)}, \beta_n\}. \tag{5.11}$$

The union bound harshly switches between the simultaneous and conditional approaches, whereas the hybrid approach smoothly interpolates between them. This is illustrated in Figure 5.4, which compares the LCBs resulting from the hybrid versus union bound approaches in the $n = 20$ dimensional Gaussian problem.

Written in terms of p-values, we easily see that the hybrid approach dominates the union bound approach, which is affirmed by Figure 5.4. Both methods reject when $p_{(1)} \leq \beta_n$. When $p_{(1)} > \beta_n$, it is quick to verify that the hybrid cutoff (5.9) will be strictly larger than the union bound cutoff (5.11), meaning hybrid will reject whenever the union bound does and more [3].

[3] the authors Andrews et al. [2023] only point out that hybrid dominates the level β classical approach, which is weaker than our statement

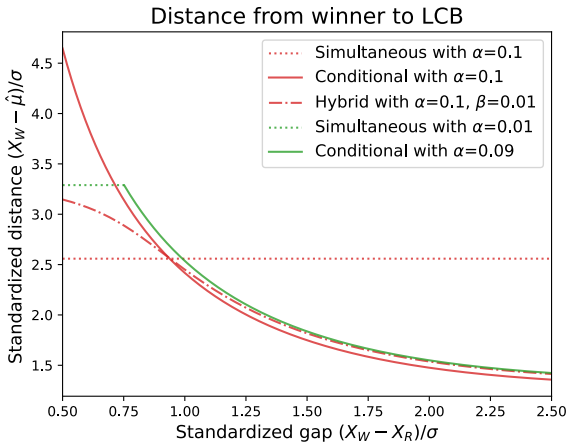


Figure 5.4: For the $n = 20$ dimensional Gaussian problem $X_i \sim N(\mu_i, \sigma^2)$ with largest observation X_W and second largest observation X_R, the standardized distance $(X_W - \hat{\mu})/\sigma$ from X_W to the level $\alpha = 0.1$ hybrid, conditional, and simultaneous LCB $\hat{\mu}$ as a function of the standardized gap $(X_W - X_R)/\sigma$ between the winner and runner-up. For hybrid we take $\beta = 0.01$. The larger of the level $\alpha = 0.09$ conditional LCB and level $\alpha = 0.01$ simultaneous LCB are shown in green (i.e., the union bound LCB).

Practically speaking, however, hybrid inference does not result in much improvement over the union bound, especially as it pertains to making discoveries. This is already somewhat evident in Figure 5.4, where we see that the hybrid LCB, although always larger than the union bound LCB, is still always very close to it. As the variance σ^2 gets large, the absolute difference $\hat{\mu}_{hyb} - \hat{\mu}_{union}$ between the hybrid and union bound LCBs grows with σ, but the relative difference $(\mu_{hyb} - \hat{\mu}_{union})/\sigma$ (in units of standard deviation) remains the same (see Appendix B.10.4). Accordingly, even when the hybrid cutoff (5.9) is larger than that of the union bound (5.11), it is provably not much larger. We detail why in Appendix B.11, where we also run a number of simulations comparing the power of the hybrid and union bound approaches. In our simulations, we are unable to find a setting where the hybrid approach results in a appreciable power gain.

Overall, we suggest viewing hybrid inference as a procedure that squeezes the remaining power out of the union bound approach. As it is not computationally more expensive and our p-value viewpoint makes it equally easy to implement, it is always worth using in place of the union bound.

5.3.3 Applying the closure principal

As was true in the conditional case, treating Corollary 3's test as a global null test and closing it allows us to make more discoveries. As we allow β to range from 0 to α, this closed procedure interpolates between Corollary 2's closed procedure and the Holm-Sidak procedure, which is the closure of Sidak's global null test.

Corollary 4 (Closed hybrid testing for winners)**.** *Suppose that p_i are n independent and selectively dominant p-values for the nulls $H_{0,i}$. Let $H_{0,(j)}$ denote the null corresponding to the jth smallest p-value (with ties broken randomly) and define $p_{(n+1)} = 1$. Fixing some $\beta \leq \alpha$, rejecting the null hypotheses $H_{0,(k)}$ for which*

$$p_{(j)} \leq \frac{\alpha - \beta}{1 - \beta} p_{(j+1)} + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_{n-j+1}$$

for every $j \leq k$ controls FWER error at level α.

This closed procedure is also best understood sequentially. We reject $H_{0,(1)}$ when $p_{(1)} \leq \frac{\alpha-\beta}{1-\beta} p_{(2)} + (1 - \frac{\alpha-\beta}{1-\beta})\beta_n$. Then, if we rejected $H_{0,(1)}$, we reject $H_{0,(2)}$ when $p_{(2)} \leq \frac{\alpha-\beta}{1-\beta} p_{(3)} + (1 - \frac{\alpha-\beta}{1-\beta})\beta_{n-1}$, so on and so forth until we fail to reject.

Chapter 6

Rank verification via selective inference

In this chapter, we use Chapter 4's framework to study a class of rank verification problems. First, we consider some motivating real world examples

Rank verification for large language models: LM Arena [Chiang et al., 2024a] is a platform that currently has over thirty-thousand daily users and ranks the performance of $n > 200$ large language models according to user preference data. Figure 6.1 provides an example of what the leader board may like. Is Gemini-2.5-Pro-Preview-05-06, the model at the top of this example leader board, actually the best model? If so, by how much?

Rank verification across multiple clinical trials: A big pharmaceutical company is interested in investing in a drug, and are examining $n = 5$ smaller pharmaceutical companies that have developed drugs targeting a similar condition. These smaller companies have each completed an early stage clinical trial, and the big company wants to decide which one of these smaller companies to partner with. Is the drug with the highest observed average treatment effect truly the best of the bunch? If so, how much better is it (i.e., how much of a premium should the company be willing to pay)?

Rank verification for machine learning benchmarks: Suppose we evaluate n different pretrained machine models on a benchmark dataset. For example, we may evaluate the five different pretrained ResNet models that PyTorch offers on the ImageNet validation dataset [Deng et al., 2009]. The results of this experiment are shown in Table 6.1. Is ResNet-152, the model with the best performance on the validation dataset, actually the best model? If so, by how much (i.e, is it worth the extra latency and compute cost)?

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	Gemini-2.5-Pro-Preview-05-06	1446	+6/-7	6115	Google	Proprietary
2	3	Gemini-2.5-Flash-Preview-05-20	1418	+10/-10	3892	Google	Proprietary
2	1	o3-2025-04-16	1409	+7/-6	7921	OpenAI	Proprietary
2	2	ChatGPT-4o-latest (2025-03-26)	1405	+6/-5	10280	OpenAI	Proprietary
3	6	Grok-3-Preview-02-24	1399	+5/-3	14840	xAI	Proprietary
4	3	GPT-4.5-Preview	1394	+5/-4	15276	OpenAI	Proprietary
6	6	Gemini-2.5-Flash-Preview-04-17	1387	+7/-8	6938	Google	Proprietary
8	6	DeepSeek-V3-0324	1368	+5/-5	9741	DeepSeek	MIT
8	6	GPT-4.1-2025-04-14	1365	+8/-8	6094	OpenAI	Proprietary
8	13	Hunyuan-Turbos-20250416	1356	+9/-7	5111	Tencent	Proprietary
9	9	DeepSeek-R1	1354	+4/-4	19339	DeepSeek	MIT
10	18	Gemini-2.0-Flash-001	1351	+4/-3	24928	Google	Proprietary

Figure 6.1: A screenshot displaying the top ten entries of the LM Arena leader board. Left-most column denotes the rank of the model on the leader board, which is determined by the model's "Arena Score" in the fourth column. The fifth column provides 95% bootstrap confidence intervals for the population arena score, but with no multiplicity correction, so these confidence intervals cannot be used to safely draw inferences about the population ranking.

Model	Top-1 accuracy
ResNet-18	0.673
ResNet-34	0.713
ResNet-50	0.801
ResNet-101	0.810
ResNet-152	**0.817**

Table 6.1: The top-1 accuracy (i.e., the fraction of examples where the model's top prediction matches the true label) of the five pretrained ResNet models on the $m = 50000$ sample ImageNet Validation set. The number denotes the number of layers, so a larger number indicates a larger model. The model with the highest observed accuracy is bolded.

All of these are examples of rank verification problems, and they can be formalized as follows. We observe a multivariate Gaussian vector $X \sim N(\mu, \Sigma)$ and see that $W = \text{argmax}_{1 \le i \le n} X_i$ is the index of the largest observation. Can we claim that $\mu_W > \max_{j \neq W} \mu_j$, i.e., that the largest observation came from the largest mean? We elaborate on how each of the above scenarios reduces to solving this problem.

Rank verification for large language models: The models on the LM Arena leader board are ranked by "arena scores". These scores are linear transformations of coefficients $\hat{\beta} \in \mathbb{R}^n$ from a Bradley-Terry model [Bradley and Terry, 1952]. Because the model is fit from i.i.d. samples, these coefficients follow a central limit theorem when the number of examples is large, i.e., $\hat{\beta} \dot{\sim} N(\beta, \Sigma)$ for some non-diagonal covariance Σ. Letting W be the index of the model with the largest fitted coefficient, we want to know, for what $\delta \in \mathbb{R}$ can we claim $\beta_W > \min_{j \neq W} \beta_j + \delta$?

Rank verification in multi-arm clinical trials: If there are enough participants in each clinical trial,

then the vector with each drug's average observed treatment effect $X \in \mathbb{R}^n$ obeys a central limit theorem, i.e., $X \dot{\sim} N(\mu, \Sigma)$ with diagonal covariance Σ. We want to know, for what $\delta \in \mathbb{R}$ can we claim that $\mu_W > \max_{j \neq W} \mu_j + \delta$?

Rank verification for machine learning benchmarks: When the benchmark dataset is large enough, the vector $X \in \mathbb{R}^n$ with each models classification accuracy obeys a central limit theorem, i.e., $X \dot{\sim} N(\mu, \Sigma)$ or some non-diagonal covariance Σ. We want to know, for what $\delta \in \mathbb{R}$ can we claim that $\mu_W > \max_{j \neq W} \mu_j + \delta$?

This chapter mainly considers a generalization of our motivating problem. Defining S to be the set containing the indices of the largest $1 \leq K < n$ entries of X, we aim to draw the inference $\min_{i \in S} \mu_i - \max_{j \notin S} \mu_j > \delta$ that the largest K observations came from means that are more than δ larger than the other means. We provide a powerful and computationally tractable error controlling procedure for drawing this inference, and can can recover a solution to our motivating problems by setting $K = 1$. Throughout our discussion, we will assume that the covariance Σ is known. In our applications, Σ is not known but can be estimated from the data.

To start the chapter, we give a high level overview of how we can use Chapter 4's framework to derive our procedure. In the process, we also discuss how our procedure dominates existing simultaneous inference competitors. Then, we give a formal statement of our procedure and characterize its behavior in some special settings. With all of our contributions in place, we contextualize our work by providing a broader literature review on rank verification. After the literature review, we provide a series of real data applications. We conclude by demonstrating how the framework from Chapter 4 can also be applied to rank verification problems in alternative settings, specifically the exponential family setting explored in Hung and Fithian [2019b].

6.1 Overview of approach

Prior to describing how we derive our selective rank verification procedure, we start with a simpler and non-selective problem. Having observed data $X \sim N(\mu, \Sigma)$, when can we verify that μ_i is strictly larger than the other means? This is a classical statistical inference problem. The inferential question has been fixed prior to observing the data, and it has not been selected in a data-dependent manner.

Defining the null hypotheses $H_{0,ij} : \mu_i \leq \mu_j$ for $i \neq j$, our goal is to reject the union null hypothesis $\cup_{j \neq i} H_{0,ij}$. One reasonable approach (from Berger [1982]) is to reject this null when we reject the individual nulls $H_{0,ij}$ for all $j \neq i$. Defining

$$v_{ij} = \sqrt{\text{Var}(X_i - X_j)} = \sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}$$

to be the standard deviation of $X_i - X_j$, Example 3.9.2 from Lehmann et al. [1986] tells us that the UMP test for $H_{0,ij}$ rejects when,

$$X_i > X_j + v_{ij} z_{1-\alpha}$$

Therefore, our earlier reasoning implies that it is safe to reject $\cup_{j \neq i} H_{0,ij}$ whenever

$$X_i > X_j + v_{ij} z_{1-\alpha} \text{ for all } j \neq i.$$

In our problem, our goal is not to reject $\cup_{j \neq i} H_{0,ij}$ for some fixed $i \in [n]$, but rather to reject the data-dependent null hypothesis $\cup_{j \neq W} H_{0,Wj}$, where W is the index of the largest observation. In other words, we only ever try and reject $\cup_{j \neq i} H_{0,ij}$ after observing that $W = i$. A simple fix is simultaneous inference. We run our above procedure for every $1 \leq i \leq n$, but to account for the fact that we are now simultaneously testing n hypotheses, we run it at level α/n rather than level α. (i.e., we apply a Bonferroni correction). Then, after observing that $W = i$, it is safe for use to reject $\cup_{j \neq W} H_{0,Wj}$ so long as we happened to reject $\cup_{j \neq i} H_{0,ij}$ at the reduced level α/n. In practice, this approach amounts to observing that $W = i$ and then running our original test for $\cup_{j \neq i} H_{0,ij}$ at level α/n instead of level α. It enables us to claim that $\mu_W > \max_{j \neq W} \mu_j$ whenever

$$X_W > X_j + v_{Wj} z_{1-\alpha/n} \text{ for all } j \neq W. \tag{6.1}$$

Note that in (6.1), because $z_{1-\alpha/n} = O(\sqrt{\log n})$ grows with the dimension n, this simultaneous procedure requires the winning observation X_W to have a larger and larger gap to the remaining observations as the dimension of the problem grows.

In our selective approach, we recognize that our above UMP test for $H_{0,ij}$ rejects exactly when the selectively dominant p-value $p_{ij} = 1 - \Phi((X_i - X_j)/v_{ij})$ is at most α. Correspondingly, our earlier procedure for rejecting $\cup_{j \neq i} H_{0,ij}$ rejects exactly when $\max_{j \neq i} p_{ij} \leq \alpha$. When try to reject $\cup_{j \neq W} H_{0,Wj}$ via our selective approach, we imagine only selecting the p-values p_{ij} for $j \neq i$ to use for inference when $W = i$. We show that the selection event $W = i$ corresponds to selecting p_{ij} to use for inference when it lives in some interval $[A_{ij}, B_{ij}]$, where A_{ij} and B_{ij} are random quantities that are independent of p_{ij}. To account for selection, we reject not when all of the p_{ij} for $j \neq i$ are at most α, but rather when all of the corrected selective p-values $(p_{ij} - A_{ij})/(B_{ij} - A_{ij})$ are at most α.

The selective procedure above controls the probability of false rejection conditional on W, but A_{ij} and B_{ij} have complicated forms, and it is incredibly difficult to parse. Crucially, by doing some additional analysis, we can show that, when $W = i$, each of the selective p-values $(p_{ij} - A_{ij})/(B_{ij} - A_{ij})$ for $j \neq i$ satisfy the bound

$$\frac{p_{ij} - A_{ij}}{B_{ij} - A_{ij}} \leq 2 \max_{j \neq i} p_{ij}.$$

This means that our selective approach is at least as powerful as the following procedure: first we observe $W = i$, and then we run our original test for $\cup_{j \neq i} H_{0,ij}$ at level $\alpha/2$ rather than at level α. This is much better than the above simultaneous approach, which required us to run this test at level α/n. In contrast to the simultaneous approach, the selective approach enables us to claim that $\mu_W > \max_{j \neq W} \mu_j$ whenever

$$X_W > X_j + v_{Wj} z_{1-\alpha/2} \text{ for all } j \neq W. \tag{6.2}$$

Unlike (6.1), the above rejection region (6.2) has $z_{1-\alpha/2}$ in place of $z_{1-\alpha/n}$, and it does not require the gap from X_W to the remaining observations to be larger for higher dimensional problems. Note that our selective procedure will make at least as many discoveries as the simultaneous approach, and strictly more when $n > 2$.

This is the crux of our contribution. A selective approach that is strictly more powerful than the simultaneous approach because it avoids a curse of dimensionality that the simultaneous approach fails to. In Appendix B.17, we demonstrate how more sophisticated and general simultaneous approaches (including Tukey's honest significant differences test) for this problem also fail to avoid this curse of dimensionality.

As a consequence, these other simultaneous approaches are also strictly dominated by our procedure when $n > 2$.

6.2 Main results

In this subsection, we provide our selective rank verification procedure along with some theoretical results that help us better understand its behavior. We imagine observing n-dimensional data $X \sim N(\mu, \Sigma)$ where Σ is known. Our only restriction on Σ is that we require X_i and X_j to be not perfectly correlated when $i \neq j$. Considering the null hypotheses $H_{ij}^{\delta} : \mu_i - \mu_j \leq \delta$, we derive an error controlling procedure for rejecting the data dependent union null $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$, where S is the set of the largest K observations' indices. Formally, a false rejection happens when we reject $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$ and μ_i is not more than δ larger than μ_j for all $i \in S$ and $j \notin S$. Tying back to our motivation, we can safely draw the inference $\min_{i \in S} \mu_i - \max_{j \notin S} \mu_j > \delta$ when we reject $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$.

Prior to stating our simplified results, we introduce some notation. First, for a pair $i \neq j$, we define

$$D_{ij}^{\delta} = \frac{X_i - X_j - \delta}{v_{ij}}, \quad v_{ij}^2 = \mathrm{Var}(X_i - X_j) = \Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}.$$

Considering another pair of indices $k \neq \ell$, we use $\rho_{ij,k\ell}$ to denote the correlation between D_{ij} and $D_{k\ell}$.

Using this notation, Theorem 9 states our method. Though it may look complicated, it is easy to implement on a computer and we will soon see that its behavior is very interpretable. In our statement of the theorem, we adopt the convention that the maximum of an empty set is $-\infty$.

Theorem 9 (Gaussian rank verification)**.** *If we reject $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$ when*

$$\max_{i \in S, j \notin S} \frac{\Phi\left(D_{ij}^{\delta} - \max\limits_{\substack{k \in S, \ell \notin S: \\ \rho_{ij,k\ell} < 0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0 \right) - \Phi(D_{ij}^{\delta})}{\Phi\left(D_{ij}^{\delta} - \max\limits_{\substack{k \in S, \ell \notin S: \\ \rho_{ij,k\ell} < 0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0 \right) - \Phi\left(D_{ij}^{\delta} - \min\limits_{\substack{k \in S, \ell \notin S: \\ \rho_{ij,k\ell} > 0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0 \right)} \leq \alpha \tag{6.3}$$

then, conditional on S, the probability of making a false rejection is at most α.

Our next result, Theorem 10, helps us make sense of Theorem 9's method. It is easiest to state in terms of the p-values $p_{ij}^{\delta} = 1 - \Phi(D_{ij}^{\delta})$, which are selectively dominant p-values for the nulls H_{ij}^{δ}.

Theorem 10 (Understanding Gaussian rank verification)**.** *For $\delta \geq 0$, the procedure from Theorem 9 is guaranteed to reject $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$ whenever*

$$\max_{i \in S, j \notin S} p_{ij}^{\delta} \leq \alpha/2. \tag{6.4}$$

Remark 1. *The proof of Theorem 10 also informs us that for $\delta < 0$, the procedure from Theorem 9 is guaranteed to reject $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$ whenever $\max_{i \in S, j \notin S} p_{ij}^0 \leq \alpha/2$.*

Essentially, Theorem 10 tells us that, for $\delta \geq 0$, Theorem 9's test will reject $\cup_{i \in S, j \notin S} H_{ij}^{\delta}$ whenever all the two-sided difference-of-means tests comparing pairs of observations inside and outside of the top K reject,

Algorithm 3 Top-1 rank verification via simultaneous inference

Input: data $X \in \mathbb{R}^n$, covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, significance level α

procedure SIMULTANEOUS_LCB(X, Σ, α)
  $\hat{\delta} \leftarrow \infty$
  $W \leftarrow$ index of largest observation in X
  for $j \neq W$ **do**
    $\hat{\delta} \leftarrow \min(X_W - X_j - v_{Wj} z_{1-\alpha/n}, \hat{\delta})$ ▷ $v_{Wj} = \sqrt{\Sigma_{WW} - 2\Sigma_{Wj} + \Sigma_{jj}}$
  return: $\hat{\delta}$

and possibly also in other situations as well. We already discussed why, in the case of $K = 1$ and $\delta = 0$, this is a strict improvement over the simultaneous inference approach discussed in Section 6.1. Appendix B.17 explains why this is a strict improvement over other simultaneous inference approaches for $\delta \geq 0$ and all values of K as well.

Theorem 10 also clarifies a sense in which the error control in Theorem 9's is tight. Fixing any covariance Σ, if we consider the mean vector $\infty = \mu_1 = \ldots \mu_{K-1} > \mu_K = \mu_{K+1} > \mu_{K+2} = \cdots = \mu_n = -\infty$, then $\cup_{i \in S, j \notin S} H_{ij}^0$ is always true and any rejection of $\cup_{i \in S, j \notin S} H_{ij}^0$ is false. In this setting, all of the two-sided difference-of-means tests in Theorem 10 reject with probability one, aside from the one comparing the observations X_K and X_{K+1}. This test will falsely reject with probability exactly α (after all, this is the setting of a vanilla two-sided test), and Theorem 10's procedure will therefore falsely reject with probability exactly α. Theorem 9's test both (1) rejects whenever Theorem 10's condition holds and (2) still maintains error control, so it must falsely reject $\cup_{i \in S, j \notin S} H_{ij}^0$ with probability exactly α as well.

We mention that the time complexity of running Theorem 9's test is $O(K^2(n-K)^2)$, which can be $O(n^4)$ in the worst case (e.g. if $K = n/2$). This will still not be prohibitive for many problems, but if it is, Theorem 10 tells us that we can safely reject $\cup_{i \in S, j \notin S} H_{ij}^\delta$ whenever the condition (6.4) holds, which only takes $O(K(n-K))$ time to check (i.e., $O(n^2)$ in the worst case). Because Theorem 9's test can reject even when this condition does not hold, however, doing so can potentially result in a loss of power (aside from in some special cases that we point out in Section 6.2.2). We mention that when $K = 1$, like in our original motivating problem, Theorem 9's test only takes $O(n^2)$ time to run and Theorem 10's condition only takes $O(n)$ time to check.

6.2.1 Getting lower confidence bounds

By inverting tests $\cup_{i \in S, j \notin S} H_{ij}^\delta$, i.e., finding the the smallest δ for which the test fails to reject, we can get a strict LCB for the difference $\min_{i \in S} \mu_i - \max_{j \notin S} \mu_j$ between the smallest mean whose observation is in S and the largest mean whose observation is not in S. For the case that $K = 1$, we can easily generalize the simultaneous inference test we described in Section 6.1 to apply for any $\delta \in \mathbb{R}$ and then invert it. This results in the LCB given by Algorithm 3. We will refer to Algorithm 3's LCB as the simultaneous LCB.

As an alternative, we can also get an LCB that is valid conditionally on S by inverting Theorem 9's test. We argue carefully that inverting Theorem 9's test results in a LCB in Appendix B.18. In practice, this LCB must be found via a binary search, as is described in Algorithm 4. We will refer to Algorithm 4's LCB as the conditional LCB.

Theorem 10 implies an important fact about the simultaneous and conditional LCBs: whenever the conditional LCB is non-positive, it is guaranteed to be larger than the simultaneous LCB. This means that

Algorithm 4 Top-K rank verification via conditional inference

Input: data $X \in \mathbb{R}^n$, covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, number of winners K, significance level α

procedure CONDITIONAL_LCB(X, Σ, K, α)

$S \leftarrow$ indices of largest K observations in X

$\hat{\delta} \leftarrow$ via binary search, find $\hat{\delta}$ such that

$$\max_{i \in S, j \notin S} \frac{\Phi\left(D_{ij}^{\hat{\delta}} - \max\limits_{\substack{k \in S, \ell \notin S: \\ \rho_{ij,k\ell} < 0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0\right) - \Phi(D_{ij}^{\hat{\delta}})}{\Phi\left(D_{ij}^{\hat{\delta}} - \max\limits_{\substack{k \in S, \ell \notin S: \\ \rho_{ij,k\ell} < 0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0\right) - \Phi\left(D_{ij}^{\hat{\delta}} - \min\limits_{\substack{k \in S, \ell \notin S: \\ \rho_{ij,k\ell} > 0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0\right)} = \alpha$$

▷ $v_{ij} = \sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}$

▷ $D_{ij}^{\eta} = (X_i - X_j - \eta)/v_{ij}$

▷ $\rho_{ij,k\ell} = \frac{\Sigma_{ik} - \Sigma_{i\ell} - \Sigma_{jk} + \Sigma_{j\ell}}{v_{ij} v_{k\ell}}$

return: $\hat{\delta}$

whenever the conditional LCB enables us to make a discovery, i.e., it allows us to say with confidence that the K largest observations come from means that are strictly larger than the rest, it is guaranteed to certify a bigger gap between the means of the K largest observations and the remaining means than the simultaneous LCB. This is not the case, however, when the conditional LCB returns a negative value (i.e., no discovery is made). In this case the conditional LCB can be much lower than the simultaneous LCB. In fact, if the top K and bottom $n - K$ observations get too close, the conditional LCB can diverge to $-\infty$, like we saw for the inference on winners problem in Chapter 1. We will observe this happen later when we discuss our real data applications.

This leaves users with a trade-off. If they want to ensure that their LCB will never be too negative, they should use the simultaneous approach. If they want to maximize their chance of making a discovery, however, and also be able to make the best possible claim about the strength of said discovery once it has been made, they should use the conditional approach. We suspect that most users care significantly more about the latter, especially since the exploding inferences of the conditional approach are relatively rare (we will see this later). One can always try and get the best of both worlds by taking a hybrid or union bound approach as in Section 5.3. Doing so will prevent the exploding intervals, but also lessen the user's ability to make discoveries. We suspect that users are more interested in making discoveries than anything else.

We re-iterate that, despite the fact that the conditional LCB can diverge to $-\infty$, it is provably guaranteed to make more discoveries in each individual problem instance (not just on average). This is more than enough to justify its use. It is also in stark contrast the to the inference on winners problem we discussed in Chapter 5, where conditional inference (Corollary 1) had no such advantage over simultaneous inference and still suffered from the exploding inference problem.

Finally, we mention that Theorem 10 and Remark 1 in tandem imply the validity of a much simpler way to make an LCB for $\min_{i \in S} \mu_i - \max_{j \notin S} \mu_j$ that is conditionally valid on S: for all $i \in S$ and $j \notin S$, compute the lower bounds $X_i - X_j - v_{ij} z_{1-\alpha/2}$ of the standard two-sided difference-of-means confidence intervals for $\mu_i - \mu_j$, and take the lowest one. If it is non-negative, return it as the LCB, otherwise return $-\infty$.

We describe this simpler procedure, which mimics the one suggested in Hung and Fithian [2019b], in

Algorithm 5 More simple top-K rank verification via conditional inference

Input: data $X \in \mathbb{R}^n$, covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$, number of winners K, significance level α

procedure SIMPLE_CONDITIONAL_LCB(X, Σ, K, α)
  $\hat{\delta} \leftarrow \infty$
  $S \leftarrow$ indices of largest K observations in X
  for $i \in S$ **do**
    for $j \notin S$ **do**
      $\hat{\delta} \leftarrow \min(X_i - X_j - v_{ij} z_{1-\alpha/2}, \hat{\delta})$ ▷ $v_{ij} = \sqrt{\Sigma_{ii} - 2\Sigma_{ij} + \Sigma_{jj}}$
  if $\hat{\delta} \geq 0$ **then**
    return: $\hat{\delta}$
  else
    return: $-\infty$

Algorithm 5. This LCB will always be less than or equal the one that results from inverting Theorem 9's test (i.e., it is worse), but it is faster to compute and easier to understand. Like was the case for Algorithm 4's LCB, whenever Algorithm 5's LCB is non-negative it will be larger than Algorithm 3's LCB.

6.2.2 Some special cases

Briefly we discuss some special cases where we can even more precisely characterize the behavior of Theorem 9. Our focus will always will always be for the $\delta = 0$ problem, where we are simply trying to verify that the top K observations came from means that are strictly larger than the rest.

Our first special case is when Σ is diagonal, i.e., the observations are independent. The below result tells us that, in this case, Theorem 10 gives an exact characterization of Theorem 9's test.

Corollary 5 (Independence)**.** *When Σ is diagonal, the procedure from Theorem 9 rejects $\cup_{i \in S, j \notin S} H_{ij}^0$ exactly when the condition (6.4) from Theorem 10 is satisfied. More specifically, if I and J are the indices of the observations inside and outside of the top K with the smallest standardized difference, so*

$$\frac{X_I - X_J}{v_{IJ}} = \min_{i \in S, j \notin S} \frac{X_i - X_j}{v_{ij}},$$

then the procedure from Theorem 9 rejects $\cup_{i \in S, j \notin S} H_{ij}^0$ exactly when

$$X_I > X_J + v_{IJ} z_{1-\alpha/2}.$$

Our next result explains that the same is true for equicorrelated data. The work of Hung and Fithian [2019b] implies the below result for the case that $K = 1$, but we generalize it to any K.

Corollary 6 (Equicorrelation)**.** *If Σ is an equicorrelation matrix, i.e., for some $\rho \in (-1, 1)$*

$$\Sigma_{ij} = \begin{cases} \sigma^2 & \text{if } i = j \\ \rho\sigma^2 & \text{if } i \neq j, \end{cases}$$

then the procedure from Theorem 9 rejects $\cup_{i \in S, j \notin S} H_{ij}^0$ exactly when the condition (6.4) from Theorem 10 is satisfied. More specifically, if I and J are the entries of the Kth and $(K+1)$st largest observations

respectively, then the procedure from Theorem 9 rejects $\cup_{i\in S, j\notin S} H_{ij}^0$ *exactly when*

$$X_I > X_J + v_{IJ} z_{1-\alpha/2}.$$

We also provide an analogous result for autocorrelated data when $K = 1$ or $K = n - 1$, provided that the amount of autocorrelation is not too large.

Corollary 7 (Autocorrelation)**.** *Suppose that* Σ *is an autocorrelation matrix with small enough autocorrelation. Specifically, the entries of* Σ *are given by* $\Sigma_{ij} = \sigma\rho^{|i-j|}$*, with* $\rho \in [\frac{1-\sqrt{5}}{2}, \frac{1}{2}]$*. For* $K = 1$ *or* $K = n - 1$*, the procedure from Theorem 9 rejects* $\cup_{i\in S, j\notin S} H_{ij}^0$ *exactly when the condition (6.4) from Theorem 10 is satisfied. More specifically, if* I *and* J *are the indices of the observations inside and outside of the top* K *with the smallest standardized difference, so*

$$\frac{X_I - X_J}{v_{IJ}} = \min_{i\in S, j\notin S} \frac{X_i - X_j}{v_{ij}},$$

then the procedure from Theorem 9 rejects $\cup_{i\in S, j\notin S} H_{ij}^0$ *exactly when*

$$X_I > X_J + v_{IJ} z_{1-\alpha/2}.$$

Finally we characterize how our method behaves if we set $K = 1$ and apply it to multinomial data. Particularly, we imagine using a multivariate Gaussian distribution to approximate the joint distribution of t multinomial trials $Y \sim \text{Multinomial}(t, \pi_1, \ldots, \pi_n)$, i.e., we define $\hat{\pi} = Y/t$ and apply our method to $\hat{\pi} \dot{\sim} N(\pi, \hat{\pi}/t - \hat{\pi}\hat{\pi}^\top/t)$.

Corollary 8 (Multinomial)**.** *Suppose* $K = 1$ *and we apply Theorem 9's procedure to data* $X = \hat{\pi}$ *with covariance* $\Sigma = \hat{\pi}/t - \hat{\pi}\hat{\pi}^\top/t$*. Defining* W *and* R *to be the indices of the largest and second largest observations in* $\hat{\pi}$ *respectively, Theorem 9's procedure will reject* $\cup_{i\in S, j\notin S} H_{ij}^0$ *exactly when*

$$X_W > X_R + v_{WR} z_{1-\alpha/2}.$$

More generally, we learn in the proof of Corollary 8 that when the covariance is $\Sigma = \hat{\pi}/t - \hat{\pi}\hat{\pi}^T/t$, the maximum v_{Wj} for $j \neq W$ is attained by v_{WR}. It is then straightforward to show that, for this multinomial problem, Algorithm 5 will return $\hat{\pi}_W - \hat{\pi}_R - v_{WR} z_{1-\alpha/2}$ as an LCB for $\pi_W - \max_{j\neq W} \pi_j$. This result along with Corollary 8 can be viewed as a recovery of the main application in Hung and Fithian [2019b].

6.3 Related work

Both Gutmann and Maymin [1987] and Stefansson et al. [1988] study our problem in the case that $K = 1$, $\delta = 0$, and the data $X_i \sim N(\mu_i, \sigma^2)$ are independent Gaussian samples with common variance. They show that drawing the inference $\min_{i\in S} \mu_i - \max_{j\notin S} \mu_j > \delta$ whenever the two-sided difference-of-means test comparing the largest and second largest observation rejects is an error controlling procedure. Our work provides a complete generalization of their result in the case of multivariate Gaussian data, allowing for any K, any covariance structure, and any $\delta \in \mathbb{R}$. Work prior to Gutmann and Maymin [1987] studied related rank verification problems in similar settings [Bechhofer, 1954, Bofinger, 1983, 1985b, Desu, 1970, Fabian,

1962, Gupta, 1965, 1956, Hsu, 1981b, 1984], but used simultaneous inference techniques and failed to avoid a multiplicity correction as Gutmann and Maymin [1987] did. Follow-up work to Gutmann and Maymin [1987] includes methods that avoid multiplicity corrections for other rank verification problems [Gutmann, 1987, Maymin and Gutmann, 1992], but in similarly restricted settings. Also, Cheng and Panchapakesan [2009] extend Gutmann and Maymin [1987]'s result to the case of independent samples from a scale family with a monotone likelihood ratio (Gutmann and Maymin [1987] themselves handle the case of independent samples from a location family with a monotone likelihood ratio).

To prove their result Gutmann and Maymin [1987], condition on the index of the winning observation. This is a similar strategy to that of modern post-selection inference, a field initiated by the seminal work Lee et al. [2016a]. By leveraging modern selective techniques, Hung and Fithian [2019a] generalize Gutmann and Maymin [1987]'s procedure to apply for exponential families with Schur concave carrier measures (for $K = 1$ and any $\delta \in \mathbb{R}$). Schur concavity requires the carrier measure to be symmetric, so for the multivariate Gaussian case Hung and Fithian [2019a] only generalize Gutmann and Maymin [1987]'s procedure from the independent to the equicorrelated setting (our Corollary 5 and Corollary 6 cover both of these cases). The main focus of Hung and Fithian [2019a] is rank verification for multionmial data. If there are enough samples, then multinomial data can be approximated as correlated multivariate Gaussian data via the central limit theorem, and our Corollary 8 recovers Hung and Fithian [2019a]'s result. Hung and Fithian [2019a] also consider rank verification for the Bradley-Terry model, but their method (1) only scales to games with roughly $n = 40$ players and (2) requires each player to play the other players the same number of times (both conditions are violated in our LM Arena example). Work that is concurrent with and independent from ours considers multivariate Gaussian rank verification in the independent and unequal variance case [Goldwasser et al., 2025]. Indeed, once restricted to this case, Theorem 9's method matches that of Goldwasser et al. [2025]. Goldwasser et al. [2025], however, do not show that the resulting method amounts to running all the two-sided difference-of-means tests comparing all the observations inside the top K with all the observations outside of the top K (i.e., they have no analog of our Theorem 10 or Corollary 5). As a consequence, they do not formally characterize the method's behavior or power.

6.4 Applications

Having developed and contextualized our method, we provide a series of real data examples that demonstrate its usefulness.

6.4.1 Ranking large language models from user preference data

Although we do not have access to the full LM Arena data, we can use a smaller dataset that LM Arena has made publicly available [Chiang et al., 2024b]. This dataset contains $m = 33000$ i.i.d examples of human preference data, where users provide two models with the same prompt and then mark which of the two models performed better. In total, the dataset has $n = 20$ models.

Following Chiang et al. [2024a], we use this data to fit a weighted Bradley-Terry model [Bradley and Terry, 1952], where each example is inversely weighted by the probability of those two models being shown to the user. Because these probabilities are not given to us, we estimate them from the data. Table 6.2 reports the arena scores, which are found by multiplying each model's Bradley-Terry coefficient by 400 and

Model name	Arena score
gpt-4	1535
claude-v1	1439
claude-instant-v1	1351
gpt-3.5-turbo	1301
guanaco-33b	1160
vicuna-13b	1132
palm-2	1104
wizardlm-13b	1093
vicuna-7b	1051
koala-13b	1019
mpt-7b-chat	896
gpt4all-13b-snoozy	889
RWKV-4-Raven-14B	873
alpaca-13b	841
oasst-pythia-12b	828
chatglm-6b	790
fastchat-t5-3b	776
stablelm-tuned-alpha-7b	691
llama-13b	657
dolly-v2-12b	656

Table 6.2: Arena scores of the $n = 20$ models from the publicly available LM Arena dataset [Chiang et al., 2024b].

then adding 1000.

Table 6.2 shows that gpt-4 is the best performing model on Chiang et al. [2024b]'s dataset. Can we affirm that it is indeed the best performing model in the population (i.e., if we had infinite samples)? Because the number of samples is so large, the central limit theorem implies our estimated coefficients $\hat{\beta} \dot{\sim} N(\beta, \Sigma)$ approximately follow a normal distribution centered around the population coefficients β. The covariance Σ is unknown, but we can estimate it via the bootstrap (we use $b = 10000$ bootstrap iterates). This enables to apply both Algorithm 3 and Algorithm 4 to get LCBs for the difference between gpt-4 and the remaining models' population arena scores. Algorithm 3's simultaneous LCB enables us to say with 95% confidence that

$$\text{ArenaScore}_{\text{gpt}-4} > \max_{\text{model} \neq \text{gpt}-4} \text{ArenaScore}_{\text{model}} - 21. \tag{6.5}$$

That is, we can affirm that gpt-4's population arena score is not too much worse than any of the other models. In contrast, Algorithm 4's conditional approach enables us to affirm with confidence that gpt-4 is indeed be the best model in the population. In particular, it allows to say with 95% confidence that

$$\text{ArenaScore}_{\text{gpt}-4} > \max_{\text{model} \neq \text{gpt}-4} \text{ArenaScore}_{\text{model}} + 19. \tag{6.6}$$

Despite having such a large sample size, we still air on the safe side and run some simulations to affirm that our conditional LCB does not undercover in this sort of real data setting. In our simulations, we use our fitted Bradley-Terry coefficients $\hat{\beta}$ as the population coefficients. For each simulated trial, we sample $m = 33000$ pairs of competing models (i, j) with $i < j$ from our original dataset (with replacement), and mark that the ith model won with probability $f(e^{\hat{\beta}_i - \hat{\beta}_j})$, where $f(\cdot)$ is the sigmoid function. To save time, we only run $B = 1000$ simulated trials and only use $b = 100$ bootstrap iterates in each of these trials. Our 95%

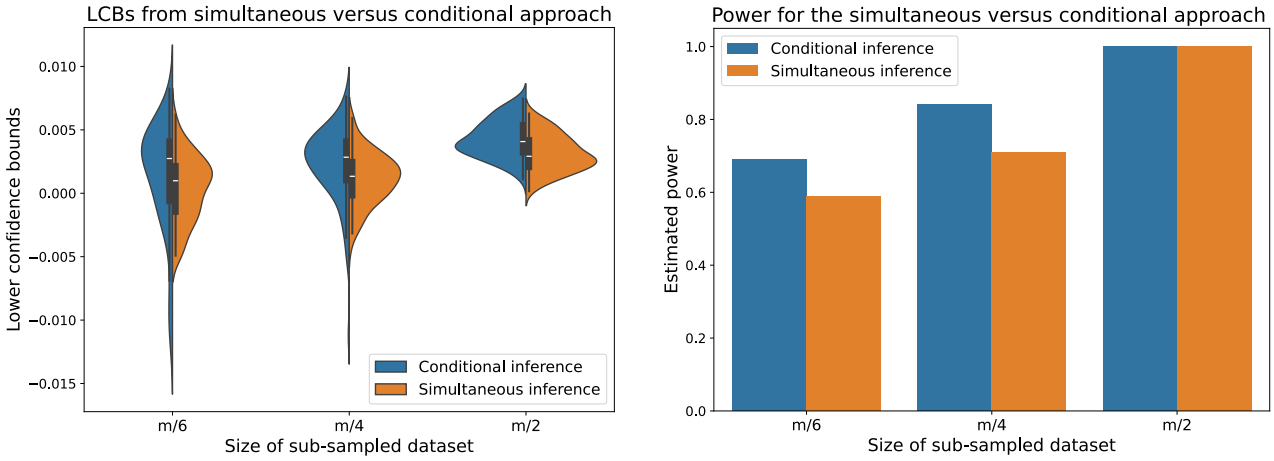


Figure 6.2: For each sub-sample size on the x-axis, the five ResNet models from Table 6.1 are evaluated on $B = 100$ different random sub-samples of the size $m = 50000$ ImageNet validation set. The plots give the distribution of the resulting conditional LCBs from Algorithm 4 and simultaneous LCBs from Algorithm 3 (left), along with the proportion of times these LCBs are non-negative (right).

conditional LCB has an empirical miscoverage rate of only 0.001. Note that this overcoverage should not be surprising. Even though our conditional procedure is strictly more powerful than simultaneous inference, it still involves taking the maximum of $n = 20$ selective p-values, so it can certainly be conservative. This may especially be the case when many observations are in close competition to be the largest, as is the case in this example.

6.4.2 Ranking ResNet models

We revisit the problem of determining whether the ResNet model with the best performance on the validation dataset is actually the best performing model in the population (see Table 6.1 for a refresher on the five different models and their performances). We will denote the validation top-1 accuracy (i.e., the fraction of validation examples where the model's top prediction matches the true label) of the five models as $\hat{q} \in \mathbb{R}^5$, with $\hat{q}_i$ being the validation accuracy of the ith largest model (so $\hat{q}_5$ is the validation accuracy of ResNet-152). Since $\hat{q}$ is a sample mean of $m = 50000$ five-dimensional i.i.d observations, it should follow roughly a normal distribution by the central limit theorem, i.e., $\hat{q} \dot{\sim} N(q, \Sigma)$ where $q \in \mathbb{R}^5$ contains the unknown population accuracies of the five models. Having observed that $\hat{q}_5$ is the largest entry of $\hat{q}$, we want to verify whether or not q_5 is the largest entry of q. Although the covariance Σ of $\hat{q}$ is unknown, $\hat{q}$ is a sample mean, so Σ is easy to estimate from the data. We point out that, because models often get the same easy examples correct and same difficult examples wrong, the covariance Σ is dense and highly non-trivial. This dense covariance precludes us from using the methods of Gutmann and Maymin [1987], Hung and Fithian [2019b], and Stefansson et al. [1988].

We compare four approaches for verifying that q_5 is the largest entry of q:

1. **Incorrect approach:** Often times, the validation set performance of models is displayed exactly as we have shown it in Table 6.1, and it is implicitly assumed that the model with the best validation performance would perform the best in the population. For example, this is the case in the original

ResNet paper, which contains a table exactly like the one we have presented [He et al., 2016]. This approach is unprincipled, and can result in false claims with high probability!

2. **Naive approach:** A naive, but statistically valid approach, is to report standard error bars for each model's validation accuracy. These sorts of error bars are what are now typically presented in contemporary machine learning literature. Using these error bars, one can apply a Bonferonni correction and get simultaneously valid confidence intervals for each model's population accuracy. For example, we can say with 95% confidence that all of the following statements simultaneously hold:

$$q_1 \in (0.667, 679), \quad q_2 \in (0.707, 0.719), \quad q_3 \in (0.796, 0.806), \quad q_4 \in (0.805, 0.815), \quad q_5 \in (0.812, 0.821)$$

Because the intervals for q_4 and q_5 overlap, however, this approach does not allow us to confidently say that ResNet-152 is the best model.

3. **Simultaneous approach:** The simultaneous approach is to use the LCB given by Algorithm 3. Unlike the previous naive approach (which is technically just a different way of doing simultaneous inference), this non-naive simultaneous approach does inference on the difference between the accuracy of two models. Because the model predictions are highly correlated (models tend to be correct and incorrect on the same examples), our estimates of the difference in performance between two models tend to have lower standard errors than the estimates of each model's individual performance. Algorithm 3's LCB allows us to state with 95% confidence that

$$q_5 > \max_{1 \leq i \leq 4} q_i + 0.0044$$

i.e., we can affirm that ResNet-152 is the best model in the population.

4. **Selective approach:** By inverting Theorem 9's test, we can state with 95% confidence that

$$q_5 > \max_{1 \leq i \leq 4} q_i + 0.0053.$$

The selective approach both (1) affirms that ResNet-152 is the best model in the population, and (2) affirms a bigger gap in population performance between ResNet-152 and the remaining models than the simultaneous approach. We re-affirm that, according to our theory, whenever the selective approach is able to affirm a positive gap, it will affirm a bigger gap than the simultaneous approach.

In this problem, we were lucky to have a large enough validation set that both the simultaneous and conditional LCBs were able to verify ResNet-152 as the best model. We are not always fortunate enough, however, to have benchmark datasets with $m = 50000$ examples, and it is worth studying how these two methods compare when the validation set size is smaller.

Figure 6.2 displays what happens when we run both Algorithm 3 and Algorithm 4 after randomly subsetting the ImageNet validation set to different sizes. The left panel shows that the conditional method tends to deliver larger LCBs than the simultaneous approach, and the right plot demonstrates what we already provably knew: the conditional method makes strictly more discoveries than the simultaneous one. We do notice in the left panel that the conditional method does sometimes (very rarely) give very low LCBs, particularly when the size of the validation set is small (in low sample settings, the winner is more likely to

be less distinguishable from the other observations). We see, however, that it is quite rare to observe one of these exploding conditional LCBs, and they are more than made up for by the fact that the conditional method makes strictly more discoveries.

Despite the sample size being so large, we air on the safe side and run some simulations to affirm that our conditional LCB does not undercover in this sort of real data setting. So that our simulated setting mimics our real data example, we treat $\hat{q}$ as the population accuracy of each "pretrained model" in our simulations. To get simulated examples that respect the correlation structure of our real data examples, we use the following three step procedure to generate each new five-dimensional validation sample.

1. Generate $Y \sim N(0, \hat{\Omega})$, where $\hat{\Omega}$ is our estimate $\hat{\Sigma}$ of the covariance of $\hat{q}$, but standardized to be a correlation matrix.

2. Transform Y into five uniformly disturbed random variables via the transformation $U_i = \Phi(Y_i)$.

3. For this sample, mark the ith "pretrained model" as correct when $U_i \leq \hat{q}_i$ and incorrect otherwise.

Over $B = 10000$ simulated trials, where in each simulated trial we generate $m = 50000$ new validation examples, our 95% conditional LCB has an empirical miscoverage rate of 0.048.

6.4.3 Ranking clutch basketball players

Player	Clutch field goal %	Clutch field goal attempts
Michael Jordan	0.462	676
Kobe Bryant	0.444	2871
LeBron James	0.523	3525
Stephen Curry	**0.558**	**1685**
Kevin Durant	0.504	2224
Kawhi Leonard	0.510	921

Table 6.3: According to `inpredictable.com` [Beuoy, 2025], the observed clutch field goal percentage and number of clutch field goals taken for LeBron James, Kobe Bryant, Michael Jordan, Stephen Curry, Kevin Durant, and Kawhi Leonard. Player with the highest clutch field goal percentage is bolded.

In basketball, a classic question is determining which player in a group is the most clutch, i.e., when it matters, who is most likely to score a basket?

Whether or not a shot should be classified as "clutch" is up to interpretation, and many definitions exist. We will leave this classification up to the popular sports analytics site `inpredictable.com`, which broadly describes clutch shots as "shots taken that have an elevated impact on win probability" [Beuoy, 2025]. Whether or not a shot fits into this category is determined internally by the site, and then final clutch field goal percentages and number of clutch shots taken are reported for each player. We provide these numbers for six well-known players in Table 6.3. We mention that Michael Jordan's numbers are taken only from 1996 onward because this is the year that the National Basketball Association started recording play-by-play data. Hence, any claims we make about Michael Jordan only apply to the Michael Jordan that played from 1996-2003, and are not representative of his whole career.

In our analysis, we make the simplifying assumption that the outcomes of all the clutch shots for the all the players are independent. This assumption may be violated if some clutch shots were taken in the same game, whether by the same or different players. We expect the extent to which this assumption is violated, however, to be small.

More specifically, since LeBron James' has taken $n_{\text{LeBron}} = 3525$ clutch shots, we assume that the number of clutch shots LeBron has made will follow a $\text{Bin}(3525, q_{\text{LeBron}})$ distribution, where LeBron's population clutch field goal percentage q_{LeBron} is unknown. Because the number of clutch shots each player has taken is different, Hung and Fithian [2019b]'s rank verification procedure cannot be applied to this problem. To proceed, we instead note that n_{LeBron} is large enough for the central limit theorem to apply, i.e., $\hat{q}_{\text{LeBron}} \dot{\sim} N(q_{\text{LeBron}}, \sigma^2)$ where $\sigma^2 = q_{\text{LeBron}}(1-q_{\text{LeBron}})/n_{\text{LeBron}}$. More generally, the vector $\hat{q} \in \mathbb{R}^6$ of empirical clutch field goal percentages for each player approximately has a multivariate normal distribution $\hat{q} \dot{\sim} N(q, \Sigma)$, where $q \in \mathbb{R}^6$ encodes the players' population clutch field goal percentages, and Σ is a diagonal covariance matrix. This enables us to apply our rank verification method to the problem. Per usual, Σ is unknown. Since the $\hat{q}_i$ are sample means, however, Σ can be easily estimated from the data.

The first question we try and answer is the most popular one. Out of Lebron James, Kobe Bryant, and Michael Jordan, who is the most clutch? Upon observing that James has the highest empirical clutch field goal percentage of the three (see Table 6.3), can we affirm that his population clutch field goal percentage is actually highest? Algorithm 3's simultaneous LCB allows us to say with 95% confidence that

$$q_{\text{LeBron}} > \max\{q_{\text{Michael}}, q_{\text{Kobe}}\} + 0.016$$

In contrast, Algorithm 4's conditional LCB allows us to say with 95% confidence that

$$q_{\text{LeBron}} > \max\{q_{\text{Michael}}, q_{\text{Kobe}}\} + 0.025$$

Via both methods, we are able to say with confidence that James is a more clutch player than Jordan and Bryant. As predicted by our theory, however, the conditional approach is able to certify a larger gap between James' and the others' abilities. Although the difference between the gaps, which is 0.9%, may seem small, a one-percent difference in field goal percentage is quite sizable in basketball, especially when one considers that star players take thousands of clutch shots over their careers.

What if we had wanted to add some more modern players into the mix? Along with Lebron James, Kobe Bryant, and Michael Jordan, we now also consider Stephen Curry, Kevin Durant, and Kahwi Leonard, three other players who have won Finals MVPs and are known to perform well in the clutch. Out of the bunch Curry has the highest empirical clutch field goal percentage. Can we verify that he is actually the most clutch? Algorithm 3's simultaneous LCB allows us to say with 95% confidence that

$$q_{\text{Stephen}} > \max\{q_{\text{LeBron}}, q_{\text{Michael}}, q_{\text{Kobe}}, q_{\text{Kevin}}, q_{\text{Kawhi}}\} - 0.0009.$$

That is, we can affirm that Curry is not that much less clutch than any of the others, but we cannot affirm that he is the most clutch. In contrast, Algorithm 4's conditional LCB allows us to say with 95% confidence that

$$q_{\text{Stephen}} > \max\{q_{\text{LeBron}}, q_{\text{Michael}}, q_{\text{Kobe}}, q_{\text{Kevin}}, q_{\text{Kawhi}}\} + 0.0084,$$

and therefore enables us to claim with confidence that Curry is indeed more clutch than the rest.

Again, despite having a large number of clutch shots per player, we run simulations to make sure that our conditional LCB does not undercover in this sort of real data setting. In our simulated trials, we treat each player's observed clutch field goal percentage $\hat{q}_{\text{player}}$ as their population clutch field goal percentage, and sample a binomial observation $\text{Bin}(n_{\text{player}}, \hat{q}_{\text{player}})$ for each player. Using these observations, we apply our procedure as described above. When we include only James, Bryant and Jordan in our simulations, our 95% selective LCB has an empirical miscoverage rate of 0.033 over $B = 10000$ trials. When we include all six players in our simulations, our 95% selective LCB has an empirical miscoverage rate of 0.024 over $B = 10000$ trials.

6.5 Rank verification in exponential families

Briefly, we revisit the rank verification problem of Hung and Fithian [2019b], where we observe data $X \in \mathbb{R}^n$ from a natural multi-parameter exponential family P_θ with density

$$g_\theta(x) = \exp(\theta_1 x_1 + \cdots + \theta_n x_n - \psi(\theta))g(x), \tag{6.7}$$

with respect to some base measure [1]. Our goal in revisiting this problem is two-fold. First, it illustrates how Hung and Fithian [2019b]'s rank verification procedure fits nicely in our framework. Second, we show that Hung and Fithian [2019b] do not appropriately handle cases where there can be ties for the winner In contrast, our selective dominance framework naturally handles these cases properly.

Like Hung and Fithian [2019b], we want to verify that θ_W is strictly larger than the remaining θ_j, where W is the index of the largest X_i. In case of ties, we follow Hung and Fithian [2019b] and set W to randomly be one of the winning indices. Since $E_\theta[X_i] = \theta_i$, the winning parameter θ_W may reasonably be the largest of the θ_i.

Considering the nulls $H_{0,ij} : \theta_i \leq \theta_j$, we want to reject the data-dependent null $\cup_{j \neq W} H_{0,Wj}$ and affirm that θ_W is strictly larger than every other parameter [2]. Considering $j \neq W$, our strategy will be to come up with a valid test for $H_{0,Wj}$, and then reject $\cup_{j \neq W} H_{0,Wj}$ whenever we reject $H_{0,Wj}$ for every $j \neq W$. Fixing $i \neq j$, we start by constructing the UMPU p-value p_{ij} for testing $H_{0,ij} : \theta_i \leq \theta_j$. Ultimately, we will only use this p-value to test $H_{0,ij}$ when i is selected as our winning index, and we will correspondingly adjust the p-value to account for this selection.

Defining the transformed sufficient statistics $Y \in \mathbb{R}^n$ by

$$Y_i = \frac{X_i - X_j}{2}, \qquad Y_j = \frac{X_i + X_j}{2}, \qquad Y_\ell = X_\ell \text{ for } \ell \neq i, j, \tag{6.8}$$

the random vector Y has an exponential family density given by

$$\tilde{g}_\theta(y) = \exp\left((\theta_i - \theta_j)y_i + (\theta_i + \theta_j)y_j + \sum_{\ell \neq i,j} \theta_\ell y_\ell - \psi(\theta)\right) \tilde{g}(y) \tag{6.9}$$

[1] We assume that $T_i(x) = x_i$ so that our discussion more closely mirrors Hung and Fithian [2019b], although we do not need to. Hung and Fithian [2019b] also assume that $g(x)$ is Schur concave for other purposes, but we leave this assumption out.

[2] If we performed the same analysis for the nulls $H^\delta_{0,ij} : \theta_i - \theta_j \leq \delta$ then we could verify that θ_W is more than δ larger than any other θ_j. Inverting these tests would result in a LCB for the difference $\theta_W - \max_{j \neq W} \theta_j$ between the winning and next largest parameter.

with respect to some other base measure. It is then well-established (see Appendix B.15.1) that if we denote the conditional left-continuous survival function of Y_i and its right-hand limit as

$$G_{ij}(y_i|y_{-i}) = P_{\theta_i=\theta_j}(Y_i \geq y_i|Y_{-i} = y_{-i}) \qquad G^+_{ij}(y_i|y_{-i}) = \lim_{u \downarrow y_i} G_{ij}(u|y_{-i}), \tag{6.10}$$

then the UMPU p-value p_{ij} for testing $H_{0,ij} : \theta_i \leq \theta_j$ is given by

$$p_{ij} = G^+_{ij}(Y_i|Y_{-i}) + U_{ij,aux}(G_{ij}(Y_i|Y_{-i}) - G^+_{ij}(Y_i|Y_{-i})), \tag{6.11}$$

where $U_{ij,aux}$ are Unif($[0, 1]$) random variables that are independent from each other and the data. By Example 4, the p-value p_{ij} is selectively dominant given Y_{-i}.

Crucially, we can tell if X_i is a winner by examining the p-value p_{ij}. It is straightforward to confirm that X_i is the sole winner exactly when $Y_i > \max_{k\neq i} Y_k - Y_j$. Equivalently, this happens when p_{ij} is strictly smaller than

$$q^+_{ij}(Y_{-i}) = G^+_{ij}(\max_{k\neq i} Y_k - Y_j|Y_{-i}). \tag{6.12}$$

Likewise, one can confirm that X_i is one of multiple winners exactly when $Y_i = \max_{k\neq i} Y_k - Y_j$, or equivalently when p_{ij} is at least q^+_{ij} but at most

$$q_{ij}(Y_{-i}) = G_{ij}(\max_{k\neq i} Y_k - Y_j|Y_{-i}). \tag{6.13}$$

Moreover, in the case that there are multiple winners, the number of winners is also a deterministic function of Y_{-i}:

$$N_i(Y_{-i}) = 1 + |\{\ell \neq i : Y_\ell = \max_{k\neq i} Y_k\}|. \tag{6.14}$$

Note that $N_i(Y_{-i})$, which is always at least two, is <u>not</u> the same as the number of winners, which can be one. Rather, it is the number of winners there will be if X_i is a winner and at least one other X_k is as well (see Appendix B.12 for details).

Leveraging these facts, we can apply our framework to come up with a valid test for $H_{0,Wj}$. Essentially, we use p_{ij} to test $H_{0,ij}$ with probability one when it is less than q^+_{ij}, and with probability $1/N_i$ (we randomly select one of the N_i winners) when it is between q^+_{ij} and q_{ij}. Explicitly, letting $p = p_{ij}$ and $Z = Y_{-j}$, we can apply our framework with the selection function

$$s(x, z) = \begin{cases} 1 & \text{if } x < q^+_{ij}(z), \\ \frac{1}{N_i(z)} & \text{if } x \in [q^+_{ij}(z), q_{ij}(z)] \\ 0 & \text{otherwise} \end{cases}.$$

This is a piece-wise linear function that is easy to integrate, and, after some computations detailed in Appendix B.13, Theorem 5 tells us to reject when the selective p-value

$$\frac{p_{ij} - \left(1 - \frac{1}{N_i}\right)(p_{ij} - q^+_{ij})_+}{q^+_{ij} + \frac{1}{N_i}(q_{ij} - q^+_{ij})} \tag{6.15}$$

is at most α.

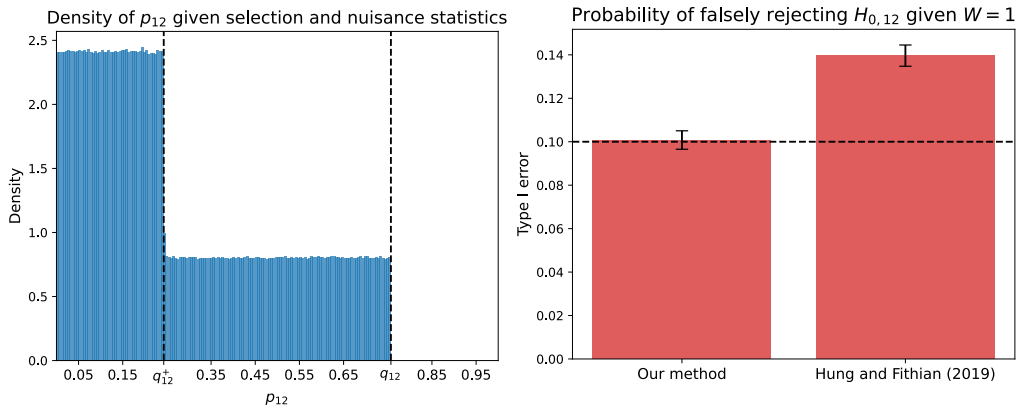


Figure 6.3: Considering three independent binomials $X_i \sim \text{Bin}(b, s_i)$ with $b = 4$ and $s_i = 1/2$, the first panel (left) depicts $N = 10^6$ draws from the conditional distribution of the p-value p_{12} (used for testing $H_{0,12} : s_1 \leq s_2$) given $W = 1$ and the nuisance statistics $(X_1 + X_2)/2 = 2$, $X_3 = 2$. When $p_{12} < q_{12}^+$, then X_1 is the sole winner and the p-value is selected for inference with probability one, but when $p_{12}^0 \in [q_{12}^+, q_{12}]$ there is a three-way tie and it is selected for inference with probability $1/3$. Hence, the p-value's conditional distribution is not uniform on $[0, q_{12}]$ as Hung and Fithian [2019b] implicitly assume. The next panel (right) displays the consequence. Conditional on $W = 1$, Hung and Fithian [2019b] do not maintain Type I error control when testing H_{W2}^0 at level $\alpha = 0.1$ (denoted by horizontal dashed line), whereas our method does. Error bars denote a 99% confidence interval.

The crucial difference between our derivation and Hung and Fithian [2019b]'s is that Hung and Fithian [2019b] use the same selective p-value when there can and cannot be ties amongst the X_k. If there cannot be ties amongst the X_k, then $q_{ij} = q_{ij}^+$ always, and the selective p-value (6.15) simplifies to p_{ij}/q_{ij}. If we use p_{ij}/q_{ij} when ties are possible, however, we will not achieve conditional error control (as is claimed in Hung and Fithian [2019b]). We give an example in Figure 6.3, where $X \in \mathbb{R}^3$ is composed of three independent binomials. The left panel of Figure 6.3 depicts the conditional distribution of p_{12} given $W = 1$ for a specific setting of the nuisances statistics Y_{-j}. It makes it clear that rejecting when $p_{ij}/q_{ij} \leq \alpha$ does not maintain conditional error control, as is affirmed in Figure 6.3's right panel.

Appendix A

Supplementary material relating to column subset selection

A.1 Additional simulation details and results

A.1.1 Missing data simulation setting

We describe the distribution used in Section 2.4.1. Recall the PCSS model:

$$\begin{aligned} X_S &\sim F \\ X_{-S} \mid X_S &\sim N(\mu_{-S} + \boldsymbol{W}(X_S - E_F[X_S]), \sigma^2 \boldsymbol{I}_{p-k}). \end{aligned} \tag{3.1, revisited}$$

We set the ground truth subset to be $S = \{1, 2, 3, 4\}$ and let X_S have a mean-zero multivariate Gaussian distribution with equicorrelated covariance $0.75 \cdot \boldsymbol{I}_4 + 0.25 \cdot 11^\top$. We set $\sigma^2 = 0.15$. The matrix $\boldsymbol{W}$ is given

below:

$$\boldsymbol{W} = \begin{bmatrix} \sqrt{17/90} & \sqrt{17/90} & \sqrt{17/90} & 0 \\ \sqrt{17/50} & \sqrt{17/50} & -\sqrt{17/50} & 0 \\ \sqrt{17/50} & -\sqrt{17/50} & \sqrt{17/50} & 0 \\ \sqrt{17/50} & -\sqrt{17/50} & -\sqrt{17/50} & 0 \\ -\sqrt{17/50} & \sqrt{17/50} & \sqrt{17/50} & 0 \\ -\sqrt{17/50} & \sqrt{17/50} & -\sqrt{17/50} & 0 \\ -\sqrt{17/50} & -\sqrt{17/50} & \sqrt{17/50} & 0 \\ -\sqrt{17/90} & -\sqrt{17/90} & -\sqrt{17/90} & 0 \\ 0 & \sqrt{17/90} & \sqrt{17/90} & \sqrt{17/90} \\ 0 & \sqrt{17/50} & \sqrt{17/50} & -\sqrt{17/50} \\ 0 & \sqrt{17/50} & -\sqrt{17/50} & \sqrt{17/50} \\ 0 & \sqrt{17/50} & -\sqrt{17/50} & -\sqrt{17/50} \\ 0 & -\sqrt{17/50} & \sqrt{17/50} & \sqrt{17/50} \\ 0 & -\sqrt{17/50} & \sqrt{17/50} & -\sqrt{17/50} \\ 0 & -\sqrt{17/50} & -\sqrt{17/50} & \sqrt{17/50}] \\ 0 & -\sqrt{17/90} & -\sqrt{17/90} & -\sqrt{17/90} \end{bmatrix}$$

It is designed so the rows have non-zero entries of equal magnitude, rows with the same non-zero entries have a different configuration of signs, and the variance of each variable is one.

A.1.2 Simulation setting for selecting subset size

We describe the distribution used in Section 3.3.1. Recall the subset factor model:

$$\begin{aligned} &X_S \sim F \qquad \epsilon \sim (0, \boldsymbol{D}) \qquad \epsilon_1 \perp, \ldots, \perp \epsilon_{p-k} \qquad \mathrm{Cov}(X_S, \epsilon) = \boldsymbol{0} \\ &X_{-S} = \boldsymbol{W}(X_S - E_F[X_S]) + \mu_{-S} + \epsilon. \end{aligned} \tag{3.2, revisited}$$

We set the ground truth subset to be $S = \{1, \ldots, 20\}$ and let X_S have a mean-zero multivariate Gaussian distribution with block diagonal covariance. Specifically there are five equally sized blocks, each of which is the equicorrelated matrix $0.5 \cdot \boldsymbol{I}_4 + 0.5 \cdot 11^\top$. The matrix $\boldsymbol{W}$ is somewhat sparse, and the non-zero entries are 1 or -1 with probability $1/2$. We document the non-zero entries of $\boldsymbol{W}$ here:

$$\boldsymbol{W}_{1:7,1:8} = \begin{bmatrix} -1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 \\ -1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\ -1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & 1 & -1 & 1 & 1 \end{bmatrix}, \boldsymbol{W}_{8:14,5:12} = \begin{bmatrix} 1 & -1 & 1 & 1 & 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & -1 & -1 & -1 & 1 & -1 \\ -1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & -1 & -1 & -1 & 1 & 1 \\ -1 & 1 & 1 & -1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\boldsymbol{W}_{15:21,9:16} = \begin{bmatrix} 1 & 1 & 1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & -1 & 1 & 1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & 1 & 1 & -1 & -1 & 1 \end{bmatrix}, \boldsymbol{W}_{22:30,13:20} = \begin{bmatrix} 1 & 1 & -1 & -1 & 1 & 1 & 1 & -1 \\ -1 & 1 & -1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & -1 & 1 \\ -1 & -1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & 1 & 1 & -1 \\ -1 & -1 & -1 & -1 & -1 & 1 & -1 & 1 \end{bmatrix}$$

To define $\boldsymbol{D}$, first we define the diagonal matrix $\widetilde{\boldsymbol{D}}$ where $\widetilde{\boldsymbol{D}}_{ii} = i \mod 6$. Then we set $\boldsymbol{D} = s \cdot \tilde{\boldsymbol{D}}$, where we use $s \in \mathbb{R}$ to control the amount of signal. The values of s we use are $s = 0.254, 0.812, 2.71, 9.2, 30.0$. The unique factors $\epsilon \sim (0, \boldsymbol{D})$ are generated independently of the principal variables. When the unique factors are non-Gaussian, ten of them are $\text{Exp}(1) - 1$ random variables, ten are Rademacher random variables that take value 1 or -1 with probability $1/2$, and ten are t-distributed random variables with 3 degrees of freedom. In the non-Gaussian case we scale the unique factors so their variance matches $\boldsymbol{D}$. The indices of the exponential random variables are $1, 11, 13, 17, 19, 20, 23, 24, 26, 27$, of the Rademacher random variables are $3, 4, 8, 9, 12, 14, 21, 22, 29, 30$, and of the t-distributed random variables are $2, 5, 6, 7, 10, 15, 16, 18, 25, 28$. These indices were chosen from a random permutation.

We also ran the same simulations in the case where $\boldsymbol{X}_S \sim N(0, \boldsymbol{I}_k)$, $\epsilon \sim N(0, \eta \boldsymbol{I}_k)$ for some $\eta > 0$, $X_S \perp \epsilon$, and the rows of $\boldsymbol{W}$ are randomly generated unit vectors. The results (not shown) are notably better than the ones we chose to display.

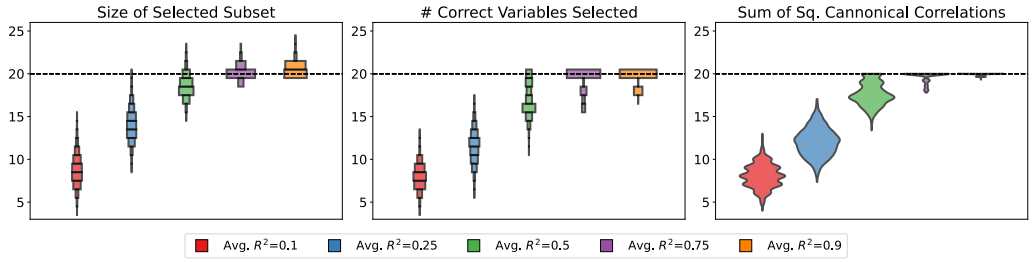
A.1.3 Additional simulations for selecting subset size

We give results from additional simulations, where we repeat the simulations from Section 3.3.1 but for different sample sizes n. Figure A.1 depicts the results for the Gaussian unique factor case and Figure A.2 for the non-Gaussian unique factor case.

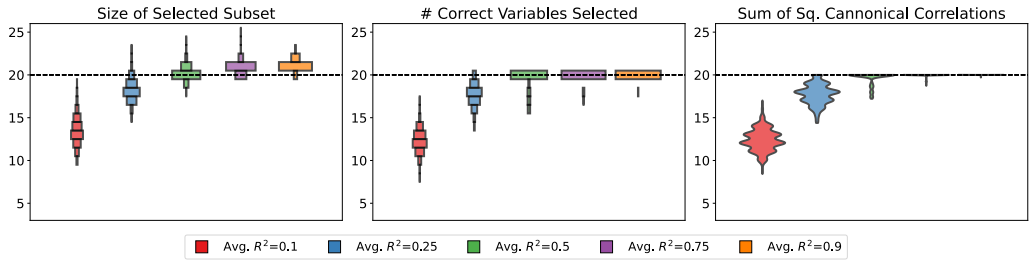
A.1.4 Selecting subset size with a forward-backward algorithm

We rerun our simulation from Section 3.3.1 but now using a forward-backward algorithm to find the smallest subset size for which we fail to reject. This is a natural choice, as forward-backward algorithms are commonly used for model selection tasks [Borboudakis and Tsamardinos, 2019]. Our algorithm completes a forward stage and then a backward stage. During each iteration of the forward stage, the algorithm greedily adds the variable that results in the largest decrease of the test statistic $T(\cdot)$ to the selected set. The forward stage continues until we find a subset for which we fail to reject. Then it begins the backward stage. On each iteration of the backward stage, the algorithm greedily removes the variable whose removal results in the smallest increase of $T(\cdot)$. The backward stage continues until removing any additional variable would result in a subset for which we reject. Conveniently, after one such forward and backward stage, our algorithm has converged, because removing any additional variables would result in rejecting, and there is no need to add more variables since we have already found a subset size for which we fail to reject.

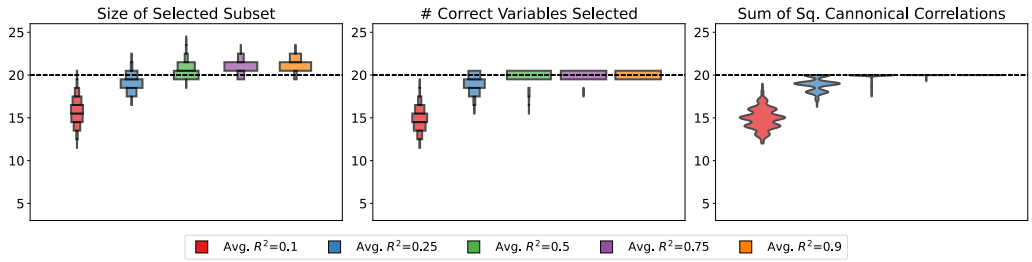
The results from replicating Section 3.3.1's simulations with our forward-backward algorithm are given in Figure A.3. Although the forward-backward algorithm is faster, it performs considerably worse than our



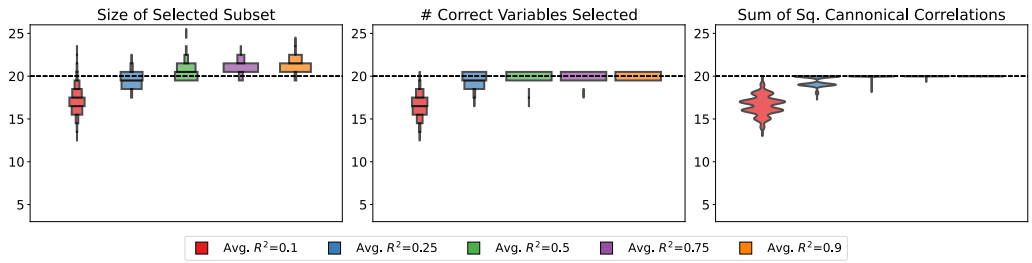
(a) $n = 100$



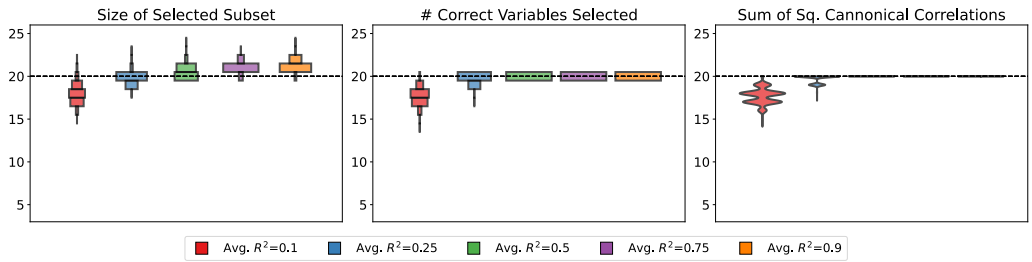
(b) $n = 200$



(c) $n = 300$

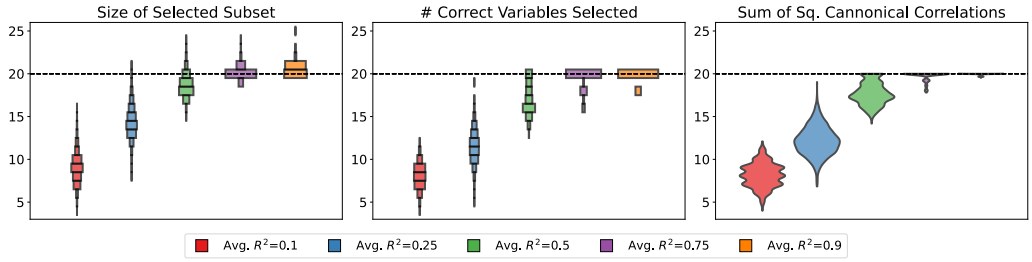


(d) $n = 400$

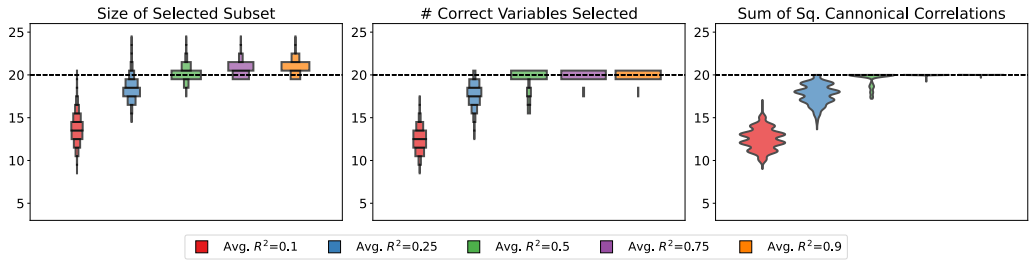


(e) $n = 500$

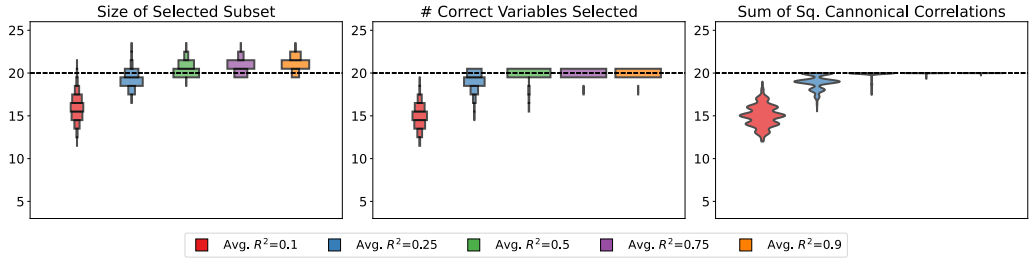
Figure A.1: Results for the case of Gaussian unique factors. For description of the plots, see the caption of Figure 3.1



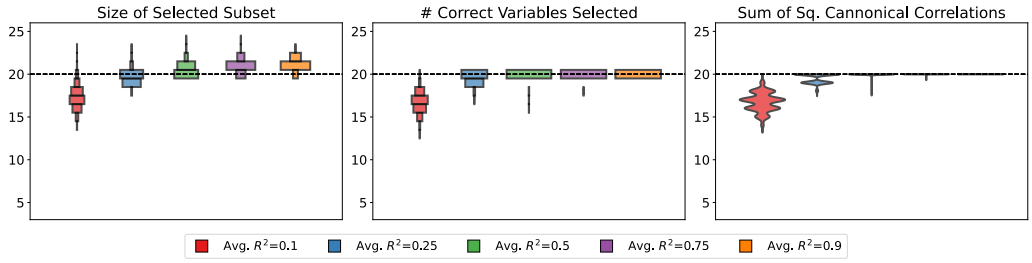
(a) $n = 100$



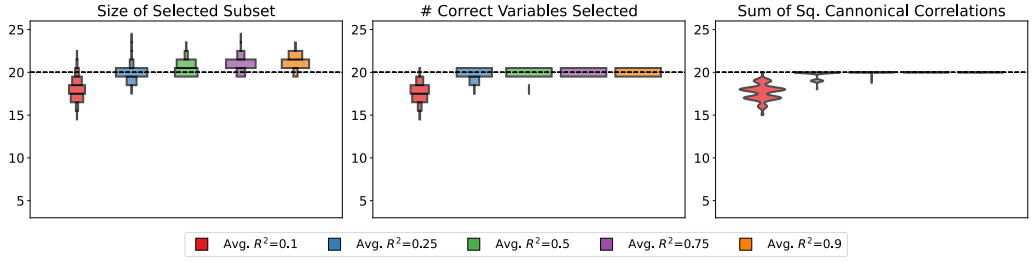
(b) $n = 200$



(c) $n = 300$

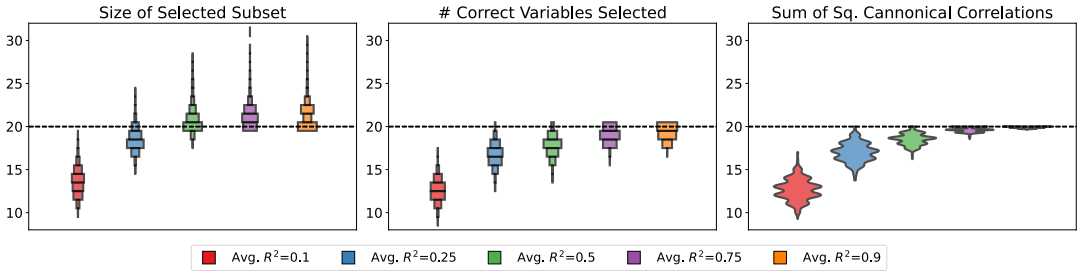


(d) $n = 400$

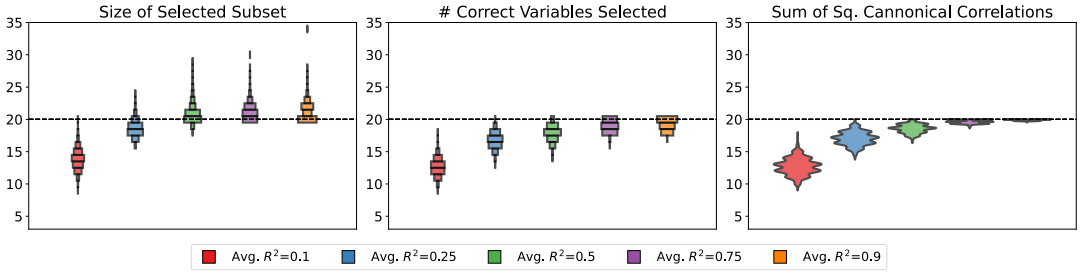


(e) $n = 500$

Figure A.2: Results for the case of non-Gaussian unique factors. For description of the plots, see the caption of Figure 3.1



(a) Gaussian unique factors



(b) Non-Gaussian unique factors

Figure A.3: Replication of the simulations whose results are in depicted Figure 3.1, but now using our forward-backward algorithm to perform subset search. For description of the plots, see caption of Figure 3.1.

swapping algorithm. Compared to the swapping algorithm, our forward-backward algorithm over selects more often and more drastically. Also, it often selects worse subsets, in the sense that the overlap between the variables it selects and the true population factor set is noticeably smaller. When we replicate the experiment for sample sizes other than $n = 200$ as we did for swapping in Appendix A.1.3, we see the same degradation in performance (results not shown).

A.2 More on selecting the appropriate subset size

In this section, we provide more details surrounding the procedures presented in Section 3.3. Recall that we want to identify k^*, the smallest k for which the data generating distribution belongs to the subset factor model, and that the procedures we consider output $\hat{k}$, the smallest k for which we are unable to reject the null that a size-k subset is sufficient. In Appendix A.2.1 we consider the restricted case where the unique factors are Gaussian and independent of the principal variables, and establish a test for this null with finite sample validity. In Appendix A.2.2, we show that the same test maintains asymptotic validity in the unrestricted case. This is immediately sufficient to imply the correctness of Theorem 4, the proof of which is detailed in Appendix A.7.12. In Appendix A.2.3, we discuss some desirable properties of the (ideally) outputted subset $\hat{S}$ and some practical considerations surrounding searching for it. Finally, in Appendix A.2.4 we give a discussion that establishes the analogous procedure for the PCSS model (3.1).

Throughout we consider $n > p$ data points $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ that are sampled from some distribution and have sample mean $\hat{\mu}$ and sample covariance $\hat{\mathbf{\Sigma}}$.

A.2.1 A finite-sample test when the unique factors are Gaussian

The goal of this section is to develop a test that is level α in finite samples for the null $\overline{H}_0 : |S| \leq k$ that the data is drawn from from a distribution in the k-dimensional subset factor model (3.2) with Gaussian unique factors ϵ that are independent of the principal variables X_S. This is the set of distributions that satisfy

$$\begin{aligned} X_S &\sim F \\ X_{-S} \mid X_S &\sim N(\mu_{-S} + \boldsymbol{W}(X_S - E_F[X_S]), \boldsymbol{D}). \end{aligned} \tag{A.1}$$

for some size-k subset S, where $\boldsymbol{D} \succ 0$ is an arbitrary diagonal matrix. In other words, it is the PCSS model (3.1) where the covariance of $X_{-S} \mid X_S$ is now allowed to be diagonal (rather than just isotropic).

Motivating a test statistic

Our test statistic more or less comes from a generalized likelihood ratio test (LRT). Specifically we compare the data's likelihood under a totally unrestricted multivariate Gaussian model to its likelihood under a restricted multivariate Gaussian model, where the restricted multivariate Gaussian model only considers distributions that are also in k-dimensional subset factor model. Suppose that $\hat{\boldsymbol{\Sigma}} \succ \boldsymbol{0}$. It is well known that the minimized negative log-likelihood, scaled by $1/n$, under the unrestricted Gaussian model is given by

$$\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}| + \frac{p}{2}(1 + \log(2\pi))$$

From [Petersen and Pedersen, 2008, 9.1.2], the above can be written as

$$\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}_U| + \frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U}| + \frac{p}{2}(1 + \log(2\pi)),$$

for any size-k subset U. Appendix A.7.5 in the proof of Theorem 1 tells us that the minimized negative log-likelihood, also scaled by $1/n$, under the restricted Gaussian model is

$$\min_{U\subset[p]:|U|=k} \frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}_U| + \frac{1}{2}\log(|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U})|) + \frac{p}{2}(1 + \log(2\pi))$$

The generalized LRT statistic is two times the difference in log likelihoods, and is thus given by

$$\min_{U\subset[p]:|U|=k} n\log\left(\frac{|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U})|}{|\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U}|}\right).$$

Our test statistic generalizes the LRT statistic by allowing for singular $\hat{\boldsymbol{\Sigma}}$:

$$T_k = \min_{U\subset[p]:|U|=k} T(U), \qquad T(U) = n\log\left(\frac{|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{+}\hat{\boldsymbol{\Sigma}}_{U,-U})|}{|\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{+}\hat{\boldsymbol{\Sigma}}_{U,-U}|}\right) \tag{3.4, revisited}$$

There may be size-k subsets U for which $|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{+}\hat{\boldsymbol{\Sigma}}_{U,-U})| = |\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{+}\hat{\boldsymbol{\Sigma}}_{U,-U}| = 0$, and we adopt the convention that $T(U) = 0$ for such U. In the case that $k = 0$, the same reasoning tells us that $T_0 = n\log\Big(|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}})|/|\hat{\boldsymbol{\Sigma}}|\Big)$.

Determining a critical value

We will come up with a critical value by studying the distribution of $T(S)$, where S is a size-k subset for which the data generating distribution satisfies (A.1). Since $T_k = \min_{U \subset [p]:|U|=k} T(U) \leq T(S)$, rejecting when T_k is larger than the appropriate quantile of $T(S)$ will be sufficient to guarantee type I error control.

To study the distribution of $T(S)$ we leverage techniques from fixed-X Ordinary Least Squares. It suffices to understand the distribution of $\hat{\boldsymbol{\Sigma}}_{-S} - \hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^+\hat{\boldsymbol{\Sigma}}_{S,-S}$, i.e, the distribution of the residual covariance from the regression of the $x_{-S}^{(i)}$ on the $x_S^{(i)}$. We are able to show that, conditional on the $x^{(i)}$, this covariance has a Wishart distribution. Then, via the Wishart distribution's Bartlett decomposition [Anderson et al., 1958, Corollary 7.2.1], we establish that the quantiles of $T(S)$ are bounded above by those of

$$n \sum_{j=2}^{p-k} \log\left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-k-j}}\right). \tag{3.5, revisited}$$

In (3.5) $\{\chi^2_\ell\}, \{\tilde{\chi}^2_\ell\}$ are mutually independent chi-squared random variables with degrees of freedom specified by their subscript. When $k = p-1$, (3.5) should be interpreted as a point mass at zero, and the same argument holds when $k = 0$.

Denoting the $(1-\alpha)$ quantile of (3.5) as $Q_{n,p,k}(1-\alpha)$, Lemma A.2.1 both summarizes this result and also establishes finite sample validity the test that rejects when $T_k > Q_{n,p,k}(1-\alpha)$.

Lemma A.2.1. *Consider $n > p$ samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution P that satisfies (A.1) for some size-k set S where $0 \leq k < p$. Then the quantiles of $T(S)$ are bounded above by the quantiles of (3.5). As a consequence, $P(T_k > Q_{n,p,k}(1-\alpha)) \leq \alpha$.*

A.2.2 An asymptotic test for the subset factor model

We now show that the test from Appendix A.2.1 maintains asymptotic validity for the more general null $H_{0,k} : |S| \leq k$ that the data is drawn from a distribution in the k-dimensional subset factor model. Recall that the k-dimensional subset factor model is the set of distributions satisfying

$$\begin{aligned} &X_S \sim F \qquad \epsilon \sim (0, \boldsymbol{D}) \qquad \epsilon_1 \perp, \ldots, \perp \epsilon_{p-k} \qquad \mathrm{Cov}(X_S, \epsilon) = \mathbf{0} \\ &X_{-S} = \boldsymbol{W}(X_S - E_F[X_S]) + \mu_{-S} + \epsilon. \end{aligned} \tag{3.2, revisited}$$

for some size-k subset S.

Lemma A.2.2 tells us that, should the data be drawn from a distribution satisfying (3.2) for some size-k set S, then $T(S)$ has a pivotal chi-squared limiting distribution. The degrees of freedom match what one would expect from Wilks' theorem [Wilks, 1938], although our setting is much more general. The proof is a careful application of the delta method and standard, but its result is surprising. Typically, we would expect the limiting distribution of a statistic like $T(S)$ to depend on fourth moments and therefore not be pivotal.

Lemma A.2.2. *Consider samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution that satisfies (3.2) for some size-k set S where $0 \leq k < p$. With p and k fixed, $T(S) \rightsquigarrow \chi^2_{(p-k)(p-k-1)/2}$ as the number of samples n tends to infinity.*

In Lemma A.2.2, $\chi^2_{(p-k)(p-k-1)/2}$ should be interpreted as a point mass at zero when $k = p-1$.

Using Lemma A.2.2, Lemma A.2.3 establishes the asymptotic validity of the same test from Appendix A.2.1 under the more general null $H_{0,k} : |S| \leq k$. Using Lemma A.2.2, we can easily show that the $(1-\alpha)$ quantile of (3.5) converges to the $(1-\alpha)$ quantile of the $\chi^2_{(p-k)(p-k-1)/2}$ distribution, which we will denote $Q_{p,k}(1-\alpha)$. Because $T(S)$ has a limiting distribution that admits a density, it is then immediate that the difference between $P(T(S) > Q_{n,p,k}(1-\alpha))$ and $P(T(S) > Q_{p,k}(1-\alpha))$ converges to zero, and this is sufficient to imply the result.

Lemma A.2.3. *For samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution P that satisfies (A.1) for some size-k set S where $0 \leq k < p$, $\limsup_{n\to\infty} P(T_k > Q_{n,p,k}(1-\alpha)) \leq \alpha$.*

Our decision to use the quantiles of (3.5) rather than those of the $\chi^2_{(p-k)(p-k-1)/2}$ distribution is intentional, even though the latter guarantee the same asymptotic validity. In many practical settings, n is not sufficiently larger than p for our asymptotic guarantees to be reflected in our finite sample results. Although rejecting according to the quantiles of (3.5) only guarantees finite sample validity when the unique factors are Gaussian (and independent of the principal variables), we may hope that this validity is somewhat robust to deviations from Gaussianity. If so, rejecting according to the quantiles of (3.5) should improve performance for smaller sample sizes while maintaining the same asymptotic guarantees.

Figure A.4 illustrates the benefits of rejecting according to the quantiles of (3.5). In Figure A.4, we consider settings with $p = 50$ variables, a size $k = 20$ factor set S, and either $n = 200$ or $n = 5000$ samples, and plot three different distributions: (1) the limiting chi-squared distribution of $T(S)$; (2) the distribution (3.5); and (3) the distribution of $T(S)$ in an example case where the unique factors are not Gaussian. In this non-Gaussian example, the thirty unique factors consist of ten centered Poi(λ) random variables with $\lambda = 1, \ldots, 10$, ten centered chi-squared random variables with $1, \ldots, 10$ degrees of freedom, and ten uniform random variables that range from $-a$ to a for $a = 1, \ldots, 10$. The principal variables are independent standard Gaussians and regression coefficient matrix $\boldsymbol{W} \in \mathbb{R}^{(p-k)\times k}$ has uniformly random unit vector rows. Figure A.4 has two important takeaways. First, at reasonable sample sizes, the limiting chi-squared distribution is a poor approximation for the actual distribution of $T(S)$, and rejecting according to its quantiles will lead to over-selection. Second, even when the unique factors are not Gaussian and the sample size is reasonably small, (3.5) is an excellent approximation for the actual distribution of $T(S)$. This remained to be the case when we varied n, p, k, and the distributions of unique factors.

A.2.3 Examination of the selected subset

Ultimately, having settled on selecting a size-$\hat{k}$ subset, we would ideally the size-$\hat{k}$ subset $\hat{S}$ that minimizes $T(U)$. Unlike in Section 3.3, $\hat{S}$ in this section refers to any actual global minimizer of $T(U)$ that would be found via exhaustive search. In what follows, we give a useful re-characterization of $\hat{S}$.

First, we re-characterize $\hat{S}$ as an MLE. Given that $T(U)$ is motivated by the LRT, this should not come as a surprise. Consider the case that every $(k+1)$ by $(k+1)$ principal sub-matrix of $\hat{\boldsymbol{\Sigma}}$ is full rank. Then we can use [Petersen and Pedersen, 2008, 9.1.2] to compute for any size-k subset U that

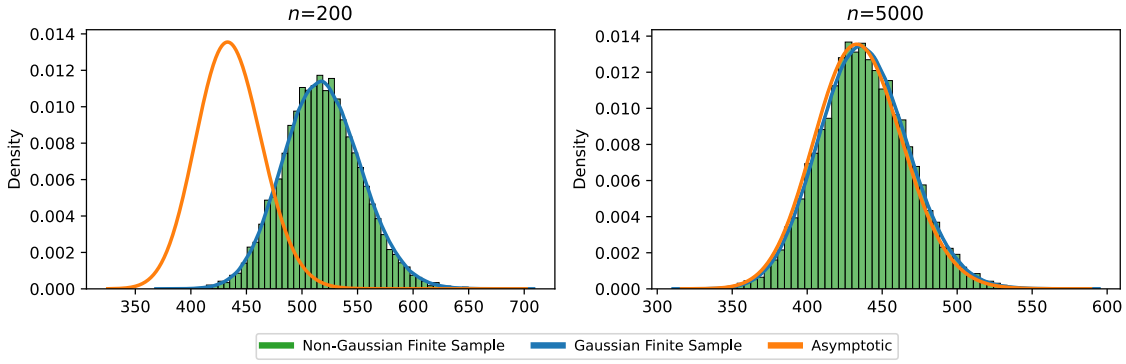


Figure A.4: In the case of $p = 50$ variables and a size $k = 20$ subset S, the exact finite sample distribution of $T(S)$ for $n = 200$ (left) and $n = 5000$ (right) samples in the cases of both Gaussian and non-Gaussian unique factors, plotted against its limiting $\chi^2_{(p-k)(p-k-1)/2}$ distribution.

$$
\begin{aligned}
n \log\left(\frac{|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U})|}{|\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U}|}\right) &= n \log\left(\frac{|\hat{\boldsymbol{\Sigma}}_U||\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U})|}{|\hat{\boldsymbol{\Sigma}}_U||\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U}|}\right) \\
&= n \log\left(\frac{|\hat{\boldsymbol{\Sigma}}_U||\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U})|}{|\hat{\boldsymbol{\Sigma}}|}\right) \\
&= n\left(\log(|\hat{\boldsymbol{\Sigma}}_U|) + \log(|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U})|)\right) \\
&\quad - n\log(|\hat{\boldsymbol{\Sigma}}|)
\end{aligned}
$$

So, in this case, every minimizer $\hat{S}$ of $T(U)$ belongs to

$$
\underset{U \subset [p]:|U|=k}{\operatorname{argmin}} \log(|\hat{\boldsymbol{\Sigma}}_U|) + \log(|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{U,-U})|) \tag{A.2}
$$

Correspondingly, in this case, the subsets belonging to (A.2) are exactly the maximum likelihood estimators for S in the restricted Gaussian model that only consists of distributions also in the $\hat{k}$-dimensional subset factor model (see Appendix A.7.5 in the proof of Theorem 1).

Solving (A.2) also handles the case that every $(k+1)$ by $(k+1)$ principal sub-matrix of $\hat{\boldsymbol{\Sigma}}$ is not full rank. In this case, there must be a size-k subset U such that $T(U) = 0$, and we should fail to reject. However, every $(k+1)$ by $(k+1)$ principal sub-matrix of $\hat{\boldsymbol{\Sigma}}$ is not full rank if and only if the minimized value of (A.2) is $-\infty$. Thus solving (A.2) is sufficient to handle both cases.

We can easily modify our algorithms from Section 2.3.1 to efficiently search for a subset that solves (A.2). We detail how to do this in Appendix A.3.

Because $\hat{S}$ results in the most empirically diagonal residual covariance, it is immediate that $\hat{S}$ asymptotically (fixed p, n goes to infinity) recovers a subset that results in an exactly diagonal population residual covariance, provided that $\hat{k} \geq k^*$. Unfortunately, this does not necessarily mean that $\hat{S}$ will recover a subset S for which the subset factor model (3.2) is satisfied, because the subset factor model requires that the unique factors are independent. We can at best guarantee that they are uncorrelated. Nonetheless, aside from some pathological cases, we can guarantee recovery.

Lemma A.2.4 uses the following less restrictive version of the subset factor model to state our recovery result:

$$\begin{aligned} &X_S \sim F \qquad \epsilon \sim (0, \boldsymbol{D}) \qquad \mathrm{Cov}(X_S, \epsilon) = \mathbf{0} \\ &X_{-S} = \boldsymbol{W}(X_S - E_F[X_S]) + \mu_{-S} + \epsilon. \end{aligned} \tag{A.3}$$

Lemma A.2.4. *Consider samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution P that satisfies (A.3) for some size-k set S. Suppose that P has population covariance $\boldsymbol{\Sigma}$ such that $\boldsymbol{\Sigma}_S \succ 0$. Then, with probability one, P eventually satisfies (A.3) with $\hat{S}$ as the number of samples n tends to infinity.*

A.2.4 The PCSS model

We provide an exactly analogous discussion for the PCSS model (3.1). Recall that the k-dimensional PCSS model is the set of distributions satisfying

$$\begin{aligned} &X_S \sim F \\ &X_{-S} \mid X_S \sim N(\mu_{-S} + \boldsymbol{W}(X_S - E_F[X_S]), \sigma^2 \boldsymbol{I}_{p-k}). \end{aligned} \tag{3.1, revisited}$$

for some size-k set S. Again, our procedure returns $\hat{k}$, the smallest k for which we fail to reject the null $\widetilde{H}_0 : |S| \leq k$ that the data generating distribution belongs to the k-dimensional PCSS model. Like before, we provide a test for this null that is valid in finite samples and we comment on some qualities of the selected subset $\hat{S}$. Because the PCSS model makes a Gaussian assumption, we do not need to discuss asymptotics.

Motivating a test statistic

Like in Appendix A.2.1, we consider a generalized LRT. Specifically, we compare the data's likelihood under a totally unrestricted multivariate Gaussian model to its likelihood under a restricted multivariate Gaussian model, where the restricted multivariate Gaussian model only considers distributions that are also in k-dimensional PCSS model. Again, suppose that $\boldsymbol{\Sigma} \succ \mathbf{0}$. Identical reasoning as in Appendix A.2.1 along with computations from the proof of Theorem 1 (see Appendix A.7.5) imply that the LRT test statistic is

$$\min_{U \subset [p] : |U| = k} n \log \left(\frac{\left(\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U} \hat{\boldsymbol{\Sigma}}_U^{-1} \hat{\boldsymbol{\Sigma}}_{U,-U}) / (p-k) \right)^{p-k}}{|\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U} \hat{\boldsymbol{\Sigma}}_U^{-1} \hat{\boldsymbol{\Sigma}}_{U,-U}|} \right)$$

Again, we will consider a generalization of this statistic that allows $\hat{\boldsymbol{\Sigma}}$ to be singular:

$$\widetilde{T}_k = \min_{U \subset [p] : |U| = k} \widetilde{T}(U), \qquad \widetilde{T}(U) = n \log \left(\frac{\left(\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U} \hat{\boldsymbol{\Sigma}}_U^{+} \hat{\boldsymbol{\Sigma}}_{U,-U}) / (p-k) \right)^{p-k}}{|\hat{\boldsymbol{\Sigma}}_{-U} - \hat{\boldsymbol{\Sigma}}_{-U,U} \hat{\boldsymbol{\Sigma}}_U^{+} \hat{\boldsymbol{\Sigma}}_{U,-U}|} \right)$$

Determining a critical value

Like in Appendix A.2.1, we will consider a size-k subset S for which the data generating distribution satisfies (3.1) and study the distribution of $\widetilde{T}(S)$. For simplicity we will assume that the principal variable distribution F admits a density. Like before, the residual covariance still has a conditional Wishart distribution, and we can use the Wishart's Bartlett decomposition to show that $\widetilde{T}(S)$ has distribution

$$n \log \left(\left(\frac{\left(\tilde{\chi}^2_{(p-k)(p-k-1)/2} + \sum_{j=1}^{p-k} \chi^2_{n-k-j} \right)}{p-k} \right)^{p-k} \Big/ \left(\prod_{j=1}^{p-k} \chi^2_{n-k-j} \right) \right) \tag{A.4}$$

In (A.4), $\{\chi^2_\ell\}, \{\tilde{\chi}^2_\ell\}$ are mutually independent chi-squared random variables with degrees of freedom specified by their subscript. Defining $\widetilde{Q}_{n,p,k}(1-\alpha)$ as the $(1-\alpha)$-quantile of (A.4), Lemma A.2.5 establishes a valid finite sample test for the null $\widetilde{H}_0 : |S| \leq k$.

Lemma A.2.5. *Consider $n > p$ samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution P that satisfies (3.1) for some size-k set S where $0 \leq k < p$. Suppose further that P admits a density. Then $\widetilde{T}(S)$ has distribution (A.4) and, as a consequence, $P(\widetilde{T}_k > \widetilde{Q}_{n,p,k}(1-\alpha)) \leq \alpha$.*

Proposition A.2.1 establishes error control of our procedure for the PCSS model. It is analogous to Theorem 4.

Proposition A.2.1. *Consider $n > p$ samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ from a distribution P that admits a density and let k^* be the smallest k for which P belongs to the k-dimensional PCSS model (3.1). If $\hat{k}$ is the smallest k for which the test statistic $\widetilde{T}_k$ is not strictly larger than the critical value $\widetilde{Q}_{n,p,k}(1-\alpha)$, then*

$$P(\hat{k} > k^*) \leq \alpha.$$

Examining the selected subset

As in Appendix A.2.3 we ultimately select the subset $\widetilde{S}$ that globally minimizes $\widetilde{T}(U)$. Like before, we can characterize this subset as an MLE. Supposing that every k by k principal sub-matrix of $\hat{\mathbf{\Sigma}}$ is full rank and no size-k subset empirically perfectly linearly reconstructs the remaining variables, similar computations to those in Appendix A.2.3 shows that a size-k subset U minimizes $\widetilde{T}(U)$ if and only if it solves

$$\underset{U \subset [p] : |U| = k}{\operatorname{argmin}} \log |\hat{\mathbf{\Sigma}}_U| + (p-k) \log(\operatorname{Tr}(\hat{\mathbf{\Sigma}}_{-U} - \hat{\mathbf{\Sigma}}_{-U,U} \hat{\mathbf{\Sigma}}_U^{-1} \hat{\mathbf{\Sigma}}_{U,-U})/(p-k)) \tag{A.5}$$

Under the same conditions, it is also true that U solves (A.5) if and only if it is the MLE for S in the restricted multivariate Gaussian model that only contains distributions also in the k-dimensional PCSS model (see Appendix A.7.5 in the proof of Theorem 1)

If the conditions from the previous paragraph are not met, it is unfortunately the case that the subset that minimizes $\widetilde{T}(U)$ may not solve (A.5). The conditions from the previous paragraph are not met, however, if and only if the minimized objective of (A.5) is $-\infty$. Solving (A.5) thus warns us if this is the case, but it is not totally sufficient as was true in Appendix A.2.3.

We show how to search for the subset that solves (A.5) using Section 2.3.1's algorithms in Appendix A.3.

Criterion/Method	Algorithm 1	Naive
McCabe's First Criterion	$O(\min\{p^2k, pk^3\})$	$O(p^4k)$ or $O(pk^4)$
McCabe's Second Criterion	$O(p^2k)$	$O(p^3k^2)$
McCabe's Third Criterion	$O(p^3k)$	$O(p^3k^2)$
McCabe's Fourth Criterion	$O(p^3k)$	$O(p^3k^2)$

Table A.1: A time complexity comparison of our algorithms to the naive algorithms that, when features are added or removed from the current subset, recompute the objective as written with no optimizations. The reported time corresponds to one complete run of Algorithm 1 or one iteration of Algorithm 2. McCabe himself pointed out that his first criteria of minimizing $|\mathbf{\Sigma}_{-S} - \mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{S,-S}|$ is equivalent to maximizing $|\mathbf{\Sigma}_S|$, so we report the naive time complexities for both versions of the objective.

Analogous to Lemma A.2.4, Lemma A.2.6 uses the following generalization of the PCSS model to state a subset recovery result:

$$\begin{aligned} &X_S \sim F \qquad \epsilon \sim (0, \sigma^2 I) \qquad \mathrm{Cov}(X_S, \epsilon) = \mathbf{0} \\ &X_{-S} = \boldsymbol{W}(X_S - E_F[X_S]) + \mu_{-S} + \epsilon. \end{aligned} \tag{A.6}$$

It says that $\widetilde{S}$ eventually will result in a isotropic population residual covariance.

Lemma A.2.6. *Consider samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution P that satisfies (A.6) for some size-k set S. Suppose further that, at the population level, some set of k variables perfectly linearly reconstruct the remaining. Then, with probability one, P eventually satisfies (A.6) with $\widetilde{S}$ as the number of samples n tends to infinity.*

A.3 Efficient subset search for other objectives

To modify Algorithm 1 and Algorithm 2 to perform subset search under a new objective we simply need to alter f in (2.3) so that $i \notin U \subset [p]$ that minimizes $f(i, \cdot)$ also minimizes the new objective over the subsets $U + i$. In this section, we explicitly give $f(i, \cdot)$ that allow us to search for subsets that minimize $T(\cdot)$ (see Section 3.3), $\widetilde{T}(\cdot)$ (see Appendix A.2.4), and McCabe's other three criteria. We generalize all of McCabe's criteria by changing inverses to pseudo-inverses, thereby allowing for singular $\mathbf{\Sigma}$. Correctness of our proposed $f.(i, \cdot)$ are implied by derivations in Appendix A.7.17.

For McCabe's criteria, Table A.1 contrasts the time complexity of our algorithms with that of naive implementations. In the naive implementation, the practitioner recomputes the objective as written with no optimizations when adding/removing elements to/from the current subset. For both our algorithms and the naive implementation, the time complexity of searching for subsets that minimize $T(\cdot)$ and $\widetilde{T}(\cdot)$ matches that of McCabe's second criterion.

Minimizing $T(\cdot)$: The discussion in Appendix A.2.3 tells us that it suffices to find a subset that minimizes $\log|\mathbf{\Sigma}_S| + \log(|\mathrm{Diag}(\mathbf{\Sigma}_{R(X_{-S}, X_S)})|)$. To perform this search, use

$$\begin{aligned} f(i, U, \mathbf{\Sigma}_{R(X,X_U)}) &= \log((\mathbf{\Sigma}_{R(X,X_U)})_{ii}) \\ &\quad + \sum_{j \notin U+i} \log\left((\mathbf{\Sigma}_{R(X,X_U)})_{jj} - (\mathbf{\Sigma}_{R(X,X_U)})_{ij}^2 / (\mathbf{\Sigma}_{R(X,X_U)})_{ii} \cdot I_{(\mathbf{\Sigma}_{R(X,X_U)})_{ii} > 0}\right) \end{aligned}$$

Minimizing $\widetilde{T}(\cdot)$: The discussion in Appendix A.2.3 suggests finding a subset that minimizes $\log|\mathbf{\Sigma}_S| + (p-k)\log(\mathrm{Tr}(\mathbf{\Sigma}_{R(X_{-S},X_S)})/(p-k))$. If the objective value is not $-\infty$ then solving this problem is sufficient. Otherwise we cannot say anything concrete. To perform this search, use

$$f(i,U,\mathbf{\Sigma}_{R(X,X_U)}) = \log((\mathbf{\Sigma}_{R(X,X_U)})_{ii}) \\ + (p-k)\cdot\log\left(\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_U)}) - \|(\mathbf{\Sigma}_{R(X,X_U)})_{\bullet i}\|_2^2/(\mathbf{\Sigma}_{R(X,X_U)})_{ii}\cdot I_{(\mathbf{\Sigma}_{R(X,X_U)})_{ii}>0}\right)$$

McCabe's first criterion: McCabe's first criterion suggests selecting a subset that minimizes $|\mathbf{\Sigma}_{-S} - \mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{S,-S}|$. We generalize this to selecting the subset which minimizes $|\mathbf{\Sigma}_{-S} - \mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^{+}\mathbf{\Sigma}_{S,-S}| = |\mathbf{\Sigma}_{R(X,X_S)}|$. For simplicity, we will assume that there exists a size-k subset such that $|\mathbf{\Sigma}_S| > 0$. If this is not the case, then every size k subset has redundant variables which can be perfectly linearly reconstructed by other variables in the subset, and the practitioner should use a smaller subset size. To perform subset search according to McCabe's first criterion use:

$$f(i,\mathbf{\Sigma}_{R(X,X_U)}) = -(\mathbf{\Sigma}_{R(X,X_U)})_{ii}$$

Alternatively, as we add/remove variables to/from U, we can keep track of $\mathbf{\Sigma}_U^+$ instead of $\mathbf{\Sigma}_{R(X,X_U)}$ using [Petersen and Pedersen, 2008, 3.2.7] and instead use

$$f(i,\mathbf{\Sigma},\mathbf{\Sigma}_U^+) = -(\mathbf{\Sigma}_{ii} - \mathbf{\Sigma}_{iU}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{Ui})$$

Lemma 1 implies that this is equivalent to our earlier suggestion, but it has smaller time complexity when $k \ll \sqrt{p}$.

McCabe's third criterion: McCabe's third criterion suggests selecting a subset that minimizes $\|\mathbf{\Sigma}_{-S} - \mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{S,-S}\|_F^2$. We generalize this to selecting the subset which minimizes $\|\mathbf{\Sigma}_{-S} - \mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^{+}\mathbf{\Sigma}_{S,-S}\|_F = \|\mathbf{\Sigma}_{R(X,X_S)}\|_F^2$. To perform subset search according to McCabe's third criterion use:

$$f(i,\mathbf{\Sigma}_{R(X,X_U)}) = [(\|\beta\|_2^2/\beta_i)^2 - 2\beta^\top\mathbf{\Sigma}_{R(X,X_U)}\beta/\beta_i]\cdot I_{\beta_i>0}$$

where $\beta = (\mathbf{\Sigma}_{R(X,X_U)})_{\bullet i}$.

McCabe's fourth criterion: McCabe's fourth criterion suggests selecting a subset that maximizes $\mathrm{Tr}(\mathbf{\Sigma}_{-S}^{-1}\mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{S,-S})$. This is equivalent to maximizing the sum of squared cannonical correlations between X_S and X_{-S}. We generalize this to selecting the subset which maximizes $\mathrm{Tr}(\mathbf{\Sigma}_{-S}^{+}\mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^{+}\mathbf{\Sigma}_{S,-S})$ which has the same interpretation in the more general case where $\mathbf{\Sigma}$ is singular. Consider a currently selected subset U. Fix an $i \notin U$ and let $V = U + i$. Take j and h to be the rank of i in V and $-U$ respectively. So, for example, $(\mathbf{\Sigma}_V)_{jj} = \mathbf{\Sigma}_{ii}$ and $(\mathbf{\Sigma}_{-U})_{hh} = \mathbf{\Sigma}_{ii}$ Then, to perform subset search in accordance with McCabe's fourth criterion use

$$f(i,U,\mathbf{\Sigma},\mathbf{\Sigma}_{R(X_{-U},X_U)},\mathbf{\Sigma}_{R(X_U,X_{-U})},\mathbf{\Sigma}_{-U}^+,\mathbf{\Sigma}_U^+) = \\ \mathrm{Tr}((\mathbf{\Sigma}_V^+)_{-j,-j}\mathbf{\Sigma}_{R(X_U,X_{-U})}) + \frac{\beta^\top\mathbf{\Sigma}_V^+\beta}{\beta_j}I_{\beta_j>0} - I_{(\mathbf{\Sigma}_{R(X_{-U},X_U)})_{hh}>0}$$

where $\beta = \mathbf{\Sigma}_{V,i} - \mathbf{\Sigma}_{V,-V}\mathbf{\Sigma}^+_{-V}\mathbf{\Sigma}_{-V,i}$, and we compute $\mathbf{\Sigma}^+_V$ from $\mathbf{\Sigma}^+_U$ and $\mathbf{\Sigma}^+_{-V}$ from $\mathbf{\Sigma}^+_{-U}$ using [Petersen and Pedersen, 2008, 3.2.7]. We keep track of $\mathbf{\Sigma}^+_U$ and $\mathbf{\Sigma}^+_{-U}$ as we add/remove elements to/from U using [Petersen and Pedersen, 2008, 3.2.7] and keep track of $\mathbf{\Sigma}_{R(X_{-U},X_U)}$ and $\mathbf{\Sigma}_{R(X_U,X_{-U})}$ using Lemma 1 as usual.

If $\mathbf{\Sigma} \succ \mathbf{0}$, then we can give an exact closed form for the replacement f given above:

$$f(i, U, \mathbf{\Sigma}, \mathbf{\Sigma}_{R(X_{-U},X_U)}, \mathbf{\Sigma}^{-1}_{-U}, \mathbf{\Sigma}^{-1}_{U}) = \frac{\beta^\top_{-j}\mathbf{\Sigma}^{-1}_U\beta_{-j}}{\beta_j} - \frac{(\alpha^\top\beta_{-j})^2}{\delta\beta_j} - \frac{\alpha^\top\mathbf{\Sigma}_{R(X_U,X_{-U})}\alpha + 2\alpha^\top\beta_{-j} - \beta_j}{\delta}$$

where $\alpha = \mathbf{\Sigma}^{-1}_U\mathbf{\Sigma}_{Ui}$, $\delta = \mathbf{\Sigma}_{ii} - \mathbf{\Sigma}_{iU}\alpha$ and β is as before, but now we can explicitly compute $\mathbf{\Sigma}^+_{-V} = \mathbf{\Sigma}^{-1}_{-V} = (\mathbf{\Sigma}^{-1}_{-U})_{-h,-h} - (\mathbf{\Sigma}_{-U})_{-h,h}(\mathbf{\Sigma}_{-U})_{h,-h}/(\mathbf{\Sigma}_{-U})_{hh}$. Also, as we add elements to U, [Khan, 2008] now allows us to keep track of $\mathbf{\Sigma}^+_U = \mathbf{\Sigma}^{-1}_U$ and $\mathbf{\Sigma}^+_{-U} = \mathbf{\Sigma}^{-1}_{-U}$ with simpler closed form updates.

A.4 The high-dimensional consistency of CSS

In this section we characterize a high-dimensional setting where the size-k CSS estimate (1.1) one gets from the sample covariance is consistent for the size-k CSS solution according to the population covariance (2.1). In particular, we fix k and imagine observing n samples from a distribution P over $\mathbb{R}^p$, where p is possibly growing with n. We denote the population covariance under P as $\mathbf{\Sigma} \in \mathbb{S}^{p\times p}_+$. Our analysis applies in the proportional asymptotic regime, where p/n converges to a constant. For the result to apply, P must satisfy three assumptions as n grows.

Assumption 1 (Light-tailed data)**.** *The sub-Gaussian norm (see [Vershynin, 2018, Definition 2.5.6]) of each variable is bounded above by a constant independent of p and n.*

Assumption 2 (Low-dimensional invertibility)**.** *The minimum of the smallest eigenvalues of all the $k \times k$ principal sub-matrices of $\mathbf{\Sigma}$ is bounded below by a positive constant independent of p and n.*

Assumption 3 (Separation)**.** *For some $\delta > 0$, the achieved CSS objective for the population best subset*

$$S^* = \operatorname*{argmin}_{S\subseteq[p],|S|=k} \operatorname{Tr}(\mathbf{\Sigma} - \mathbf{\Sigma}_{\bullet S}\mathbf{\Sigma}^+_S\mathbf{\Sigma}_{\bullet S})$$

is well separated from the objective achieved by any other subset

$$\min_{S\neq S^*} \operatorname{Tr}(\mathbf{\Sigma} - \mathbf{\Sigma}_{\bullet S}\mathbf{\Sigma}^+_S\mathbf{\Sigma}_{\bullet S}) - \operatorname{Tr}(\mathbf{\Sigma} - \mathbf{\Sigma}_{\bullet S^*}\mathbf{\Sigma}^+_{S^*}\mathbf{\Sigma}_{\bullet S^*}) = \Omega(p^{1/2+\delta})$$

Note that Assumption 2 is much weaker than assuming a lower bound on the minimum eigenvalue of $\mathbf{\Sigma}$. Under Assumption 2 , $\mathbf{\Sigma}$ can still be singular.

Under the above three assumptions, Theorem 11 guarantees that the in-sample CSS solution is consistent for the population CSS solution.

Theorem 11 (High-dimensional setting where CSS is consistent)**.** *Suppose we observe samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ with sample covariance $\hat{\mathbf{\Sigma}} \in \mathbb{S}^{p\times p}_+$ from a distribution P with population covariance $\mathbf{\Sigma} \in \mathbb{S}^{p\times p}_+$. Fix k, and consider the proportional asymptotic regime where p/n converges to a constant as $n, p \to \infty$. If P satisfies*

Assumption 1, Assumption 2, and Assumption 3 as $n, p \to \infty$, then any in-sample CSS solution

$$\hat{S} \in \operatorname*{argmin}_{S \subseteq [p], |S|=k} \operatorname{Tr}(\hat{\boldsymbol{\Sigma}} - \hat{\boldsymbol{\Sigma}}_{\bullet S} \hat{\boldsymbol{\Sigma}}_S^{+} \hat{\boldsymbol{\Sigma}}_{\bullet S})$$

is consistent for the population CSS solution

$$S^* \in \operatorname*{argmin}_{S \subseteq [p], |S|=k} \operatorname{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet S} \boldsymbol{\Sigma}_S^{+} \boldsymbol{\Sigma}_{\bullet S}),$$

i.e.,

$$\lim_{n \to \infty} P(\hat{S} = S^*) = 1.$$

Note that, by Assumption 3, the population CSS solution S^* is eventually unique.

The rest of this section is a proof of Theorem 11. We use $\|\cdot\|_2$ to denote the spectral norm of a matrices. We denote the sorted eigenvalues of a positive semi-definite matrix $\boldsymbol{A} \in \mathbb{S}_+^{q \times q}$ as $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_q(\boldsymbol{A}) \geq 0$. For a matrix $\boldsymbol{M} \in \mathbb{R}^{q \times k}$ we denote its sorted top-k singular values as $s_1(\boldsymbol{M}) \geq \cdots \geq s_k(\boldsymbol{M}) \geq 0$. We use c_i to denote positive constants that are independent of n and p. If we introduce a new c_i, that means that such a c_i exists. Also, we make use of the stochastic calculus notation of [Vaart, 1998, Chapter 2].

A.4.1 Central argument

First we consider the case that $X \sim (0, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma} \in \mathbb{S}_+^{p \times p}$ is mean-zero and let $\widetilde{\boldsymbol{\Sigma}}_{ij} = n^{-1} \sum_{\ell=1}^n X_i^{(\ell)} X_j^{(\ell)}$ be our estimate of $\boldsymbol{\Sigma} = E[XX^\top]$. We will use the notation

$$\operatorname{Obj}(\boldsymbol{\Omega}, U) = p^{-1/2-\delta} \operatorname{Tr}(\boldsymbol{\Omega} - \boldsymbol{\Omega}_{\bullet U} \boldsymbol{\Omega}_U^{+} \boldsymbol{\Omega}_{U \bullet}).$$

to denote the CSS objective for the covariance $\boldsymbol{\Omega} \in \mathbb{S}_+^{p \times p}$, scaled by $p^{-1/2-\delta}$. Our goal is to show that $\tilde{S}$, which is a minimizer (break ties randomly) of $\operatorname{Obj}(\widetilde{\boldsymbol{\Sigma}}, S)$ for size-k subsets S, is consistent for S^*, which is the minimizer of $\operatorname{Obj}(\boldsymbol{\Sigma}, S)$ for size-k subsets S. By Assumption 3, $\operatorname{Obj}(\boldsymbol{\Sigma}, S)$ eventually has a unique minimizer.

Assumption 3 tells us exactly that

$$\min_{S \neq S^*} \operatorname{Obj}(\boldsymbol{\Sigma}, S) - \operatorname{Obj}(\boldsymbol{\Sigma}, S^*) = \Omega(1),$$

If we show that

$$\max_{S \subseteq [p] : |S|=k} \widetilde{\Delta}(S) = o_p(1), \qquad \widetilde{\Delta}(S) = |\operatorname{Obj}(\boldsymbol{\Sigma}, S) - \operatorname{Obj}(\widetilde{\boldsymbol{\Sigma}}, S)|$$

then $\tilde{S}$ will be consistent for S^* via the following argument:

$$\begin{aligned}
P(\tilde{S} = S^*) &= P(\operatorname{Obj}(\widetilde{\boldsymbol{\Sigma}}, S^*) < \min_{S \neq S^*} P(\operatorname{Obj}(\widetilde{\boldsymbol{\Sigma}}, S)) \\
&\geq P(\max_{S \subseteq [p] : |S|=k} \widetilde{\Delta}(S) < \frac{1}{2}(\min_{S \neq S^*} \operatorname{Obj}(\boldsymbol{\Sigma}, S) - \operatorname{Obj}(\boldsymbol{\Sigma}, S^*))) \\
&= 1 - o(1)
\end{aligned}$$

Considering an arbitrary size-k subset S, we fix $t > 0$ and devote the rest of the argument to showing that

$$P(\widetilde{\Delta}(S) > t) = o\left(\binom{p}{k}^{-1}\right). \tag{A.7}$$

Then, the fact that $\max_{S\subseteq[p]:|S|=k} \widetilde{\Delta}(S) = o_p(1)$ follows trivially from a union bound.

We start by re-writing $\widetilde{\Delta}(S)$. For two square positive semi-definite matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, von Neumann's trace inequality [Mirsky, 1975] implies that $|\mathrm{Tr}(\boldsymbol{AB})| \leq |\mathrm{Tr}(\boldsymbol{A})||\mathrm{Tr}(\boldsymbol{B})|$. Then, for four square positive semi-definite matrices $\widetilde{\boldsymbol{A}}$, $\widetilde{\boldsymbol{B}}$, $\boldsymbol{A}$, and $\boldsymbol{B}$, we have that

$$\begin{aligned}
|\mathrm{Tr}(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{B}} - \boldsymbol{AB})| &= |\mathrm{Tr}(\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{B}} - \widetilde{\boldsymbol{A}}\boldsymbol{B} + \widetilde{\boldsymbol{A}}\boldsymbol{B} - \boldsymbol{AB})| \\
&\leq |\mathrm{Tr}(\widetilde{\boldsymbol{A}}(\widetilde{\boldsymbol{B}} - \boldsymbol{B}))| + |\mathrm{Tr}((\widetilde{\boldsymbol{A}} - \boldsymbol{A})\boldsymbol{B})| \\
&\leq |\mathrm{Tr}(\widetilde{\boldsymbol{A}})| \cdot |\mathrm{Tr}(\widetilde{\boldsymbol{B}} - \boldsymbol{B})| + |\mathrm{Tr}(\widetilde{\boldsymbol{A}} - \boldsymbol{A})| \cdot |\mathrm{Tr}(\boldsymbol{B})|.
\end{aligned}$$

Using this fact we can reconfigure $\widetilde{\Delta}(S)$

$$\begin{aligned}
&\widetilde{\Delta}(S) \\
&= p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}} - \widetilde{\boldsymbol{\Sigma}}_{\bullet S}\widetilde{\boldsymbol{\Sigma}}_S^+\widetilde{\boldsymbol{\Sigma}}_{S\bullet}) - \mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet S}\boldsymbol{\Sigma}_S^+\boldsymbol{\Sigma}_{S\bullet})| \\
&\leq p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})| + p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{S\bullet}\widetilde{\boldsymbol{\Sigma}}_{\bullet S}\widetilde{\boldsymbol{\Sigma}}_S^+ - \boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S}\boldsymbol{\Sigma}_S^+)| \\
&\leq p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})| + p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{S\bullet}\widetilde{\boldsymbol{\Sigma}}_{\bullet S})| \cdot |\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^+ - \boldsymbol{\Sigma}_S^+)| + p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{S\bullet}\widetilde{\boldsymbol{\Sigma}}_{\bullet S} - \boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\boldsymbol{\Sigma}_S^+)| \\
&\leq p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})| + p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{S\bullet}\widetilde{\boldsymbol{\Sigma}}_{\bullet S} - \boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^+ - \boldsymbol{\Sigma}_S^+)| \\
&\quad + p^{-1/2-\delta}|\mathrm{Tr}(\boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^+ - \boldsymbol{\Sigma}_S^+)| + p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{S\bullet}\widetilde{\boldsymbol{\Sigma}}_{\bullet S} - \boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\boldsymbol{\Sigma}_S^+)|
\end{aligned}$$

We can further bound some of these terms. We know from Assumption 1 that $\max_{i\in[p]} \boldsymbol{\Sigma}_{ii} \leq c_1$ and therefore also $\max_{i\in[p],j\in[p]} |\boldsymbol{\Sigma}_{ij}| \leq c_1$, so $|\mathrm{Tr}(\boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S})| \leq kp \max_{i\in[p],j\in[p]} \boldsymbol{\Sigma}_{ij}^2 \leq c_2 p$. We also know from Assumption 2 that $|\mathrm{Tr}(\boldsymbol{\Sigma}_S^+)| \leq k\|\boldsymbol{\Sigma}_S^{-1}\|_2 \leq c_3$. Hence, we have

$$\begin{aligned}
\widetilde{\Delta}(S) \leq \underbrace{p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})|}_{\widetilde{\Delta}_1(S)} &+ p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{S\bullet}\widetilde{\boldsymbol{\Sigma}}_{\bullet S} - \boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^+ - \boldsymbol{\Sigma}_S^+)| \\
&+ c_2 \underbrace{p^{1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^+ - \boldsymbol{\Sigma}_S^+)|}_{\widetilde{\Delta}_3(S)} + c_3 \underbrace{p^{-1/2-\delta}|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_{S\bullet}\widetilde{\boldsymbol{\Sigma}}_{\bullet S} - \boldsymbol{\Sigma}_{S\bullet}\boldsymbol{\Sigma}_{\bullet S})|}_{\widetilde{\Delta}_2(S)}
\end{aligned}$$

If we provide appropriate tail bounds for the $\widetilde{\Delta}_i(S)$, i.e. show that for any $t > 0$ that

$$P(\widetilde{\Delta}_i(S) > t) = o\left(\binom{p}{k}^{-1}\right),$$

where the right-hand-side bound is independent of S. Then it is straightforward to establish the tail bound (A.7) for $\widetilde{\Delta}(S)$. We devote the rest of the argument to establishing this tail bound for $\widetilde{\Delta}_1(S)$, $\widetilde{\Delta}_2(S)$, and $\widetilde{\Delta}_3(S)$.

A.4.2 Establishing tail bounds

Tail bound for $\widetilde{\Delta}_1(S)$

We start by getting a concentration bound for the entries of $\widetilde{\mathbf{\Sigma}}_{ij}$. Using the Orlicz norm notation from Section 2.7.1 of Vershynin [2018], Assumption 1 implies that $\|X_i\|_{\psi_2} \leq c_4$. We can use [Vershynin, 2018, Lemma 2.7.7] to see that $\|X_iX_j\|_{\psi_1} \leq \|X_i\|_{\psi_2}\|X_j\|_{\psi_2} \leq c_5$ implying that all the X_iX_j are sub-exponential. For any $x \geq 0$, [Vershynin, 2018, Corollary 2.8.3] gives the following concentration bound for the entries $\widetilde{\mathbf{\Sigma}}_{ij}$:

$$P(|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| \geq x) \leq 2\exp\left(-c_6 \min\left(x^2, x\right) n\right) \tag{A.8}$$

Now, we use the union bound along with our sub-exponential concentration result

$$\begin{aligned}
P(\Delta_1(S) > t) &\leq P(p^{-1/2-\delta}\sum_{i=1}^{p}|\widetilde{\mathbf{\Sigma}}_{ii} - \mathbf{\Sigma}_{ii}| > t) && \text{(triangle inequality)}\\
&\leq \sum_{i=1}^{p} P\left(|\widetilde{\mathbf{\Sigma}}_{ii} - \mathbf{\Sigma}_{ii}| > p^{-1/2+\delta}t\right) && \text{(union bound)}\\
&\leq 2p\exp\left(-c_6\min\left(np^{-1+2\delta}t^2, np^{-1/2+\delta}t\right)\right) && \text{(sub-exponential concentration)}\\
&= o\left(\binom{p}{k}^{-1}\right)
\end{aligned}$$

Tail bound for $\widetilde{\Delta}_2(S)$

To get a tail bound for $\Delta_2(S)$ we need a concentration result for $\widetilde{\mathbf{\Sigma}}_{ij}^2$. Fix some constant $c_7 > 0$ and define $\widetilde{B}_{ij} = \max\{|2\mathbf{\Sigma}_{ij} + c_7|, |2\mathbf{\Sigma}_{ij} - c_7|\}$. Two things are true. First, if $|\widetilde{\mathbf{\Sigma}}_{ij} + \mathbf{\Sigma}_{ij}| > \widetilde{B}_{ij}$, then it must be the case that $|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > c_7$. Second $\widetilde{B}_{ij} \leq c_8$ for some constant. Considering some $x \geq 0$, we can now get a concentration result for $\widetilde{\mathbf{\Sigma}}_{ij}^2$.

$$\begin{aligned}
&P\left(|\widetilde{\mathbf{\Sigma}}_{ij}^2 - \mathbf{\Sigma}_{ij}^2| > x\right)\\
&= P\left(|\widetilde{\mathbf{\Sigma}}_{ij}^2 - \mathbf{\Sigma}_{ij}^2| > x, |\widetilde{\mathbf{\Sigma}}_{ij} + \mathbf{\Sigma}_{ij}| \leq \widetilde{B}_{ij}\right) + P\left(|\widetilde{\mathbf{\Sigma}}_{ij}^2 - \mathbf{\Sigma}_{ij}^2| > x, |\widetilde{\mathbf{\Sigma}}_{ij} + \mathbf{\Sigma}_{ij}| > \widetilde{B}_{ij}\right)\\
&\leq P\left(|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > x/\widetilde{B}_{ij}\right) + P\left(|\widetilde{\mathbf{\Sigma}}_{ij} + \mathbf{\Sigma}_{ij}| > \widetilde{B}_{ij}\right)\\
&\leq P\left(|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > x/c_8\right) + P\left(|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > c_7\right) && \text{(definition of } \widetilde{B}_{ij})\\
&\leq 2\exp(-c_9\min\left(x^2, x\right) n) + 2\exp(-c_{10}n) && \text{(sub-exponential concentration)}
\end{aligned}$$

With this result under our belt, getting a tail bound for $\widetilde{\Delta}_2(S)$ is straight-forward.

$$
\begin{aligned}
&P(\widetilde{\Delta}_2(S) > t) \\
&\leq P(p^{-1/2-\delta} \sum_{i\in S}\sum_{j=1}^{p} |\widetilde{\boldsymbol{\Sigma}}_{ij}^2 - \boldsymbol{\Sigma}_{ij}^2| > t) && \text{(triangle inequality)} \\
&= \sum_{i\in S}\sum_{j=1}^{p} P\left(|\widetilde{\boldsymbol{\Sigma}}_{ij}^2 - \boldsymbol{\Sigma}_{ij}^2| > k^{-1}p^{-1/2+\delta}t\right) && \text{(union bound)} \\
&\leq 2pk\exp\left(-c_{11}\min\left(np^{-1+2\delta}t^2, np^{-1/2+\delta}\right)\right) + 2pk\exp(-c_{10}n) && \text{(concentration for } \widetilde{\boldsymbol{\Sigma}}_{ij}^2\text{)} \\
&= o\left(\binom{p}{k}^{-1}\right)
\end{aligned}
$$

Tail bound for $\widetilde{\Delta}_3(S)$

We will get a tail bound for $\Delta_3(S)$ via a concentration inequality for the singular values of $\boldsymbol{X}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2} \in \mathbb{R}^{n\times k}$, where $\boldsymbol{X}_{\bullet S} \in \mathbb{R}^{n\times k}$ is random matrix with i.i.d samples of X_S as rows. Essentially, concentration of these singular values this will imply concentration of the eigenvalues of $\widetilde{\boldsymbol{\Sigma}}_S^{-1}$.

Using Assumption 1 and Assumption 2, we can show that the rows of $\boldsymbol{X}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2} \in \mathbb{R}^{n\times k}$ are isotropic sub-gaussian random vectors with bounded sub-gaussian norm:

$$
\begin{aligned}
\|\boldsymbol{\Sigma}_S^{-1/2}X_S\|_{\psi_2} &= \sup_{\|v\|_2=1} \|(\boldsymbol{\Sigma}_S^{-1/2}X_S)^\top v\|_{\psi_2} \\
&= \sup_{\|v\|_2=1} \left(\left\| X_S^\top \left(\frac{\boldsymbol{\Sigma}_S^{-1/2}v}{\|\boldsymbol{\Sigma}_S^{-1/2}v\|_2} \right) \right\|_{\psi_2} \|\boldsymbol{\Sigma}_S^{-1/2}v\|_2 \right) \\
&\leq \|\boldsymbol{\Sigma}_S^{-1/2}\|_2 \sup_{\|\tilde{v}\|_2=1} \|\tilde{v}^\top X_S\|_{\psi_2} \\
&\leq \|\boldsymbol{\Sigma}_S^{-1}\|_2^{1/2} \sup_{\|\tilde{v}\|_2=1} \sum_{i=1}^{k} |v_i| \|X_S\|_{\psi_2} \\
&\leq \|\boldsymbol{\Sigma}_S^{-1}\|_2^{1/2} \max_{i\in[p]} \|X_i\|_{\psi_2} \sup_{\|\tilde{v}\|_2=1} \|\tilde{v}\|_1 \\
&\leq c_{12}.
\end{aligned}
$$

Note that in the above $v, \tilde{v} \in \mathbb{R}^k$ have fixed (non-growing) dimension because k is fixed. [Vershynin, 2018, Theorem 4.6.1], then tells us that there is a constant c_{13} such that for any $x \geq c_{13}$,

$$
P\left(\max_{i\in[k]} |s_i(\boldsymbol{X}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - n^{1/2}| \geq x\right) \leq 2\exp(-c_{14}(x - c_{13})^2), \tag{A.9}
$$

To apply this result we need to note two facts. First, whenever the first k singular values of $\boldsymbol{X}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}$ are positive, that means that $\boldsymbol{\Sigma}_S^{-1/2}\widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{1/2}$ is invertible, so $\widetilde{\boldsymbol{\Sigma}}_S$ must be invertible. Second, whenever $\widetilde{\boldsymbol{\Sigma}}_S$ is

invertible,

$$\begin{aligned}
|\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^{+} - \boldsymbol{\Sigma}_S^{+})| &= |\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^{-1} - \boldsymbol{\Sigma}_S^{-1})| \\
&= |\mathrm{Tr}((\widetilde{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S - \boldsymbol{I}_k)\boldsymbol{\Sigma}_S^{-1})| \\
&\leq |\mathrm{Tr}(\widetilde{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S - \boldsymbol{I}_k)||\mathrm{Tr}(\boldsymbol{\Sigma}_S^{-1})| \\
&\leq k\|\boldsymbol{\Sigma}_S^{-1}\|_2|\mathrm{Tr}(\boldsymbol{\Sigma}_S^{1/2}\widetilde{\boldsymbol{\Sigma}}_S^{-1}\boldsymbol{\Sigma}_S^{1/2} - \boldsymbol{I}_k)| \\
&\leq k\|\boldsymbol{\Sigma}_S^{-1}\|_2|\mathrm{Tr}((\boldsymbol{\Sigma}_S^{-1/2}\widetilde{\boldsymbol{\Sigma}}_S\boldsymbol{\Sigma}_S^{-1/2})^{-1} - \boldsymbol{I}_k)| \\
&\leq k^2\|\boldsymbol{\Sigma}_S^{-1}\|_2|\max_{i\in[k]}|s_i^{-2}(n^{-1/2}\boldsymbol{X}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| \\
&\leq c_{15}\max_{i\in[k]}|s_i^{-2}(n^{-1/2}\boldsymbol{X}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1|
\end{aligned}$$

where we have again made use of Assumption 2. To finalize our argument, we will need to make use of the following lemma.

Lemma A.4.1. *Fix some $a \in (0,1)$. There exists some $b > 0$ such that for all $x > 0$*

$$|x-1| < a \implies b|x-1| \geq |x^{-2} - 1| \tag{A.10}$$

Proof. If $x = 1$ the right hand inequality is always true, so consider $x \neq 1$. If $|x-1| < a$, then

$$\begin{aligned}
\frac{|x^{-2}-1|}{|x-1|} &= \frac{|1-x^2|}{x^2|x-1|} \\
&= \frac{1+x}{x^2} \\
&\leq \frac{2+a}{(1-a)^2}
\end{aligned}$$

So we can take $b = (2+a)/(1-a)^2$. □

Now we can prove the tail bound for $\widetilde{\Delta}_3(S)$. Fix some $c_{16} \in (0,1)$. Then for large enough n,

$$
\begin{aligned}
&P\left(\widetilde{\Delta}_3 > t\right) \\
&\le P\left(p^{1/2-\delta}|\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_S^+ - \mathbf{\Sigma}_S^+)| > t, \max_{i\in[k]}|s_i(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| < c_{16}\right) \\
&\quad + P\left(\max_{i\in[k]}|s_i(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| \ge c_{16}\right) \\
&\le P\left(\max_{i\in[k]}|s_i^{-2}(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| > c_{17}p^{-1/2+\delta}t, \max_{i\in[k]}|s_i(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| < c_{16}\right) \\
&\quad + P\left(\max_{i\in[k]}|s_i(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| \ge c_{16}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\widetilde{\mathbf{\Sigma}}_S \text{ is invertible on the event}) \\
&\le P\left(\max_{i\in[k]}|s_i(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| > c_{18}p^{-1/2+\delta}t, \max_{i\in[k]}|s_i(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| < c_{16}\right) \\
&\quad + P\left(\max_{i\in[k]}|s_i(n^{-1/2}\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - 1| \ge c_{16}\right) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Lemma A.4.1 with } a = c_{16}) \\
&\le P\left(\max_{i\in[k]}|s_i(\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - n^{1/2}| > c_{18}n^{1/2}p^{-1/2+\delta}t\right) + P\left(\max_{i\in[k]}|s_i(\boldsymbol{X}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - n^{1/2}| \ge n^{1/2}c_{16}\right) \\
&\le 2\exp(-c_{19}(n^{1/2}p^{-1/2+\delta}t - c_{20})^2) + 2\exp(-c_{21}(n^{1/2} - c_{22})^2) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(apply Equation (A.9) once } n \text{ large enough)} \\
&= o_p\left(\binom{p}{k}^{-1}\right).
\end{aligned}
$$

A.4.3 Estimation with mean

Now we allow $X \sim (\mu, \mathbf{\Sigma}), \mathbf{\Sigma} \in \mathbb{S}_+^{p\times p}$ to have a non-zero mean. Now with

$$\widetilde{\mathbf{\Sigma}} = n^{-1}\sum_{\ell=1}^{n}(X_i^{(\ell)} - \mu)(X_i^{(\ell)} - \mu)^T,$$

we still let $\widetilde{S}$ be a minimizer of $\mathrm{Obj}(\widetilde{\mathbf{\Sigma}}, S)$ (with ties broken randomly). Also still we let S^* be the (eventually unique) minimizer of $\mathrm{Obj}(\mathbf{\Sigma}, S)$ for size-k subsets S. Now we define $\hat{S}$ to be the minimizer of $\mathrm{Obj}(\hat{\mathbf{\Sigma}}, S)$ (with ties broken randomly) where

$$\widetilde{\mathbf{\Sigma}} = n^{-1}\sum_{\ell=1}^{n}(X_i^{(\ell)} - \hat{\mu})(X_i^{(\ell)} - \hat{\mu})^T, \qquad \hat{\mu} = n^{-1}\sum_{\ell=1}^{n}X_i^{(\ell)},$$

and we want to show that $\hat{S}$ is consistent for S^*.

By our earlier arguments, we know that

$$\max_{S\subseteq[p]:|S|=k}\widetilde{\Delta}(S) = o_p(1), \qquad \widetilde{\Delta}(S) = p^{-1/2+\delta}|\mathrm{Obj}(\widetilde{\mathbf{\Sigma}}, S) - \mathrm{Obj}(\mathbf{\Sigma}, S)|$$

If we can show that also

$$\max_{S\subseteq[p]:|S|=k} \widehat{\Delta}(S) = o_p(1), \qquad \widehat{\Delta}(S) = p^{-1/2+\delta}|\mathrm{Obj}(\hat{\mathbf{\Sigma}}, S) - \mathrm{Obj}(\widetilde{\mathbf{\Sigma}}, S)|.$$

then we have consistency, because

$$\begin{aligned}
&P(\hat{S} = S^*) \\
&= P(\mathrm{Obj}(\hat{\mathbf{\Sigma}}, S^*) < \min_{S\neq S^*} P(\mathrm{Obj}(\hat{\mathbf{\Sigma}}, S)) \\
&\geq P\left(\max_{S\subseteq[p]:|S|=k} \widehat{\Delta}(S) < \frac{1}{4}(\min_{S\neq S^*} \mathrm{Obj}(\mathbf{\Sigma}, S) - \mathrm{Obj}(\mathbf{\Sigma}, S^*)), \max_{S\subseteq[p]:|S|=k} \widetilde{\Delta}(S) < \frac{1}{4}(\min_{S\neq S^*} \mathrm{Obj}(\mathbf{\Sigma}, S) - \mathrm{Obj}(\mathbf{\Sigma}, S^*))\right) \\
&= 1 - o(1)
\end{aligned}$$

Again, we consider an arbitrary size-k subset S, fix $t > 0$, and devote the rest of the argument to showing that

$$P(\widehat{\Delta}(S) > t) = o\left(\binom{p}{k}^{-1}\right).$$

Then, the fact that $\max_{S\subseteq[p]:|S|=k} \widehat{\Delta}(S) = o_p(1)$ follows trivially from a union bound.

Again we reconfigure $\widehat{\Delta}(S)$ and bound it, similar to how we did before.

$$\begin{aligned}
&\widehat{\Delta}(S) \\
&= p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_{\bullet S}\hat{\mathbf{\Sigma}}_S^+\hat{\mathbf{\Sigma}}_{S\bullet}) - \mathrm{Tr}(\widetilde{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}}_{\bullet S}\widetilde{\mathbf{\Sigma}}_S^+\widetilde{\mathbf{\Sigma}}_{S\bullet})| \\
&\leq p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}})| + p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S}\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S}\widetilde{\mathbf{\Sigma}}_S^+)| \\
&\leq p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}})| + p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S})| \cdot |\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)| + p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S} - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S})| \cdot |\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_S^+)| \\
&\leq p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}})| + p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S} - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S})| \cdot |\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)| \\
&\quad + p^{-1/2+\delta}|\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S} - \mathbf{\Sigma}_{S\bullet}\mathbf{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)| + p^{-1/2-\delta}|\mathrm{Tr}(\mathbf{\Sigma}_{S\bullet}\mathbf{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)| \\
&\quad + p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S} - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S})| \cdot |\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_S^+ - \mathbf{\Sigma}_S^+)| + p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S} - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S})| \cdot |\mathrm{Tr}(\mathbf{\Sigma}_S^+)| \\
&\leq \underbrace{p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}} - \widetilde{\mathbf{\Sigma}})|}_{\widehat{\Delta}_1(S)} + \underbrace{p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S} - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S})|}_{\widetilde{\Delta}_2(S)} \cdot |\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)| \\
&\quad + p^{-1/2+\delta}|\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S} - \mathbf{\Sigma}_{S\bullet}\mathbf{\Sigma}_{\bullet S})| \cdot |\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)| + c_2\underbrace{p^{1/2+\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)|}_{\widehat{\Delta}_3(S)} \\
&\quad + p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S} - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S})| \cdot \underbrace{|\mathrm{Tr}(\widetilde{\mathbf{\Sigma}}_S^+ - \mathbf{\Sigma}_S^+)|}_{\widetilde{\Delta}_3(S)} + c_3\underbrace{p^{-1/2-\delta}|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{S\bullet}\hat{\mathbf{\Sigma}}_{\bullet S} - \widetilde{\mathbf{\Sigma}}_{S\bullet}\widetilde{\mathbf{\Sigma}}_{\bullet S})|}_{\widehat{\Delta}_2(S)}
\end{aligned}$$

Our above results already imply tail bounds for the $\widetilde{\Delta}_i(S)$, i.e., we know for any $t > 0$ that

$$P(\widetilde{\Delta}_i(S) > t) = o\left(\binom{p}{k}^{-1}\right),$$

We know show the same tail bounds for the $\widehat{\Delta}_i(S)$, i.e., we show for any $t > 0$ that

$$P(\widehat{\Delta}_i(S) > t) = o\left(\binom{p}{k}^{-1}\right),$$

It is then straightforward to establish the required tail bound (A.7) for $\widehat{\Delta}(S)$.

Prior to proceeding, we note it is easy to verify that

$$\widetilde{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}} = \mu(\mu - \hat{\mu})^T + \hat{\mu}(\hat{\mu} - \mu)^T \implies \widetilde{\mathbf{\Sigma}}_{ij} - \hat{\mathbf{\Sigma}}_{ij} = (\hat{\mu}_i - \mu_i)(\hat{\mu}_j - \mu_j).$$

Tail bound for $\widehat{\Delta}_1(S)$

By Assumption 1, the X_i are sub-Guassian and $\max_i \|X_i\|_{\psi_2}$ is bounded by some constant c_{23}. Therefore, [Vershynin, 2018, Theorem 2.6.3] gives us sub-Gaussian concentration, i.e., for any $x > 0$:

$$P(|\hat{\mu}_i - \mu_i| \geq x) \leq 2\exp\left(-c_{24}t^2 n\right).$$

We can then get our tail bound by applying a union bound.

$$\begin{aligned}
P(\widehat{\Delta}_1(S) > t) &= P(p^{-1/2-\delta}\sum_{i=1}^{p}(\hat{\mu}_i - \mu_i)^2 > t) \\
&\leq \sum_{i=1}^{p} P\left(|\hat{\mu}_i - \mu| > (tp^{-1/2+\delta})^{\frac{1}{2}}\right) && \text{(union bound)} \\
&\leq 2p\exp\left(-c_{24}tnp^{-1/2+\delta}\right) && \text{(sub-Gaussian concentration)} \\
&= o\left(\binom{p}{k}^{-1}\right).
\end{aligned}$$

Tail bound for $\widehat{\Delta}_2(S)$

As was the case for $\widetilde{\Delta}_2(S)$, it suffices to examine the different $\hat{\mathbf{\Sigma}}_{ij}^2 - \widetilde{\mathbf{\Sigma}}_{ij}^2$. Fix some constant $c_{25} > 0$ and define $\widehat{B}_{ij} = \max\{2\mathbf{\Sigma}_{ij} + c_{25}, 2\mathbf{\Sigma}_{ij} - c_{25}\}$. Two things are true. First, rewriting

$$\hat{\mathbf{\Sigma}}_{ij} + \widetilde{\mathbf{\Sigma}}_{ij} = 2\mathbf{\Sigma}_{ij} + 2(\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}) + (\hat{\mathbf{\Sigma}}_{ij} - \widetilde{\mathbf{\Sigma}}_{ij})$$

makes it clear that $|\hat{\mathbf{\Sigma}}_{ij} + \widetilde{\mathbf{\Sigma}}_{ij}| > \widehat{B}_{ij}$ implies that either $2|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > c_{25}/2$ or $|\hat{\mathbf{\Sigma}}_{ij} - \widetilde{\mathbf{\Sigma}}_{ij}| > c_{25}/2$. Second, $\widehat{B}_{ij} \leq c_{26}$ for some constant. Thus for any $x \geq 0$,

$$
\begin{aligned}
&P\left(|\hat{\mathbf{\Sigma}}_{ij}^2 - \widetilde{\mathbf{\Sigma}}_{ij}^2| > x\right) \\
&= P\left(|\hat{\mathbf{\Sigma}}_{ij}^2 - \widetilde{\mathbf{\Sigma}}_{ij}^2| > x, |\hat{\mathbf{\Sigma}}_{ij} + \widetilde{\mathbf{\Sigma}}_{ij}| \le \widehat{B}_{ij}\right) + P\left(|\hat{\mathbf{\Sigma}}_{ij}^2 - \widetilde{\mathbf{\Sigma}}_{ij}^2| > x, |\hat{\mathbf{\Sigma}}_{ij} + \widetilde{\mathbf{\Sigma}}_{ij}| > \widehat{B}_{ij}\right) \\
&\le P\left(|\hat{\mathbf{\Sigma}}_{ij} - \widetilde{\mathbf{\Sigma}}_{ij}| > x/\widehat{B}_{ij}\right) + P\left(|\hat{\mathbf{\Sigma}}_{ij} + \widetilde{\mathbf{\Sigma}}_{ij}| > \widehat{B}_{ij}\right) \\
&\le P\left(|\hat{\mathbf{\Sigma}}_{ij} - \widetilde{\mathbf{\Sigma}}_{ij}| > x/c_{26}\right) + P\left(|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > c_{25}/4\right) + P\left(|\hat{\mathbf{\Sigma}}_{ij} - \widetilde{\mathbf{\Sigma}}_{ij}| > c_{25}/2\right) \\
&\qquad\qquad \text{(definition of } \widehat{B}_{ij} \text{ and union bound)} \\
&= P\left(|\hat{\mu}_i - \mu_i| \cdot |\hat{\mu}_j - \mu_j| > x/c_{26}\right) + P\left(|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > c_{25}/4\right) + P\left(|\hat{\mu}_i - \mu_i| \cdot |\hat{\mu}_j - \mu_j| > c_{25}/2\right) \\
&\le P(|\hat{\mu}_i - \mu_i| > (x/c_{26})^{1/2}) + P(|\hat{\mu}_j - \mu_j| > (x/c_{26})^{1/2}) + P\left(|\widetilde{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > c_{25}/4\right) \\
&\qquad + P\left(|\hat{\mu}_i - \mu_i| > (c_{25}/2)^{1/2}\right) + P\left(|\hat{\mu}_j - \mu_j| > (c_{25}/2)^{1/2}\right) \\
&\qquad\qquad \text{(union bound)} \\
&\le 4\exp(-c_{27}xn) + 5\exp(-c_{28}n) \\
&\qquad\qquad \text{(sub-Gaussian and sub-exponential concentration)}
\end{aligned}
$$

We can finish the argument the same way as we did for $\widetilde{\Delta}_2(S)$.

$$
\begin{aligned}
&P(\widehat{\Delta}_2(S) > t) \\
&\le P(p^{-1/2-\delta} \sum_{i \in S} \sum_{j=1}^{p} |\hat{\mathbf{\Sigma}}_{ij}^2 - \widetilde{\mathbf{\Sigma}}_{ij}^2| > t) && \text{(triangle inequality)} \\
&= \sum_{i \in S} \sum_{j=1}^{p} P\left(|\hat{\mathbf{\Sigma}}_{ij}^2 - \widetilde{\mathbf{\Sigma}}_{ij}^2| > k^{-1} p^{-1/2+\delta} t\right) && \text{(union bound)} \\
&\le 4pk\exp(-c_{29}tnp^{-1/2+\delta}) + 5pk\exp(-c_{30}n) && \text{(concentration for } \hat{\mathbf{\Sigma}}_{ij}^2\text{)} \\
&= o\left(\binom{p}{k}^{-1}\right)
\end{aligned}
$$

Tail bound for $\widehat{\Delta}_3(S)$

We first develop the following bound, which holds whenever both $\hat{\mathbf{\Sigma}}_S$ and $\widetilde{\mathbf{\Sigma}}_S$ are invertible.

$$
\begin{aligned}
|\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^+ - \widetilde{\mathbf{\Sigma}}_S^+)| &= |\mathrm{Tr}(\hat{\mathbf{\Sigma}}_S^{-1} - \widetilde{\mathbf{\Sigma}}_S^{-1})| \\
&= |\mathrm{Tr}(\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_S^{1/2}(\hat{\mathbf{\Sigma}}_S^{-1} - \widetilde{\mathbf{\Sigma}}_S^{-1})\mathbf{\Sigma}_S^{1/2})| \\
&\le k\|\mathbf{\Sigma}_S^{-1}\|_2 \cdot \left|\mathrm{Tr}\left((\mathbf{\Sigma}_S^{-1/2}\widetilde{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2} + \mathbf{\Sigma}_S^{-1/2}(\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S)\mathbf{\Sigma}_S^{-1/2})^{-1} - (\mathbf{\Sigma}_S^{-1/2}\hat{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2})^{-1}\right)\right| \\
&\le c_{31}\left|\mathrm{Tr}\left((\mathbf{\Sigma}_S^{-1/2}\widetilde{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2} + \mathbf{\Sigma}_S^{-1/2}(\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S)\mathbf{\Sigma}_S^{-1/2})^{-1} - (\mathbf{\Sigma}_S^{-1/2}\hat{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2})^{-1}\right)\right|
\end{aligned}
$$

Crucially, we can show that $\|\mathbf{\Sigma}_S^{-1/2}(\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S)\mathbf{\Sigma}_S^{-1/2}\|_2$ concentrates to zero. For any $x \geq 0$, we have

$$
\begin{aligned}
&P(\|\mathbf{\Sigma}_S^{-1/2}(\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S)\mathbf{\Sigma}_S^{-1/2}\|_2 \geq x) \\
&\leq P(\|\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S\|_2 \geq \|\mathbf{\Sigma}_S^{-1}\|_2^{-1}x) \\
&\leq P(\|\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S\|_F \geq \|\mathbf{\Sigma}_S^{-1}\|_2^{-1}x) \\
&\leq \sum_{i\in S}\sum_{j\in S} P(|\hat{\mu}_i - \mu_i| \cdot |\hat{\mu}_j - \mu_j| \geq k^{-2}\|\mathbf{\Sigma}_S^{-1}\|_2^{-2}x^2) && \text{(triangle inequality + union bound)} \\
&\leq \sum_{i\in S}\sum_{j\in S} P(|\hat{\mu}_i - \mu_i| \geq k^{-1}\|\mathbf{\Sigma}_S^{-1}\|_2^{-1}x) + P(|\hat{\mu}_j - \mu_j| \geq k^{-1}\|\mathbf{\Sigma}_S^{-1}\|_2^{-1}x) && \text{(union bound)} \\
&\leq 4\exp(-c_{32}x^2 n) && \text{(sub-Gaussian concentration)}
\end{aligned}
$$

Henceforth we refer to $\|\mathbf{\Sigma}_S^{-1/2}(\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S)\mathbf{\Sigma}_S^{-1/2}\|_2$ as $\boldsymbol{E}$. We will need the following lemma.

Lemma A.4.2. *Fix some $a > 0$. There exists $b > 0$ such that for all $x, y \geq c$*

$$|x^{-1} - y^{-1}| \leq b|x - y| \tag{A.11}$$

Proof.

$$|x^{-1} - y^{-1}| = \frac{|x-y|}{xy} \leq \frac{1}{c^2}|x - y|$$

So one can take $b = 1/c^2$. □

The Bauer-Fike theorem [Bauer and Fike, 1960] tells us that every eigenvalue of $\mathbf{\Sigma}_S^{-1/2}\widetilde{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2} + \boldsymbol{E}$ is within $\|\boldsymbol{E}\|_2$ of an eigenvalue of $\mathbf{\Sigma}_S^{-1/2}\widetilde{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2}$. Note that because $\mathbf{\Sigma}_S$ is always full-rank, whenever $\mathbf{\Sigma}_S^{-1/2}\widetilde{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2} + \boldsymbol{E}$ is invertible, so is $\hat{\mathbf{\Sigma}}_S$. Using Bauer-Fike, we find whenever $\widetilde{\mathbf{\Sigma}}_S$ and $\hat{\mathbf{\Sigma}}_S$ are both invertible that

$$
\begin{aligned}
&\left|\mathrm{Tr}\left((\mathbf{\Sigma}_S^{-1/2}\widetilde{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2} + \mathbf{\Sigma}_S^{-1/2}(\hat{\mathbf{\Sigma}}_S - \widetilde{\mathbf{\Sigma}}_S)\mathbf{\Sigma}_S^{-1/2})^{-1} - (\mathbf{\Sigma}_S^{-1/2}\hat{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2})^{-1}\right)\right| \\
&\leq k\left(\frac{1}{\lambda_k(\mathbf{\Sigma}_S^{-1/2}\hat{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2}) - \|\boldsymbol{E}\|_2} - \frac{1}{\lambda_1(\mathbf{\Sigma}_S^{-1/2}\hat{\mathbf{\Sigma}}_S\mathbf{\Sigma}_S^{-1/2}) + \|\boldsymbol{E}\|_2}\right) && \text{(Bauer-Fike)} \\
&= k\left(\frac{1}{s_k^2(\widetilde{\boldsymbol{X}}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) - \|\boldsymbol{E}\|_2} - \frac{1}{s_1^2(\widetilde{\boldsymbol{X}}_{\bullet S}\mathbf{\Sigma}_S^{-1/2}) + \|\boldsymbol{E}\|_2}\right)
\end{aligned}
$$

where $\widetilde{\boldsymbol{X}}_{\bullet S} = \boldsymbol{X}_{\bullet S} - 1\mu_S^T$.

Now we can get our tail bound. Fix some $c_{33} \in (0,1)$ and also some $c_{34} \in (0, (1-c_{33})^2)$. We see for large enough n that

$$
\begin{aligned}
&P\left(\widehat{\Delta}_3 > t\right) \\
&\leq P\left(p^{1/2-\delta}|\mathrm{Tr}(\widehat{\boldsymbol{\Sigma}}_S^{+} - \widehat{\boldsymbol{\Sigma}}_S^{+})| > t, \max_{i\in[k]}|s_i(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| \leq c_{33}, \|\boldsymbol{E}\|_2 \leq c_{34}\right) \\
&\quad + P(\max_{i\in[k]}|s_i(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| > c_{33}) + P(|\boldsymbol{E}\|_2 > c_{34}) \\
&\hfill \text{(union bound)} \\
&\leq P\left(|s_1^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - s_k^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2})| + 2\|\boldsymbol{E}\|_2 > c_{35}k^{-1}p^{-1/2+\delta}t, \max_{i\in[k]}|s_i(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| \leq c_{33}, \|\boldsymbol{E}\|_2 \leq c_{34}\right) \\
&\quad + P(\max_{i\in[k]}|s_i(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - n^{1/2}| > c_{33}n^{1/2}) + P(|\boldsymbol{E}\|_2 > c_{34}) \\
&\hfill \text{(both } \hat{\boldsymbol{\Sigma}}_S \text{ and } \widetilde{\boldsymbol{\Sigma}}_S \text{ invertible on event, apply Lemma A.4.2)} \\
&\leq P\left(|s_1^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - s_k^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2})| > c_{36}p^{-1/2+\delta}t\right) \\
&\quad + P(\|\boldsymbol{E}\|_2 > c_{37}p^{-1/2+\delta}t) + P(\max_{i\in[k]}|s_i(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - n^{1/2}| > n^{1/2}c_{33}) + P(|\boldsymbol{E}\|_2 > c_{34}) \\
&\hfill \text{(union bound)} \\
&\leq P\left(|s_1^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - s_k^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2})| > c_{36}p^{-1/2+\delta}t\right) \\
&+ 4\exp(-c_{38}p^{-1+\delta}nt^2) + 2\exp(-c_{39}(n^{1/2} - c_{40})^2) + 4\exp(-c_{41}n) \\
&\hfill \text{(earlier concentration bounds once } n \text{ is large enough)}
\end{aligned}
$$

The last three terms are $o\left(\binom{p}{k}^{-1}\right)$, so all that remains to show is that the first term is as well. By a union bound

$$
\begin{aligned}
&P\left(|s_1^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - s_k^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2})| > c_{36}p^{-1/2+\delta}t\right) \\
&\leq P\left(|s_1^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| > c_{42}p^{-1/2+\delta}t\right) + P\left(|s_k^2(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1|) > c_{42}p^{-1/2+\delta}t\right)
\end{aligned}
$$

Thus, letting $c_{43} = \max\{|2 - c_{44}|, |2 + c_{44}|\}$, it suffices to see that

$$
\begin{aligned}
&P\left(\max_{i\in[k]} |s_i^2(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| > c_{42}p^{-1/2+\delta}t\right) \\
&= P\left(\max_{i\in[k]} |s_i^2(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| > c_{42}p^{-1/2+\delta}t, \max_{i\in[k]} |s_i(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) + 1| \le c_{43}\right) \\
&\quad + P\left(\max_{i\in[k]} |s_i(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) + 1| > c_{43}\right) \\
&\le P\left(\max_{i\in[k]} |s_i(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| > c_{45}p^{-1/2+\delta}t\right) + P\left(\max_{i\in[k]} |s_i(n^{-1/2}\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - 1| > c_{44}\right) \\
&\le P\left(\max_{i\in[k]} |s_i(\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - n^{1/2}| > c_{45}n^{1/2}p^{-1/2+\delta}t\right) + P\left(\max_{i\in[k]} |s_i(|\widetilde{\boldsymbol{X}}_{\bullet S}\boldsymbol{\Sigma}_S^{-1/2}) - n^{1/2}| > c_{44}n^{1/2}\right) \\
&\le 2\exp(-c_{46}(n^{1/2}p^{-1/2+\delta}t - c_{47})^2) + 2\exp(-c_{48}(n^{1/2} - c_{49})^2) \\
&\qquad\qquad \text{(Apply Equation (A.9) once } n \text{ large enough)} \\
&= o\left(\binom{p}{k}^{-1}\right)
\end{aligned}
$$

where we have applied the same technique as we did for $\widetilde{\Delta}_2(S)$.

A.5 Additional theoretical results

A.5.1 Estimation error of regression coefficients

In this section, we use the finite sample bounds we developed in Appendix A.4 to study the regression coefficients from the regression of X_{-U} on X_U, where U is some size-k subset. More concretely, we aim to measure the quality of the observed regression coefficients $\hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1}$ as an estimate of the population regression coefficients $\boldsymbol{\Sigma}_{-U,U}\boldsymbol{\Sigma}_U^{-1}$. We will use the 2-to-$\infty$ norm to measure our estimation error,

$$\|\hat{\boldsymbol{\Sigma}}_{-U,U}\hat{\boldsymbol{\Sigma}}_U^{-1} - \boldsymbol{\Sigma}_{-U,U}\boldsymbol{\Sigma}_U^{-1}\|_{2\to\infty} = \max_{j\notin U} \|\hat{\boldsymbol{\Sigma}}_U^{-1}\hat{\boldsymbol{\Sigma}}_{Uj} - \boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{Uj}\|_2.$$

Considering variables $j \notin U$, the 2-to-∞ norm captures the worst-case L^2 estimation error for the regression coefficients from the regression of X_j on X_U. Because our analysis comes from finite sample bounds, it applies even in the high-dimensional where the number of variables p and the sample size n are comparable in size. Along with providing a bound that holds for a specific subset U, we also provide a bound that holds uniformly over all size-k subsets. Because it is uniform, it is informative even for a subset that is chosen by a possibly complicated and non-deterministic CSS algorithm.

Proposition A.5.1 states our error control result. In Proposition A.5.1, we adopt two of the assumptions from Appendix A.4, this time also requiring that constants be independent of k:

Assumption 4 (Light-tailed data)**.** *The sub-Gaussian norm (see [Vershynin, 2018, Definition 2.5.6]) of each variable is bounded above by a constant independent of n, p, and k.*

Assumption 5 (Low-dimensional invertibility)**.** *The minimum of the smallest eigenvalues of all the $k \times k$ principal sub-matrices of $\boldsymbol{\Sigma}$ is bounded below by a positive constant independent of n, p, and k.*

Proposition A.5.1. *Consider observing samples $x^{(1)}, \dots, x^{(n)} \in \mathbb{R}^p$ from a mean-zero distribution $X \sim P$ satisfying Assumption 4 and Assumption 5, and let $\hat{\mathbf{\Sigma}} = n^{-1} \sum_{i=1}^n x^{(i)} (x^{(i)})^\top$ denote the sample covariance. For a size-k subset U, let M be the sub-Gaussian norm of X_U and set $B = \max_{j \notin U} \|\mathbf{\Sigma}_{Uj}\|_2$. Then for any $t \in (0,1)$,*

$$P(\|\hat{\mathbf{\Sigma}}_{-U,U} \hat{\mathbf{\Sigma}}_U^{-1} - \mathbf{\Sigma}_{-U,U} \mathbf{\Sigma}_U^{-1}\|_{2\to\infty} > t) \le c_1 pk \exp\left(- \left(\frac{c_2 t n^{1/2}}{k^{1/2} \max(BM^2, M^2, 1)} - 1 \right)_+^2 \right)$$

for some constants $c_1, c_2 > 0$ that are independent of n, p, k. As a consequence,

$$P\left(\max_{U \subseteq [p] : |U| = k} \|\hat{\mathbf{\Sigma}}_{-U,U} \hat{\mathbf{\Sigma}}_U^{-1} - \mathbf{\Sigma}_{-U,U} \mathbf{\Sigma}_U^{-1}\|_{2\to\infty} > t \right) \le c_1 p^{k+1} k \exp\left(- \left(\frac{c_3 t n^{1/2}}{k^2} - 1 \right)_+^2 \right),$$

where $c_3 > 0$ is another constant independent of n, p, k.

In Proposition A.5.1, the light-tail assumption is necessary to ensure that our estimates concentrate, and the invertibility assumption is necessary to ensure that the estimate $\hat{\mathbf{\Sigma}}_U^{-1}$ of $\mathbf{\Sigma}_U^{-1}$ is well-behaved. We assume mean-zero data for simplicity. By applying identical arguments to those in Appendix A.4.3, we can modify Proposition A.5.1 to accommodate data with non-zero mean.

While Proposition A.5.1 may appear hard to parse at first, what it tells us is quite simple: excluding log factors, the 2-to-∞ regression coefficient estimation error for a subset U scales (up to logarithmic factors) at worst as $\sqrt{k/n} \max(BM^2, M^2, 1)$, where M is the sub-Gaussian norm of X_U and $B = \max_{j \notin U} \|\Sigma_{Uj}\|_2$ is the maximum L^2 norm of the covariance between the variables U and variables $j \notin U$. If the B and M^2 can be bounded by constants that have no dependence on k, n, or p, then the estimation error scales (up to logarithmic factors) at worst as $\sqrt{k/n}$.

A situation where we may reasonably believe that the estimation error scales at worst (up to logarithmic factors) as $\sqrt{k/n}$ is if our data comes from the k-dimensional PCSS model (3.1) and we consider coefficient estimates for population subset S. It is not unreasonable to imagine that the variables in S have little dependence, as if the principal variables are highly correlated and explain redundant variation, a smaller and less dependent principal variable set may better model the data. Hence, we can reasonably restrict to the case where the sub-Gaussian norm of X_S and the maximum eigenvalue of $\mathbf{\Sigma}_S$ are bounded by a constant independent of n, p, and k. One simple case where this is true is when the principal variables are exactly independent (because of Assumption 4) . Henceforth, when we say bounded by a constant, we mean a constant independent of n, p and k. In this case, we have immediately that M is bounded by a constant. To deal with B, we note that by Assumption 4, the maximum variance of the variables in $-S$ is bounded above by a constant. Because of the covariance structure induced by the PCSS model, this means that $\max_{j \notin S} \mathbf{\Sigma}_{jS} \mathbf{\Sigma}_S^{-1} \mathbf{\Sigma}_{Sj} = \max_{j \notin S} \|\mathbf{\Sigma}_S^{-1/2} \mathbf{\Sigma}_{Sj}\|_2^2$ and therefore $\max_{j \notin S} \|\mathbf{\Sigma}_S^{-1/2} \mathbf{\Sigma}_{Sj}\|_2$ are also bounded above by a constants. Since

$$B = \max_{j \notin S} \|\mathbf{\Sigma}_{Sj}\|_2^2 \le \|\mathbf{\Sigma}_S^{1/2}\|_2 \max_{j \notin S} \|\mathbf{\Sigma}_S^{-1/2} \mathbf{\Sigma}_{Sj}\|_2 = \|\mathbf{\Sigma}_S\|_2^{1/2} \max_{j \notin S} \|\mathbf{\Sigma}_S^{-1/2} \mathbf{\Sigma}_{Sj}\|_2,$$

and $\|\mathbf{\Sigma}_S\|_2^{1/2}$ is bounded above by a constant, so is B. Hence, if the principal variables are sufficiently non-dependent, we can reasonably imagine that the regression coefficient error for the principal variable set S (up to logarithmic factors) scales at worst as $\sqrt{k/n}$.

In general, however, for a set of more correlated variables U, the sub-Gaussian norm of X_U or the maximum eigenvalue of $\mathbf{\Sigma}_U$ can grow with k. Still, even in the most general case, it is straightforward to use Assumption 4 to bound $M \leq c_4\sqrt{k}$ and $B \leq c_4\sqrt{k}$ for some constant $c_4 > 0$. Hence we still find that the estimation error (up to logarithmic factors) scales at worst as $k^2/\sqrt{n}$. In fact, Proposition A.5.1 tells us that this error bound holds uniformly over all size-k subsets, so it applies even for a subset selected via a complex and possibly non-deterministic CSS algorithm.

Since k is typically much smaller than n and p, the estimation error is quite small in both cases, even when the number of variables p and sample size n are comparable in magnitude.

A.5.2 Selecting the best subset

We identify a handful of settings where our greedy (Algorithm 1) and/or swapping (Algorithm 2) algorithms find the best subset according to the CSS objective (2.1). In the below, we consider performing CSS on a covariance matrix $\mathbf{\Sigma}$ and suppose we are aiming to select a size-k subset. Proofs of the claims can be found in Appendix A.7.19.

1. **Independence**: Supposing that $\mathbf{\Sigma}$ is a diagonal matrix, both our greedy and swapping algorithms find an optimal solution.

2. **Perfect reconstruction**: Suppose $\mathbf{\Sigma}$ is such that there exists a set of k variables that perfectly linearly reconstruct the remaining. Then both our greedy and swapping algorithms will find such a set of variables, and therefore find an optimal solution.

3. **Equicorrelation**: If $\mathbf{\Sigma}$ is a equicorrelation matrix, then all subsets have the same objective value by symmetry, and thus our algorithms trivially find an optimal solution. Note that our covariance view from Section 2.2 greatly simplifies our analysis of this setting.

4. **Block diagonal**: Suppose $\mathbf{\Sigma}$ is a block diagonal correlation with k blocks, i.e., there are k groups of variables. Let $M_1 \in \mathbb{R}^{q\times q}, \ldots, M_k \in \mathbb{R}^{q\times q}$ be the different blocks. Supposing that each group has a strong enough representative, i.e., there exists a row/column i_j in each of the $j = 1, \ldots, k$ blocks such that $\|M_{\bullet i_j}\|_2^2 > q/2$, then our greedy algorithm will find the optimal solution.

Empirically we verified that when $\mathbf{\Sigma}_{ij} = \rho^{|i-j|}$ (covariance of an AR(1) process) our greedy and swapping algorithms are not guaranteed to find the optimal solution.

A.6 L^2 Random variables as a Hilbert space

The results of this section are generally known but hard to find covered in full generality. We give detailed account of them as they are particularly useful for proving the results in this article. Consider a probability space $(\Omega, \mathcal{F}, P)$. Upon identifying random variables which are equal almost surely in the same equivalence class, the random variables in $L^2(\Omega, \mathcal{F}, P)$ form a complete Hilbert space. Equality between random variables denotes membership to the same equivalence class, i.e., almost sure equality. For two real-valued random variables $X, Y \in L^2$, the inner product $\langle X, Y\rangle_{\mathcal{H}}$, given by $E[XY]$, induces a norm $\|X\|_{\mathcal{H}}^2 = E[X^2]$. As abuse of notation, when dealing with random vectors $Z \in \mathbb{R}^k$ with entries Z_i in L^2, we let $\|Z\|_{\mathcal{H}}^2 = \sum_{i=1}^k \|Z_i\|_{\mathcal{H}}^2$, so $\|Z\|_{\mathcal{H}}^2 = \mathrm{Tr}(E[ZZ^\top])$.

Since we're working in a Hilbert space, we can project random variables onto linear subspaces of our Hilbert space. We only project onto finite dimensional subspaces, so we don't concern ourselves with the subtleties of infinite dimensional subspaces. If $\mathcal{G}$ is a subspace of $\mathcal{H}$, we denote the projection of X onto $\mathcal{G}$ as $\text{Proj}_{\mathcal{G}}(X)$. For random vectors, $\text{Proj}_{\mathcal{G}}(Z)$ is the random vector with ith entry $\text{Proj}_{\mathcal{G}}(Z_i)$. We denote the subspace spanned by the entries of Z as $\text{Span}(Z)$, but as an abuse of notation we write $\text{Proj}_Z(\cdot)$ instead of $\text{Proj}_{\text{Span}(Z)}(\cdot)$. The following Lemma helps us compute projections:

Lemma A.6.1. *For random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$,*

$$\{E[YX^\top]E[XX^\top]^+ + \boldsymbol{E} : \boldsymbol{E} \in \mathbb{R}^{q\times p} \text{ such that } \boldsymbol{E}X = 0\} = \underset{\boldsymbol{B}\in\mathbb{R}^{q\times p}}{\text{argmin}}\, E[\|Y - \boldsymbol{B}X\|_2^2].$$

Theorem 12 provides some useful facts about these projections.

Theorem 12. *Consider random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$ and define $P = E[YX^\top]E[XX^\top]^+X$ and $R = Y - P$. Then,*

(i) $P = \text{Proj}_X(Y)$,

(ii) $E[PR^\top] = \mathbf{0}$ *and* $E[XR^\top] = \mathbf{0}$, *i.e., each entry of X and P is orthogonal to each entry of R,*

(iii) $\|Y\|_{\mathcal{H}}^2 = \|P\|_{\mathcal{H}}^2 + \|R\|_{\mathcal{H}}^2$,

(iv) $E[PP^\top] = E[YX^\top]E[XX^\top]^+E[XY^\top]$ *and* $E[RR^\top] = E[YY^\top] - E[YX^\top]E[XX^\top]^+E[XY^\top]$.

The following is a useful corollary of Theorem 12.

Corollary 9. *Considering random vectors $X \in \mathbb{R}^p$, $Y \in \mathbb{R}^q$, $Y = \boldsymbol{W}X + \epsilon$ with $E[X\epsilon^\top] = 0$ if and only if $\boldsymbol{W} = E[YX^\top]E[XX^\top]^+ + \boldsymbol{E}$ where $\boldsymbol{E} \in \mathbb{R}^{q\times p}$ such that $\boldsymbol{E}X = 0$.*

A.7 Proofs and derivations

A.7.1 Proof of Lemma A.6.1

Minimizing $E[\|Y-\boldsymbol{B}X\|_2^2] = E[\text{Tr}((Y-\boldsymbol{B}X)(Y-\boldsymbol{B}X)^\top] = \text{Tr}(E[YY^\top]-2\boldsymbol{B}E[XY^\top]+\boldsymbol{B}E[XX^\top]\boldsymbol{B}^\top)$ with respect to $\boldsymbol{B}$ is an unconstrained convex minimization problem. Thus $\boldsymbol{B}$ is a minimizer if and only if it sets the first derivative to 0. Using [Petersen and Pedersen, 2008, 2.5.2] to compute the first derivative, we get $\boldsymbol{B}$ is a solution if and only if $\boldsymbol{B}E[XX^\top] - E[YX^\top] = 0$.

We check that $\boldsymbol{B} = E[YX^\top]E[XX^\top]^+$ is a solution. If $v \in \mathbb{R}^p$ is in the null space of $E[XX^\top]$, then $0 = v^\top E[XX^\top]v = E[(X^\top v)^2]$, so $X^\top v = 0$. Then also $E[YX^\top]v = E[Y(X^\top v)] = 0$. Then, for our choice of v, $(E[YX^\top]E[XX^\top]^+E[XX^\top] - E[YX^\top])v = 0$. If $v \in \mathbb{R}^p$ is not in the null space of $E[XX^\top]$, then $E[XX^\top]^+E[XX^\top]v = v$, so $(E[YX^\top]E[XX^\top]^+E[XX^\top] - E[YX^\top])v = 0$. Thus $(E[YX^\top]E[XX^\top]^+E[XX^\top] - E[YX^\top])v = 0$ for all $v \in \mathbb{R}^p$, so $E[YX^\top]E[XX^\top]^+E[XX^\top] - E[YX^\top] = 0$.

For any $\boldsymbol{E} \in \mathbb{R}^{q\times p}$ such that $\boldsymbol{E}X = 0$ it is obvious that $E[YX^\top]E[XX^\top]^+ + \boldsymbol{E}$ is a solution. If $\tilde{\boldsymbol{B}}$ is also a solution, then $\tilde{\boldsymbol{B}}E[XX^\top] - E[YX^\top] = 0$ and thus $(E[YX^\top]E[XX^\top]^+ - \tilde{\boldsymbol{B}})E[XX^\top] = 0$. Thus $\tilde{\boldsymbol{B}} = E[YX^\top]E[XX^\top]^+ + \boldsymbol{E}$ for some $\boldsymbol{E} \in \mathbb{R}^{q\times p}$ such that $\boldsymbol{E}E[XX^\top] = \mathbf{0}$. Since also $\boldsymbol{E}E[XX^\top]\boldsymbol{E}^\top = \mathbf{0}$, we have that $\text{Tr}(E[\boldsymbol{E}X(\boldsymbol{E}X)^\top])) = E[\|\boldsymbol{E}X\|_2^2] = 0$ so $\boldsymbol{E}X = 0$.

A.7.2 Proof of Theorem 12

We prove each result in the order the are presented.

(i) The ith entry of $\text{Proj}_X(Y)$ is given by $Y_i - (b^*)^\top X$, where b^* minimizes $\|Y_i - b^\top X\|_H^2 = E[(Y_i - b^\top X)^2]$ over b. By Lemma A.6.1, b^* must be given by the ith row of $E[YX^\top]E[XX^\top]^+ + \boldsymbol{E}$ for some $\boldsymbol{E} \in \mathbb{R}^{q\times p}$ such that $\boldsymbol{E}X = 0$. Thus, up to almost sure equality, that the ith entry of $\text{Proj}_X(Y)$ is given by $Y_i - (E[YX^\top]E[XX^\top]^+)_{i\bullet}X$.

(ii) From (i) we know that P_j is $\text{Proj}_X(Y_j)$ and $R_i = Y_i - \text{Proj}_X(Y_i)$. From standard properties of projections in Hilbert spaces, we know that $P_j \in \text{Span}(X)$ and R_i is orthogonal to anything in $\text{Span}(X)$. thus P_j and R_i are orthogonal, meaning exactly that $E[PR^\top] = \mathbf{0}$. Same argument applies for X.

(iii) From (ii) we know that $E[P_iR_i] = 0$. Thus, since $Y_i = P_i + R_i$ we have $\|Y_i\|_{\mathcal{H}}^2 = E[Y_i^2] = E[R_i^2] + E[P_i^2] = \|P_i\|_{\mathcal{H}}^2 + \|R_i\|_{\mathcal{H}}^2$.

(iv) We can compute $E[PP^\top] = E[YX^\top]E[XX^\top]^+E[XX^\top]E[XX^\top]^+E[XY^\top] = E[YX^\top]E[XX^\top]^+E[XY^\top]$. Since $Y = P+R$ and we know $E[PR^\top] = \mathbf{0}$ from (ii), we immediately get $E[YY^\top] = E[PP^\top]+E[RR^\top]$. So, $E[RR^\top] = E[YY^\top] - E[PP^\top] = E[YY^\top] - E[YX^\top]E[XX^\top]^+E[XY^\top]$.

A.7.3 Proof of Corollary 9

The reverse direction follows as if $\boldsymbol{W} = E[YX^\top]E[XX^\top]^+ + \boldsymbol{E}$, then $\boldsymbol{W}X = P$ and $\epsilon = R$ as in the statement of Theorem 12. Thus by Theorem 12(ii), $E[X\epsilon^\top] = 0$. For the forward direction, by the properties of projections in Hilbert spaces, it is only possible for ϵ_i to be uncorrelated with X if $\boldsymbol{W}_{i\bullet}X$ equals $\text{Proj}_X(Y_i) = P_i$. Therefore we must have $\boldsymbol{W} = E[YX^\top]E[XX^\top]^+ + \boldsymbol{E}$ where $\boldsymbol{E} \in \mathbb{R}^{q\times p}$ such that $\boldsymbol{E}X = 0$ by Theorem 12(i).

A.7.4 Proof of Proposition 1 and Proposition 2

By Lemma A.6.1 and Theorem 12(iv),

$$\min_{\boldsymbol{B}\in\mathbb{R}^{p\times k}} E[\|X - \boldsymbol{B}X_S\|_2^2] = \text{Tr}(E[XX^\top] - E[XX^\top]_{\bullet S}E[XX^\top]_S^+E[XX^\top]_{S\bullet}).$$

Taking X to be mean-zero, this immediately implies Proposition 2. To prove Proposition 1, let X be uniformly random over the rows of the matrix $\boldsymbol{X}$ and evaluate the expectations on both sides.

A.7.5 Proof of Theorem 1

We provide a proof of Theorem 1. We then provide the analogous result when we allow the covariance of $X_{-S}|X_S$ to be diagonal (see (A.1)). Lastly we provide the analogous results for both these cases when we assume that the principal variable distribution F is Gaussian.

PCSS model

Suppose we restrict F to have a probability density bounded by $K < \infty$. Our model is parameterized by $\theta = (S, f, \mu_{-S}, \boldsymbol{W}, \sigma^2)$ where f is the density of F. Let $\hat{\mu}$ and $\hat{\boldsymbol{\Sigma}}$ be the sample mean and sample covariance

of our data. Fix a size-k subset S. The negative log-likelihood, scaled by $1/n$, under our model is given by

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\ell(x^{(i)};\theta)\\
&=-\frac{1}{n}\sum_{i=1}^{n}\Big(\ell(x_S^{(i)};\theta)+\ell(x_{-S}^{(i)}|x_S^{(i)};\theta)\Big)\\
&=\frac{1}{n}\sum_{i=1}^{n}\log(f(x_S^{(i)}))+\frac{1}{2}\cdot\frac{1}{n}\sum_{i=1}^{n}\Big((p-k)\log(2\pi)+(p-k)\log\sigma^2+\sigma^{-2}\|x_{-S}^{(i)}-\mu_S-\boldsymbol{W}(x_S^{(i)}-\mu_{-S})\|_2^2\Big)\\
&=-\frac{1}{n}\sum_{i=1}^{n}f(x_S^{(i)})+\frac{p-k}{2}\log\sigma^2+\frac{1}{2\sigma^2}\cdot\sum_{j=1}^{p-k}\left(\frac{1}{n}\sum_{i=1}^{n}((x_{-S}^{(i)})_j-(\mu_S)_j-\boldsymbol{W}_{j\bullet}(x_S^{(i)}-\mu_{-S}))^2\right)+\frac{(p-k)\log(2\pi)}{2}.
\end{aligned}$$

We know that $-\frac{1}{n}\sum_{i=1}^{n}\log(f(x_S^{(i)}))$ is bounded below by $-\log(K)$. By adding together n bump functions are narrow enough and attain their maximum height of K at each $x^{(i)}$ and placing any remaining mass in the tails of f, we can find an f that attains this bound. Although $\mu_S = E_F[X_S]$ is not explicitly known, we can absorb $(\mu_S)_j$ into $\boldsymbol{W}_{j\bullet}\mu_{-S}$ for each j, so it is an inconsequential parameter. So, without loss of generality we set $\mu_S = \hat{\mu}_S$ in the above. For $c > 0$, the function $\log x + \frac{c}{x}$ is minimized at $x = c$ and has minimized value $\log(c)+1$. Thus we should pick $(\mu_{-S})_j$ and $\boldsymbol{W}_{j\bullet}$ to minimize $\frac{1}{n}\sum_{i=1}^{n}((x_{-S}^{(i)})_j-(\mu_{-S})_j-\boldsymbol{W}_{j\bullet}(x_S^{(i)}-\hat{\mu}_S))^2$. This is a convex problem. From the first order condition one can compute the minimizing value for $(\mu_{-S})_j$ is $(\hat{\mu}_{-S})_j$, and then apply Lemma A.6.1 to compute that the minimizing value of $\boldsymbol{W}_{j\bullet}$ is given by $(\hat{\boldsymbol{\Sigma}}_{-S,S})_{j\bullet}\hat{\boldsymbol{\Sigma}}_S^{+}$. Plugging in these minimizing values gives

$$\frac{1}{n}\sum_{i=1}^{n}((x_{-S}^{(i)})_j-(\mu_{-S})_j-\boldsymbol{W}_{j\bullet}(x_S^{(i)}-\hat{\mu}_S))^2=(\hat{\boldsymbol{\Sigma}}_{-S})_{jj}-(\hat{\boldsymbol{\Sigma}}_{-S,S})_{j\bullet}\hat{\boldsymbol{\Sigma}}_S^{+}(\hat{\boldsymbol{\Sigma}}_{S,-S})_{\bullet j}.$$

Since no k variables perfectly linearly reconstruct the remaining, we know at least one of $(\hat{\boldsymbol{\Sigma}}_{-S})_{jj}-(\hat{\boldsymbol{\Sigma}}_{-S,S})_{j\bullet}\hat{\boldsymbol{\Sigma}}_S^{+}(\hat{\boldsymbol{\Sigma}}_{S,-S})_{\bullet j}$ is positive. The minimizing value of σ^2 is thus attained at

$$\sigma^2=\frac{1}{p-k}\sum_{j=1}^{p-k}(\hat{\boldsymbol{\Sigma}}_{-S})_{jj}-(\hat{\boldsymbol{\Sigma}}_{-S,S})_{j\bullet}\hat{\boldsymbol{\Sigma}}_S^{-1}(\hat{\boldsymbol{\Sigma}}_{S,-S})_{\bullet j}=\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_{-S}-\hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{+}\hat{\boldsymbol{\Sigma}}_{S,-S})/(p-k).$$

Plugging in all our minimizing values, the negative log likelihood is

$$-\log(K)+\frac{p-k}{2}\log\Big(\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_{-S}-\hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{+}\hat{\boldsymbol{\Sigma}}_{S,-S})/(p-k)\Big)+\frac{p-k}{2}(1+\log(2\pi)).$$

Thus the S which minimizes the negative log-likelihood is the one which minimizes $\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_{-S}-\hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{+}\hat{\boldsymbol{\Sigma}}_{S,-S}) = \mathrm{Tr}(\hat{\boldsymbol{\Sigma}}-\hat{\boldsymbol{\Sigma}}_{\bullet S}\hat{\boldsymbol{\Sigma}}_S^{+}\hat{\boldsymbol{\Sigma}}_{S\bullet})$, completing the claim.

In the case that F is completely unrestricted, our model is parameterized by $\theta = (S, F, \mu_{-S}, \boldsymbol{W}, \sigma^2)$. Fix

a size-k subset S. The negative log hybrid likelihood is given by

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\ell(x^{(i)};\theta)\\
&=-\frac{1}{n}\left(\sum_{i=1}^{n}\ell(x_S^{(i)};\theta)-\ell(x_{-S}^{(i)}|x_S^{(i)};\theta)\right)\\
&=-\frac{1}{n}\sum_{i=1}^{n}\log(F(x_S^{(i)}))-\frac{1}{n}\sum_{i=1}^{n}\ell(x_{-S}^{(i)}|x_S^{(i)};\theta).
\end{aligned}$$

By [Owen, 2001, Theorem 2.1], the optimal choice of F is that which puts weight $1/n$ on each $x^{(i)}$. For this minimizing F, the negative log hybrid likelihood is $\log(n)-\frac{1}{n}\sum_{i=1}^{n}\ell(x_{-S}^{(i)}|x_S^{(i)};\theta)$. The argument is then identical to the one above.

Diagonal conditional covariance

We show the analogous result for the when we allow the covariance of $X_{-S}|X_S$ to be diagonal (see (A.1)). We will assume that every $(k+1)$ by $(k+1)$ principal sub-matrix of $\hat{\boldsymbol{\Sigma}}$ is full rank.

In place of $\sigma^2>0$ we have the diagonal matrix $\boldsymbol{D}\succ\mathbf{0}$ as a parameter. Fix a subset S and again decompose the likelihood:

$$-\frac{1}{n}\sum_{i=1}^{n}\ell(x^{(i)};\theta)=-\frac{1}{n}\sum_{i=1}^{n}\left(\ell(x_S^{(i)};\theta)+\ell(x_{-S}^{(i)}|x_S^{(i)};\theta)\right).$$

The maximum likelihood estimate of the density f is the same as before, and finding the remaining maximizers is similar to before:

$$\begin{aligned}
&-\frac{1}{n}\sum_{i=1}^{n}\ell(x_{-S}^{(i)}|x_S^{(i)};\theta)\\
&=\frac{1}{2}\sum_{j=1}^{p-k}\left(\log\boldsymbol{D}_{jj}+\frac{1}{\boldsymbol{D}_{jj}}\cdot\left(\frac{1}{n}\sum_{i=1}^{n}((x_{-S}^{(i)})_j-(\mu_S)_j-\boldsymbol{W}_{j\bullet}(x_S^{(i)}-\mu_{-S}))^2\right)\right)+\frac{(p-k)\log(2\pi)}{2}.
\end{aligned}$$

Following the reasoning above, we set $\boldsymbol{W}_{j\bullet}$ to minimize $\frac{1}{n}\sum_{i=1}^{n}((x_{-S}^{(i)})_j-(\mu_{-S})_j-\boldsymbol{W}_{j\bullet}(x_S^{(i)}-\hat{\mu}_S))^2$ resulting in a minimized value of $(\hat{\boldsymbol{\Sigma}}_{-S})_{jj}-(\hat{\boldsymbol{\Sigma}}_{-S,S})_{j\bullet}\hat{\boldsymbol{\Sigma}}_S^{-1}(\hat{\boldsymbol{\Sigma}}_{S,-S})_{\bullet j}>0$ where the inequality and invertability of $\hat{\boldsymbol{\Sigma}}_S^{-1}$ come from our new assumption. Like before, this minimized value is the minimizing choice for $\boldsymbol{D}_{jj}$, implying that the minimizing $\boldsymbol{D}$ is given by $\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-S}-\hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{-1}\hat{\boldsymbol{\Sigma}}_{S,-S})$. The minimized negative log-likelihood is then given by

$$-\log(K)+\frac{1}{2}\log(|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-S}-\hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{-1}\hat{\boldsymbol{\Sigma}}_{S,-S})|)+\frac{p-k}{2}(1+\log(2\pi)).$$

So a subset minimizes the negative log-likelihood if and only if it minimizes $\log(|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-S}-\hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{-1}\hat{\boldsymbol{\Sigma}}_{S,-S})|)$.

The argument for minimizing the negative hybrid log-likelihood is then identical to the one above.

Gaussian principal variable distribution F

Now we derive the analogous results when we require that the principal variable distribution F is a Gaussian, so $X_S \sim N(\mu_S, \boldsymbol{C})$ where $\boldsymbol{C} \succ \mathbf{0}$. First consider the PCSS model with this restriction. Along with the assumption that no k variables perfectly reconstruct the rest, we require that every k by k principal sub-matrix of $\boldsymbol{\Sigma}$ is full rank. The model is parameterized by $(S, \mu, \boldsymbol{C}, \boldsymbol{W}, \sigma^2)$. Fix a subset S. The negative log-likelihood, scaled by $\frac{1}{n}$, is given by

$$-\frac{1}{n}\sum_{i=1}^{n} \ell(x^{(i)}; \theta) = -\frac{1}{n}\sum_{i=1}^{n} \left(\ell(x_S^{(i)}; \theta) + \ell(x_{-S}^{(i)} | x_S^{(i)}; \theta)\right).$$

Fix a size-k subset S. We can ignore the dependence of $-\frac{1}{n}\sum_{i=1}^{n} \ell(x_{-S}^{(i)} | x_S^{(i)}; \theta)$ on μ_S for the same reason in the proof of Theorem 1. Minimizing $-\frac{1}{n}\sum_{i=1}^{n}(\ell(x_S^{(i)}; \theta)$ with respect to $\boldsymbol{C}$ and μ_S is equivalent finding the unrestricted MLE for a multivariate Gaussian. Since $\hat{\boldsymbol{\Sigma}}_S$ is invertible by our new assumption, the minimizing values of C and μ_S are $\hat{\boldsymbol{\Sigma}}_S$ and $\hat{\mu}_S$. The minimizing values of W, σ^2 are then identical to those in the proof of Theorem 1 by the same computations, and the minimized value of the negative log-likelihood is

$$\frac{1}{2}\log|\hat{\boldsymbol{\Sigma}}_S| + \frac{p-k}{2}\log\left(\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_{-S} - \hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{-1}\hat{\boldsymbol{\Sigma}}_{S,-S})/(p-k)\right) + \frac{p}{2}(1 + \log(2\pi)).$$

So the subset which minimizes the negative log-likelihood is one which minimizes $\log|\hat{\boldsymbol{\Sigma}}_S| + (p-k)\log(\mathrm{Tr}(\hat{\boldsymbol{\Sigma}}_{-S} - \hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{-1}\hat{\boldsymbol{\Sigma}}_{S,-S})/(p-k)$. This is sufficient to prove the claim.

In the case that the principal variable distribution F is a Gaussian and the covariance of $X_{-S} \mid X_S$ is allowed to be diagonal, we again assume that every $(k+1)$ by $(k+1)$ principal sub-matrix of $\hat{\boldsymbol{\Sigma}}$ is full rank. Then proof is identical to above. The subset which minimizes the negative log-likelihood is one which minimizes $\log|\boldsymbol{\Sigma}_S| + \log(|\mathrm{Diag}(\hat{\boldsymbol{\Sigma}}_{-S} - \hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^{-1}\hat{\boldsymbol{\Sigma}}_{S,-S})|)$.

A.7.6 Proof of Theorem 2

Throughout this proof, we denote the sorted eigenvalues of a positive semi-definite matrix $\boldsymbol{A} \in \mathbb{S}_+^{q\times q}$ as $\lambda_1(\boldsymbol{A}) \geq \cdots \geq \lambda_q(\boldsymbol{A}) \geq 0$. For a matrix $\boldsymbol{M} \in \mathbb{R}^{q\times k}$ we denote its sorted top-k singular values as $s_1(\boldsymbol{M}) \geq \cdots \geq s_k(M) \geq 0$. We also let $\|\cdot\|_2$ denote the spectral norm of a matrix.

In Appendix A.4, we identified a high-dimensional setting where the in-sample CSS solution (which is exactly the MLE $\hat{S}$) is consistent for the population CSS solution. This is the subject of Theorem 11, whose proof is the bulk of Appendix A.4. If we can show that (1) the assumptions of Theorem 11 are satisfied and (2) that the population subset S is also the population CSS solution once n is large enough, then Theorem 2 is immediately implied by Theorem 11. We recall the assumptions of Theorem 11 below.

Assumption 1 (Light-tailed data)**.** *The sub-Gaussian norm (see [Vershynin, 2018, Definition 2.5.6]) of each variable is bounded above by a constant independent of p and n.*

Assumption 2 (Low-dimensional invertibility)**.** *The minimum of the smallest eigenvalues of all the $k \times k$ principal sub-matrices of $\boldsymbol{\Sigma}$ is bounded below by a positive constant independent of p and n.*

Assumption 3 (Separation)**.** *For some $\delta > 0$, the achieved CSS objective for the population best subset*

$$S^* = \underset{U \subseteq [p], |U|=k}{\operatorname{argmin}} \operatorname{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet U} \boldsymbol{\Sigma}_U^+ \boldsymbol{\Sigma}_{\bullet U})$$

is well separated from the objective achieved by any other subset

$$\min_{U \neq S^*} \operatorname{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet U} \boldsymbol{\Sigma}_U^+ \boldsymbol{\Sigma}_{\bullet U}) - \operatorname{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet S^*} \boldsymbol{\Sigma}_{S^*}^+ \boldsymbol{\Sigma}_{\bullet S^*}) = \Omega(p^{1/2+\delta})$$

To show that the conditions of Theorem 2 imply these assumptions and also imply that the population subset S is eventually the same as the population CSS minimizer S^*, we will need two linear algebra lemmas whose results are well-known [joriki, Pavel].

Lemma A.7.1. *For any positive semi-definite matrix $\boldsymbol{A}$ written in block matrix notation*

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}, \end{bmatrix}$$

we have $\lambda_1(\boldsymbol{A}) \leq \lambda_1(\boldsymbol{A}_{11}) + \lambda_1(\boldsymbol{A}_{22})$.

Proof. We recall the proof from joriki. Let v be a unit vector. We can write $v = \begin{bmatrix} v_1 & v_2 \end{bmatrix}^T$ as $v = \|v_1\| \begin{bmatrix} v_1/\|v_1\| & 0 \end{bmatrix}^T + \|v_2\|_2 \begin{bmatrix} 0 & v_2/\|v_2\|_2 \end{bmatrix}^T$. Thus we have written v as $c_1 u_2 + c_2 u_2$ where u_1, u_2, and $c = \begin{bmatrix} c_1, c_2 \end{bmatrix}^T \in \mathbb{R}^2$ are all unit vectors. Defining the matrix

$$\boldsymbol{B} = \begin{bmatrix} u_1^T \boldsymbol{A}_{11} u_1 & u_1^T \boldsymbol{A}_{12} u_2 \\ u_2^T \boldsymbol{A}_{11} u_1 & u_2^T \boldsymbol{A}_{22} u_2. \end{bmatrix}$$

it is the case that $v^T \boldsymbol{A} v = c^T \boldsymbol{B} c$ by construction. Therefore,

$$\begin{aligned} v^T \boldsymbol{A} v &= c^T \boldsymbol{B} c \\ &\leq \lambda_1(\boldsymbol{B}) \\ &\leq \operatorname{Tr}(\boldsymbol{B}) \\ &= u_1^T \boldsymbol{A}_{11} u_1 + u_2^T \boldsymbol{A}_{22} u_2 \\ &\leq \lambda_1(\boldsymbol{A}_{11}) + \lambda_1(\boldsymbol{A}_{22}) \end{aligned}$$

Because the inequality holds for all unit vectors v, it must be the case that $\lambda_1(\boldsymbol{A}) \leq \lambda_1(\boldsymbol{A}_1 1) + \lambda_1(\boldsymbol{A}_2 2)$ as desired. □

Lemma A.7.2. *For any positive definite matrix $\boldsymbol{A} \in \mathbb{S}_{++}^{k \times k}$ written in block matrix notation*

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22}, \end{bmatrix}$$

the minimum eigenvalue of the Schur compelment $\boldsymbol{A}_{22} - \boldsymbol{A}_{21} \boldsymbol{A}_{11}^{-1} \boldsymbol{A}_{12}$ is at least as large as the minimum eigenvalue of $\boldsymbol{A}$.

Proof. We recall the proof from Pavel.

$$\begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12} \\ \boldsymbol{A}_{21} & \boldsymbol{A}_{22} \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_{k_1} & 0 \\ \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1} & \boldsymbol{I}_{k_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{A}_{11} & 0 \\ 0 & \boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_{k_1} & \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \\ 0 & \boldsymbol{I}_{k_2} \end{bmatrix},$$

Let

$$\mathcal{V} := \left\{ \begin{bmatrix} x \\ y \end{bmatrix} \in \mathbb{R}^k : x + \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12}y = 0 \right\}.$$

Note that if $z = \begin{bmatrix} x^T & y^T \end{bmatrix}^T \in \mathcal{V}$, then

$$\begin{bmatrix} \boldsymbol{I}_{k_1} & \boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12} \\ 0 & \boldsymbol{I}_{k_2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}.$$

so

$$z^T\boldsymbol{A}z = y^T(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})y \tag{A.12}$$

Also $\|z\|_2^2 \geq \|y\|_2^2$. This is sufficient to imply that

$$\min_{z\in\mathbb{R}^k, z\neq 0} \frac{z^T\boldsymbol{A}z}{z^Tz} \leq \min_{z\in\mathcal{V}, z\neq 0} \frac{z^T\boldsymbol{A}z}{z^Tz} \leq \min_{y\in\mathbb{R}^{k_2}\neq 0} \frac{y^T(\boldsymbol{A}_{22} - \boldsymbol{A}_{21}\boldsymbol{A}_{11}^{-1}\boldsymbol{A}_{12})y}{y^Ty}$$

which establishes the claim □

Now we catalogue what the conditions of Theorem 2 tell us:

1. Because all the sub-gaussian norm of each variable is bounded above by a constant, there must be some constant $C > 0$ such that all the entries of $\boldsymbol{\Sigma}$ have absolute value bounded above by C.
2. There are constants $m > 0$ and $v > 0$ such that $\lambda_k(\boldsymbol{\Sigma}_S) > m$ and $\sigma^2 \geq v$.
3. There are constants $\eta > 0$ and $c_1, \ldots, c_k > 0$ such that, if N_i is the number of entries of $\boldsymbol{W}_{\bullet i}$ whose absolute value is strictly less than η, then eventually $N_i \geq c_i p^{1/2+\delta}$. As a consequence, there is some constant $c > 0$ such that eventually $\min_{i\in[k]} N_i \geq cp^{1/2+\delta}$.

Of course, all these constants are independent of n and p. With this information in hand, we can show both that the assumptions hold and that S will eventually be the same as S^*.

Sub-Gaussianity: Assumption 1 is itself assumed in the satement of Theorem 2.

Low-dimensional invertibility: Consider any size-k subset U that is entirely disjoint from S. We can write the joint covariance of all the variables in U and S in block matrix form as

$$\boldsymbol{\Sigma}_{S\cup U} = \begin{bmatrix} \boldsymbol{\Sigma}_S & \boldsymbol{\Sigma}_{SU} \\ \boldsymbol{\Sigma}_{US} & \boldsymbol{\Sigma}_{US}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{SU} + \sigma^2\boldsymbol{I}_k \end{bmatrix}$$

One can confirm by direct computation that

$$\boldsymbol{\Sigma}_{S\cup U}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_S^{-1} + \sigma^{-2}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{SU}\boldsymbol{\Sigma}_{US}\boldsymbol{\Sigma}_S^{-1} & -\sigma^{-2}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{SU} \\ -\sigma^{-2}\boldsymbol{\Sigma}_{US}\boldsymbol{\Sigma}_S^{-1} & \sigma^{-2}\boldsymbol{I}_k \end{bmatrix}.$$

We can bound $\|\boldsymbol{\Sigma}_{SU}\boldsymbol{\Sigma}_{US}\|_2 \leq \mathrm{Tr}(\boldsymbol{\Sigma}_{SU}\boldsymbol{\Sigma}_{US}) \leq k^2C^2$. Then, applying Lemma A.7.1, we have

$$\begin{aligned} \lambda_1(\boldsymbol{\Sigma}_{S\cup U}^{-1}) &\leq \lambda_1(\boldsymbol{\Sigma}_S^{-1} + \sigma^{-2}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{SU}\boldsymbol{\Sigma}_{US}\boldsymbol{\Sigma}_S^{-1}) + \lambda_1(\sigma^{-2}\boldsymbol{I}_k) \\ &= \|\boldsymbol{\Sigma}_S^{-1} + \sigma^{-2}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{SU}\boldsymbol{\Sigma}_{US}\boldsymbol{\Sigma}_S^{-1}\|_2 + \sigma^{-2} \\ &\leq \|\boldsymbol{\Sigma}_S^{-1}\|_2 + \sigma^{-2}\|\boldsymbol{\Sigma}_S^{-1}\|_2^2\|\boldsymbol{\Sigma}_{SU}\boldsymbol{\Sigma}_{US}\|_2 + \sigma^{-2} \\ &\leq m^{-1} + v^{-1}m^{-2}k^2C^2 + v^{-1}. \end{aligned}$$

As a consequence, we have that

$$\lambda_k(\boldsymbol{\Sigma}_{S\cup U}) \geq (m^{-1} + v^{-1}m^{-2}k^2C^2 + v^{-1})^{-1} > 0.$$

Now if we consider any size-k subset $\widetilde{U}$, the matrix $\boldsymbol{\Sigma}_{\widetilde{U}}$ is a principal sub-matrix of $\boldsymbol{\Sigma}_{S\cup U}$ for some choice of size-k subset U. Therefore by Cauchy's eigenvalue interlacing theorem Bhatia [1997],

$$\lambda_k(\boldsymbol{\Sigma}_{\widetilde{U}}) \geq (m^{-1} + v^{-1}m^{-2}k^2C^2 + v^{-1})^{-1} > 0.$$

and we have established that Assumption 2 holds.

Separation: Now we establish both that Assumption 3 holds and that S^* and S are eventually the same. Recall that $X_\ell = \boldsymbol{W}_{\ell\bullet}X_S + \epsilon$ for $\ell \notin S$, where for ease of notation, we index the $p-k$ rows of $\boldsymbol{W}$ by elements of $-S$. Fix any size-k subset U that is not equal to S, so that there is some $i \in S$ that is not in U. To establish Assumption 3 and that S and S^* are eventually equal, it will suffice to argue the following

- If $\ell \notin S \cup U$ then
$$(\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell}) - (\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell}) \geq 0.$$
- If $\ell \notin S \cup U$ and $\boldsymbol{W}_{\ell i} \geq \eta$, then
$$(\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell}) - (\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell}) \geq \gamma$$
where $\gamma > 0$ is some constant.

To see why this is sufficient, note that

$$\begin{aligned}
&\big(\mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\bullet})\big) - \big(\mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\bullet})\big) \\
&\geq \sum_{\ell \notin S\cup U, \boldsymbol{W}_{\ell i} \geq w} \big[(\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell}) - (\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell})\big] \\
&\quad + \sum_{\ell \notin S\cup U, \boldsymbol{W}_{\ell i} < w} \big[(\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell}) - (\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell})\big] \\
&\quad - \left| \sum_{\ell \in S\cup U} (\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell}) - (\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell}) \right| \\
&\geq \gamma(N_i - k) - 4kC \\
&\geq \gamma \min_{j\in[k]} N_j - \gamma k - 4kC \\
&\geq \gamma c p^{\frac{1}{2}+\delta} - \gamma k - 4kC
\end{aligned}$$

where the last inequality holds for large enough n and we have also bounded

$$\left| \sum_{\ell \in S\cup U} \mathrm{Tr}(\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell}) - \mathrm{Tr}(\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell}) \right| \leq \sum_{\ell \in S\cup U} 2|\mathrm{Tr}(\boldsymbol{\Sigma}_{\ell\ell})| \leq 4kC.$$

This bound has no dependence on U, and therefore it holds that for large enough n that

$$\min_{U\neq S} \mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\bullet}) - \mathrm{Tr}(\boldsymbol{\Sigma} - \boldsymbol{\Sigma}_{\bullet S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\bullet}) \geq \gamma c p^{\frac{1}{2}+\delta} - \gamma k - 4kC$$

Since p is growing by assumption, this is sufficient to imply both Assumption 3 and that S is the eventually the same as the population CSS minimizer S^*.

We devote the rest of the proof to showing the claim. Let $\boldsymbol{L}$ be the Cholesky decomposition of $\boldsymbol{\Sigma}_S$, so $\boldsymbol{\Sigma} = \boldsymbol{L}^T\boldsymbol{L}$. If we write the population covariance in block matrix form with the subset S as the first block and the remaining variables as the second block, then it is straightforward to check it admits the following Cholesky decomposition

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{L}^T & \boldsymbol{L}^T\boldsymbol{W}^T \\ \boldsymbol{0} & \sigma\boldsymbol{I}_{p-k} \end{bmatrix}^T \begin{bmatrix} \boldsymbol{L}^T & \boldsymbol{L}^T\boldsymbol{W}^T \\ \boldsymbol{0} & \sigma\boldsymbol{I}_{p-k} \end{bmatrix}$$

Let $v_1, \ldots, v_p \in \mathbb{R}^p$ be the columns of this Cholesky decomposition. For a subset $T \subseteq [p]$, denote the span of $\{v_j : j \in T\}$ as $\mathrm{Span}(T)$, the projection of $x \in \mathbb{R}^p$ onto the span as $\mathrm{Proj}_T(x)$, and the component of x that is orthogonal to this subspace as $R_T(x) = x - \mathrm{Proj}_T(x)$. By construction, the columns satisfy $v_s^T v_t = \boldsymbol{\Sigma}_{st}$. Also they satisfy for $\ell \notin S$ that

$$v_\ell = n_\ell + \sum_{j\in S} \boldsymbol{W}_{\ell j} v_j = n_\ell + \tilde{v}_\ell \tag{A.13}$$

where

1. the n_ℓ are orthogonal to each other and the subspace $\mathrm{Span}(S)$,
2. the n_ℓ all have norm $\|n_\ell\|_2^2 = \sigma^2$,
3. the $\tilde{v}_\ell = \sum_{j\in S} \boldsymbol{W}_{\ell j} v_j$ are in $\mathrm{Span}(S)$.

It is also easy to confirm that, for a subset $T \subseteq [p]$, the norm $\|R_T(v_j)\|_2^2$ is exactly the Schur complement $\boldsymbol{\Sigma}_{jj} - \boldsymbol{\Sigma}_{jT}\boldsymbol{\Sigma}_T^{-1}\boldsymbol{\Sigma}_{Tj}$. As a consequence, recalling the universal lower bound $m > 0$ for $\lambda_k(\boldsymbol{\Sigma}_S)$, Lemma A.7.2 tells us that

$$\|R_{S-j}(v_j)\|_2^2 \geq \lambda_k(\boldsymbol{\Sigma}_S) \geq m,$$

where $S - j$ is the subset S with j removed.

Now fix a size-k subset U that is not equal to S. There must be at least some $i \in S$ that is not in U. Considering $\ell \notin S \cup U$, our goal is to lower bound the difference between

$$\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell} = \min_{\beta \in \mathbb{R}^k} \|v_\ell - \sum_{j \in U} \beta_j v_j\|_2^2,$$

where for notational simplicity we index the entries of β by the $j \in U$, and $\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell} = \sigma^2$. We can rewrite the objective

$$\begin{aligned}
&\|v_\ell - \sum_{j \in U} \beta_j v_j\|_2^2 \\
&= \|\tilde{v}_\ell - \sum_{j \in U \cap S} \beta_j v_j - \sum_{j \in U \setminus S} \beta_j \tilde{v}_j\|_2^2 + \sigma^2 + \sigma^2 \sum_{j \in U \setminus S} \beta_j^2 \\
&= \|\mathrm{Proj}_{U \cap S}(\tilde{v}_\ell) - \sum_{j \in U \setminus S} \beta_j \mathrm{Proj}_{U \cap S}(\tilde{v}_j) - \sum_{j \in U \cap S} \beta_j v_j\|_2^2 + \|R_{U \cap S}(\tilde{v}_\ell) - \sum_{j \in U \setminus S} \beta_j R_{U \cap S}(\tilde{v}_j)\|_2^2 + \sigma^2 + \sigma^2 \sum_{j \in U \setminus S} \beta_j^2
\end{aligned}$$

We see immediately that $\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell} \geq \sigma^2$. No matter what values of β_j we pick for $j \in U \setminus S$, the quantity

$$\mathrm{Proj}_{U \cap S}(\tilde{v}_\ell) - \sum_{j \in U \setminus S} \beta_j \mathrm{Proj}_{U \cap S}(\tilde{v}_j)$$

will always been in $\mathrm{Span}(S)$, so we will always set β_j for $j \in U \setminus S$ to make the left-most term equal to zero. Thus, letting $q = |U \cap S|$ we are left with

$$\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell} = \operatorname*{argmin}_{\beta \in \mathbb{R}^q} \|R_{U \cap S}(\tilde{v}_\ell) - \sum_{j \in U \setminus S} \beta_j R_{U \cap S}(\tilde{v}_j)\|_2^2 + \sigma^2 \sum_{j \in U \setminus S} \beta_j^2 + \sigma^2$$

The right hand side is a ridge regression problem with a unique solution β^* that has a closed form. Let we let $\boldsymbol{A} \in \mathbb{R}^{p \times q}$ be the matrix whose columns are given by the vectors $R_{U \cap S}(\tilde{v}_i)$. Because

$$\|R_{U \cap S}(\tilde{v}_i)\|_2^2 \leq \|\tilde{v}_i\|_2^2 \leq \|v_i\|_2^2 \leq C \tag{A.14}$$

We know that the trace and therefore largest eigenvalue of $\boldsymbol{A}^T\boldsymbol{A}$ is at most kC. This implies the largest singular value of $\boldsymbol{A}$ is at most $k^{1/2}C^{1/2}$. Let $\boldsymbol{U} \in \mathbb{R}^{p \times p}$ be the orthogonal matrix with columns u_i, whose first q columns are the top q left singular vectors of $\boldsymbol{A}$, [Hastie et al., 2009, Equation 3.47] implies for the

optimal β^* that

$$\begin{aligned}\|R_{U\cap S}(\tilde{v}_\ell) - \sum_{j\in U\setminus S}\beta_j^* R_{U\cap S}(\tilde{v}_i)\|_2^2 &= \sum_{i=1}^{q}\left(1-\frac{s_i(\boldsymbol{A})^2}{s_i(\boldsymbol{A})^2+\sigma^2}\right)^2 (u_i^T R_{U\cap S}(\tilde{v}_\ell))^2 + \sum_{i=q+1}^{p}(u_i^T R_{U\cap S}(\tilde{v}_\ell))^2 \\ &\geq \left(1-\frac{kC}{kC+\sigma^2}\right)^2 \sum_{i=1}^{p}(u_i^T R_{U\cap S}(\tilde{v}_\ell))^2 \\ &= \left(1-\frac{kC}{kC+\sigma^2}\right)^2 \|R_{U\cap S}(\tilde{v}_\ell)\|_2^2\end{aligned}$$

Now we just need to lower bound $\|R_{U\cap S}(\tilde{v}_\ell)\|_2^2$. Recall that $i \in S$ but $i \notin U$. Then we can decompose

$$\begin{aligned}R_{U\cap S}(\tilde{v}_\ell) &= R_{U\cap S}(\sum_{j\in S}\boldsymbol{W}_{\ell j}v_j) \\ &= \sum_{j\in S\setminus U}\boldsymbol{W}_{\ell j}R_{U\cap S}(v_j) \\ &= \boldsymbol{W}_{\ell i}R_{S-i}(R_{U\cap S}(v_i)) + \boldsymbol{W}_{\ell i}\mathrm{Proj}_{S-i}(R_{U\cap S}(v_i)) + \sum_{j\in S\setminus U, j\neq i}\boldsymbol{W}_{\ell j}R_{U\cap S}(v_j)\end{aligned}$$

The first vector is orthogonal to $\mathrm{Span}(S-i)$, whereas the remaining vectors in the expression live in $\mathrm{Span}(S-i)$. Because $\mathrm{Span}(U\cap S)\subseteq \mathrm{Span}(S-i)$, we also have that $R_{S-i}(R_{U\cap S}(v_i)) = R_{S-i}(v_i)$, and therefore that

$$\begin{aligned}\|R_{U\cap S}(\tilde{v}_\ell)\|_2^2 &\geq \|\boldsymbol{W}_{\ell i}R_{S-i}(R_{U\cap S}(v_i))\|_2^2 \\ &\geq \|\boldsymbol{W}_{\ell i}R_{S-i}(v_i)\|_2^2 \\ &\geq \boldsymbol{W}_{\ell i}^2 m\end{aligned}$$

where we have recalled from above that $\|R_{S-i}(v_i)\|_2^2 \geq m$. If ℓ is such that $|W_{\ell i}| \geq \eta$, then putting everything together gives that

$$(\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell U}\boldsymbol{\Sigma}_U^{-1}\boldsymbol{\Sigma}_{U\ell}) - (\boldsymbol{\Sigma}_{\ell\ell} - \boldsymbol{\Sigma}_{\ell S}\boldsymbol{\Sigma}_S^{-1}\boldsymbol{\Sigma}_{S\ell}) \geq \left(1-\frac{kC}{kC+\sigma^2}\right)^2 \eta^2 m > 0.$$

A.7.7 Proof of Theorem 3

The reverse direction is obvious. We show the forward direction. For simplicity we'll assume X is mean zero. For the general case apply the same argument to $X-\mu$. Since $Z\in\mathbb{R}^k$ has covariance $\boldsymbol{I}_k$, if $Z=\boldsymbol{B}X$ then $\boldsymbol{B}\in\mathbb{R}^{k\times p}$ must be rank k. Plugging $Z=\boldsymbol{B}X$ into our factor model equation, we get that $X = \boldsymbol{W}\boldsymbol{B}X+\epsilon$, where ϵ is uncorrelated with $\boldsymbol{B}X$. By Corollary 9, $\boldsymbol{W} = \boldsymbol{\Sigma}_{X,\boldsymbol{B}X}\boldsymbol{\Sigma}_{\boldsymbol{B}X}^{-1} = \boldsymbol{\Sigma}\boldsymbol{B}^\top(\boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}^\top)^{-1}$. Our expression $X=\boldsymbol{W}\boldsymbol{B}X+\epsilon$ follows the definition of regression component decomposition given in [Schönemann and Steiger, 1976]. Referencing [Schönemann and Steiger, 1976, Equation 2.6] and the discussion following it, the covariance $\boldsymbol{D}$ of ϵ has rank exactly $p-k$. Since $\boldsymbol{D}$ is diagonal, this means exactly k of its entries are 0. Let $S\subseteq[p]$ be the subset of indices such if $i\in S$ then $\boldsymbol{D}_{ii}=0$ and thus $\epsilon_i=0$. Then $X_S = \boldsymbol{W}_{S\bullet}\boldsymbol{B}X$. It clear that since ϵ is uncorrelated with $\boldsymbol{B}X$, it must also be uncorrelated with X_S. Since X_S and $\boldsymbol{B}X$ have full rank covariances, $\boldsymbol{W}_{S\bullet}$ must be full rank and thus invertible, and we can write $\boldsymbol{B}X = \boldsymbol{W}_{S\bullet}^{-1}X_S$.

Plugging this into our original factor model we have $X_{-S} = \widetilde{\boldsymbol{W}} X_S + \epsilon$ where $\widetilde{\boldsymbol{W}} = \boldsymbol{W}_{-S\bullet}\boldsymbol{W}_{S\bullet}^{-1}$ and X_S is uncorrelated with ϵ.

A.7.8 Proof of Lemma 1

By Lemma A.6.1 the residual $R(Y, Z)$ is given by $Y - \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_Z^+ Z$. By standard properties of projections in finite dimensional Hilbert spaces, $R(X, X_S) = R(R(X, X_U), R(X_i, X_U))$. We can therefore compute,

$$\begin{aligned}
\boldsymbol{\Sigma}_{R(X,X_S)} &= \boldsymbol{\Sigma}_{R(R(X,X_U),R(X_i,X_U)))} \\
&= \boldsymbol{\Sigma}_{R(X,X_U)} - \boldsymbol{\Sigma}_{R(X,X_U),R(X_i,X_U)}\boldsymbol{\Sigma}^+_{R(X_i,X_U)}\boldsymbol{\Sigma}_{R(X_i,X_U),R(X,X_U)} \\
&= \boldsymbol{\Sigma}_{R(X,X_U)} - (\boldsymbol{\Sigma}_{R(X,X_U)})_{\bullet i}(\boldsymbol{\Sigma}_{R(X,X_U)})_{i\bullet}/(\boldsymbol{\Sigma}_{R(X,X_U)})_{ii} \cdot I_{(\boldsymbol{\Sigma}_{R(X,X_U)})_{ii}>0}.
\end{aligned}$$

Where the second equality follows from Theorem 12(iv). Also using Theorem 12(iv) we can compute that $(\boldsymbol{\Sigma}_{R(X,X_U)})_{\bullet i} = \boldsymbol{\Sigma}_{\bullet i} - \boldsymbol{\Sigma}_{\bullet T}\boldsymbol{\Sigma}_T^+\boldsymbol{\Sigma}_{T,i}$.

A.7.9 Proof of Lemma A.2.1

Suppose $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ are drawn from a distribution satisfying (A.1) for a size-k subset S, $k < p$, and population mean μ.

First we show that the quantiles of $T(S)$ from (3.4) are bounded by those of (3.5). We will perform our analysis conditional on $x_S^{(1)}, \ldots, x_S^{(n)}$ (hence, the independence between X_S and ϵ is crucial). Let $\boldsymbol{X}_S \in \mathbb{R}^{n\times k}$ and $\boldsymbol{X}_{-S} \in \mathbb{R}^{n\times(p-k)}$ be the matrices with rows $x_S^{(i)}$ and $x_{-S}^{(i)}$. Defining $e^{(i)} = (x_{-S}^{(i)} - \mu_{-S}) - \boldsymbol{W}(x_S^{(i)} - \mu_S)$ and $\hat{e}^{(i)} = (x_{-S}^{(i)} - \hat{\mu}_{-S}) - \hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^+(x_S^{(i)} - \hat{\mu}_S)$, let the matrices $\boldsymbol{E}, \hat{\boldsymbol{E}} \in \mathbb{R}^{n\times(p-k)}$ have rows $e^{(i)}$ and $\hat{e}^{(i)}$. Note that $e^{(i)}$ are the unique factors ϵ for the ith sample. Having conditioned on the $x_S^{(i)}$, we are exactly in the setting of fixed-X Ordinary Least Squares (OLS). In particular, we are running a multi-response regression of $\boldsymbol{X}_{-S}$ on $\boldsymbol{X}_S$ where the responses are independent. The $\hat{\boldsymbol{E}}$ are the residuals from OLS, while $\boldsymbol{E}$ are the residuals one would get from using the population regression coefficients. With this in mind, let $\boldsymbol{H} \in \mathbb{R}^{n\times n}$ be the orthogonal projection matrix onto the span of the columns of $\boldsymbol{X}_S$ and $\mathbf{1} \in \mathbb{R}^n$, a n-dimensional vector of ones. By construction, $\boldsymbol{H}$ has rank $r \le k+1$. Therefore,

$$\begin{aligned}
\hat{\boldsymbol{\Sigma}}_{-S} - \hat{\boldsymbol{\Sigma}}_{-S,S}\hat{\boldsymbol{\Sigma}}_S^+\hat{\boldsymbol{\Sigma}}_{S,-S} &= \frac{1}{n}\hat{\boldsymbol{E}}^\top\hat{\boldsymbol{E}} \\
&= \frac{1}{n}\boldsymbol{X}_{-S}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{X}_{-S} \\
&= \frac{1}{n}((\boldsymbol{X}_S - \mathbf{1}\mu_S^\top)\boldsymbol{W}^\top + \mathbf{1}\mu_{-S}^\top + \boldsymbol{E})^\top(\boldsymbol{I}_n - \boldsymbol{H})((\boldsymbol{X}_S - \mathbf{1}\mu_S^\top)\boldsymbol{W}^\top + \mathbf{1}\mu_{-S}^\top + \boldsymbol{E}) \\
&= \frac{1}{n}\boldsymbol{E}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}.
\end{aligned}$$

Let $\boldsymbol{N} \in \mathbb{R}^{n\times(p-k)}$ and $\boldsymbol{M} \in \mathbb{R}^{(n-r)\times(p-k)}$ be random matrices with i.i.d $N(0,1)$ entries and $\boldsymbol{R} \in \mathbb{R}^{n\times n}$ be a rotation such that $\boldsymbol{R}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{R}$ is diagonal matrix with 1 on the first $n-r$ diagonal entries and 0

otherwise. Then we can compute

$$\begin{aligned} n \log\left(\frac{|\mathrm{Diag}(\boldsymbol{E}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E})|}{|\boldsymbol{E}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}|}\right) &= n \log\left(\frac{|\mathrm{Diag}(\boldsymbol{D}^{-1/2}\boldsymbol{E}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}\boldsymbol{D}^{-1/2})|}{|\boldsymbol{D}^{-1/2}\boldsymbol{E}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}\boldsymbol{D}^{-1/2}|}\right) \\ &\overset{d}{=} n \log\left(\frac{|\mathrm{Diag}(\boldsymbol{N}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{N}|)}{|\boldsymbol{N}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{N}|}\right) \\ &\overset{d}{=} n \log\left(\frac{|\mathrm{Diag}(\boldsymbol{N}^\top\boldsymbol{R}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{R}\boldsymbol{N}|)}{|\boldsymbol{N}^\top\boldsymbol{R}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{R}\boldsymbol{N}|}\right) \\ &\overset{d}{=} n \log\left(\frac{|\mathrm{Diag}(\boldsymbol{M}^\top\boldsymbol{M})|}{|\boldsymbol{M}^\top\boldsymbol{M}|}\right) \\ &\overset{d}{=} n \log\left(\frac{|\mathrm{Diag}(W_{p-k}(I_{p-k}, n-r))|}{|W_{p-k}(I_{p-k}, n-r)|}\right) \\ &\overset{d}{=} n \sum_{j=2}^{p-k} \log\left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-r-j+1}}\right). \end{aligned}$$

where $W_{p-k}(I_{p-k}, n-r)$ is a Wishart distribution and $\{\chi^2_\ell\}, \{\tilde{\chi}^2_\ell\}$ are mutually independent chi-squared random variables with degrees of freedom specified by their subscript. The final equality in distribution follows from the Bartlett decomposition of Wisharts [Anderson et al., 1958, Corollary 7.2.1]. For clarity, we specify that the final distribution is a point mass at zero when $k = p-1$.

Recall that these equalities in distribution are conditional. Since $r \leq k+1$, it is clear that the quantiles of $n\sum_{j=2}^{p-k} \log\left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-k-j}}\right)$ are always at least as large as those of $n\sum_{j=2}^{p-k} \log\left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-r-j+1}}\right)$. Since the conditional quantiles of $T(S)$ are always bounded above by the quantiles of $n\sum_{j=2}^{p-k} \log\left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-k-j}}\right)$, the marginal quantiles must be as well.

Now we can easily establish the final claim. Recall that $Q_{n,p,k}(1-\alpha)$ is the $(1-\alpha)$-quantile of $n\sum_{j=2}^{p-k} \log\left(1 + \frac{\tilde{\chi}^2_{j-1}}{\chi^2_{n-k-j}}\right)$. Since $T_k \leq T(S)$,

$$P(T_k > Q_{n,p,k}(1-\alpha)) \leq P(T(S) > Q_{n,p,k}(1-\alpha)) \leq \alpha.$$

A.7.10 Proof of Lemma A.2.2

First, some preliminaries. Define the function $\mathrm{vec}(\cdot)$ that maps a symmetric matrix $M \in \mathbb{R}^{q\times q}$ to a vector $m \in \mathbb{R}^{q(q+1)/2}$ that contains all it's unique entries. If $m = \mathrm{vec}(M)$, then we will index m by m_{ij} for $i \leq j$ so that $m_{ij} = M_{ij}$. The inverse map is denoted $\mathrm{vec}^{-1}(\cdot)$. We will use the stochastic calculus notation of [Vaart, 1998, Chapter 2] and also freely refer to results from this chapter.

Suppose we observe $x^{(1)}, \ldots, x^{(n)}$ drawn from a distribution that satisfies (3.2) for some size-k subset S where $k < p-1$. In the case that $k = p-1$, $T(S)$ is a point mass at zero and the result is obvious. Recall $\boldsymbol{E}$ and $\boldsymbol{H}$ from the proof of Lemma A.2.1 and consider the function $g(M) = \log(|\mathrm{Diag}(M)|) - \log(|M|)$. Computations in the proof of Lemma A.2.1 imply that $T(S) = ng(\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}/n)$, and it thus suffices to show that $ng(\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}/n) \rightsquigarrow \chi^2_{(p-k)(p-k-1)/2}$. We will show this result in steps.

Step One - CLT for $\mathrm{vec}(\boldsymbol{E}^T\boldsymbol{E}/n)$: Let $a = \mathrm{vec}(\boldsymbol{E}^T\boldsymbol{E}/n)$ and $\tilde{a} \in \mathbb{R}^{p(p-1)/2}$ be only the entries of a that correspond to the non-diagonal entries of $\boldsymbol{E}^T\boldsymbol{E}/n$. We index $\tilde{a}$ as $\tilde{a}_{ij}$, $i < j$. The standard multivariate

CLT tells us that $\sqrt{n}\tilde{a} \rightsquigarrow N(0, \boldsymbol{V})$ where $\boldsymbol{V} \in \mathbb{R}^{p(p-1)/2 \times p(p-1)/2}$ is such that $\boldsymbol{V}_{ij,st} = \boldsymbol{D}_{ii}\boldsymbol{D}_{jj}$ if $i = s$, $j = t$ and 0 otherwise. Furthermore, by the weak law of large numbers $a_{ii} - \boldsymbol{D}_{ii} = o_p(1)$.

Step Two - Controlling vec($\boldsymbol{E^T HE/n}$): We want to control $\boldsymbol{E}^T\boldsymbol{H}\boldsymbol{E}/n$, the difference between $\boldsymbol{E}^T\boldsymbol{E}/n$ and $\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}/n$. Particularly we want to show that $(\boldsymbol{E}^T\boldsymbol{H}\boldsymbol{E}/n)_{ij} = o_p(n^{-1/2})$, $i \leq j$. To do so it suffices to show that $\|\boldsymbol{E}^T\boldsymbol{H}\boldsymbol{E}\|_F^2 = O_p(1)$. Noting that $\boldsymbol{H}$ is the sum of the outer product of at most $k+1$ unit vectors, it suffices to show that $\|\boldsymbol{E}^T vv^T\boldsymbol{E}\|_F^2 = \|\boldsymbol{E}^T v\|_2^4 = O_p(1)$ for an arbitrary sequence of unit vectors $v \in \mathbb{R}^n$. To show this, it suffices to show that $\sum_{i=1}^n e_\ell^{(i)} v_i = (\boldsymbol{E}^T v)_\ell = O_p(1)$ for all $1 \leq \ell \leq p$. But $\mathrm{Var}(\sum_{i=1}^n e_\ell^{(i)} v_i) = \boldsymbol{D}_{\ell\ell}\sum_{i=1}^n v_i^2 = \boldsymbol{D}_{\ell\ell}$ and $\boldsymbol{E}[\sum_{i=1}^n e_\ell^{(i)} v_i] = 0$, so this must be the case by Chebyshev.

Step Three - CLT for vec($\boldsymbol{E^T(I_n - H)E/n}$): Let $a_H = \mathrm{vec}(\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}/n)$ and $\tilde{a}_H \in \mathbb{R}^{p(p-1)/2}$ be only the entries of a that correspond to the non-diagonal entries of $\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}/n$. Step Two implies that $\sqrt{n}(\tilde{a} - \tilde{a}_H) = o_p(1)$, so $\sqrt{n}\tilde{a}_H \rightsquigarrow N(0, \boldsymbol{V})$ by Slutsky's lemma and Step One, where $\boldsymbol{V}$ is as in Step One. Furthermore, since $a_{ii} - (a_H)_{ii} = o_p(1)$, we also have $(a_H)_{ii} - D_{ii} = o_p(1)$ by Slutsky's.

Step Four - Delta Method for $g(\boldsymbol{E^T(I_n - H)E/n})$: Define the function $f : \mathbb{R}^{q(q+1)/2} \to \mathbb{R}$ by $f(m) = g(\mathrm{vec}^{-1}(m))$. We will consider the domain of f to be the open set of m such that $\mathrm{vec}^{-1}(m)$ is a positive definite matrix. On this domain f is C^∞. We will take a second order Taylor expansion of the function f around an $d \in \mathbb{R}^{(p-k)(p-k+1)/2}$ such that $\mathrm{vec}^{-1}(d) = \boldsymbol{D}$. Note then $d_{ij} = 0$ for $i < j$. We can compute the first and second order partial derivatives of f using [Petersen and Pedersen, 2008, Section 2]:

$$\frac{\partial f(m)}{\partial m_{ij}} = \begin{cases} -2(\mathrm{vec}^{-1}(m))_{ij}^{-1} & i \neq j \\ m_{ii}^{-1} - (\mathrm{vec}^{-1}(m))_{ii}^{-1} & i = j, \end{cases}$$

$$\frac{\partial^2 f(m)}{\partial m_{ij}\partial m_{st}} = \begin{cases} 2((\mathrm{vec}^{-1}(m))_{is}^{-1}(\mathrm{vec}^{-1}(m))_{jt}^{-1} + (\mathrm{vec}^{-1}(m))_{it}^{-1}(\mathrm{vec}^{-1}(m))^{-1})_{js} & i \neq j, s \neq t \\ 2(\mathrm{vec}^{-1}(m))_{is}^{-1}(\mathrm{vec}^{-1}(m))_{js}^{-1} & i \neq j, s = t \\ ((\mathrm{vec}^{-1}(m))_{is}^{-1})^2 - m_{ii}^{-2} I_{i=s} & i = j, s = t. \end{cases}$$

By repeated application of [Petersen and Pedersen, 2008, 9.1.2], it's clear that $f \geq 0$ on its domain. Further $f(d) = 0$. In line with this, all the first order partial derivatives of f are zero at d. Many of the second order partial derivatives are also zero. Taylor's theorem gives the following expansion of f around d:

$$f(m) = \sum_{i<j} \boldsymbol{D}_{ii}^{-1}\boldsymbol{D}_{jj}^{-1} m_{ij}^2 + \sum_{i_1 \leq j_1, i_2 \leq j_2, i_3 \leq j_3} c_{i_1 j_1 i_2 j_2 i_3 j_3} \frac{\partial f^3(d + t(m-d))}{\partial m_{i_1 j_1}\partial m_{i_2 j_2}\partial m_{i_3 j_3}} (m_{i_1 j_1} - d_{i_1 j_1})(m_{i_2 j_2} - d_{i_2 j_2})(m_{i_3 j_3} - d_{i_3 j_3}).$$

for some $t \in [0, 1]$ and constants $c_{i_1 j_1 i_2 j_2 i_3 j_3}$. It is then clear that

$$ng(\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}/n) = nf(\mathrm{vec}(\boldsymbol{E}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{E}/n)) = \sum_{i<j} \boldsymbol{D}_{ii}^{-1}\boldsymbol{D}_{jj}^{-1}(\sqrt{n}(a_H)_{ij})^2 + o_p(1),$$

and by continuous mapping theorem $ng(\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}/n) \rightsquigarrow \chi^2_{(p-k)(p-k-1)/2}$.

A.7.11 Proof of Lemma A.2.3

The case that $k = p - 1$ is obvious, so we consider $k < p - 1$ without loss of generality.

First we will argue that the $Q_{n,p,k}(1-\alpha)$, the $(1-\alpha)$-quantile of (3.5), converges to $Q_{p,k}(1-\alpha)$, the $(1-\alpha)$-quantile of a $\chi^2_{(p-k)(p-k-1)/2}$ distribution as the number of samples n tends to infinty. Consider observing $x^{(1)}, \ldots, x^{(n)}$ samples drawn from a distribution that satisfies (3.2) for a size-k subset S, and suppose the principal variable distribution F admits a density. Following the proof of Lemma A.2.1, we see that $T(S)$ will have distribution exactly (3.5) since $\hat{\boldsymbol{\Sigma}}_S$ will be full-rank with probability one. But Lemma A.2.2 tells us that $T(S) \rightsquigarrow \chi^2_{(p-k)(p-k-1)/2}$. This implies that $Q_{n,p,k}(1-\alpha)$ tends to $Q_{p,k}(1-\alpha)$ as $n \to 0$.

Now, by Slutsky's and Lemma A.2.2 we know that $T(S) - Q_{n,p,k}(1-\alpha) \rightsquigarrow \chi^2_{(p-k)(p-k-1)/2} - Q_{p,k}(1-\alpha)$. Since 0 is a continuity point of $\chi^2_{(p-k)(p-k-1)/2} - Q_{p,k}(1-\alpha)$ we have that

$$\begin{aligned}\limsup_{n\to\infty} P(T_k - Q_{n,p,k}(1-\alpha) > 0) &\leq \limsup_{n\to\infty} P(T(S) - Q_{n,p,k}(1-\alpha) > 0) \\ &= P(\chi^2_{(p-k)(p-k-1)/2} - Q_{p,k}(1-\alpha) > 0) \\ &= \alpha.\end{aligned}$$

A.7.12 Proof of Theorem 4

First we show the asymptotic result. Consider a sample $x^{(1)}, \ldots, x^{(n)}$ from a distribution P and suppose that k^* is the smallest k for which P belongs to the subset factor model (3.2). Note that every distribution belongs to the $(k-1)$-dimensional factor model so $k^* \leq k + 1$. Recall that $\hat{k}$ is the smallest k for which $T_k \leq Q_{n,p,k}(1-\alpha)$. Then, using Lemma A.2.3,

$$\limsup_{n\to\infty} P(\hat{k} > k^*) \leq \limsup_{n\to\infty} P(T_{k^*} > Q_{n,p,k^*}(1-\alpha)) \leq \alpha$$

Now, for the finite sample result. Let k^* be the smallest k for which P satisfies the model (A.1) for some size-k subset S, and take $\hat{k}$ as before. It is now possible that $k^* = p$, but $P(\hat{k} > k) = 0$ if so. Thus, without loss of generality, we can suppose that $k^* < p$. Then, from lemma A.2.1

$$P(\hat{k} > k^*) \leq P(T_{k^*} > Q_{n,p,k^*}(1-\alpha)) \leq \alpha.$$

A.7.13 Proof of Lemma A.2.4

First we show that a distribution P satisfies (A.3) with a size-k set S if and only if, for the population covariance $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}_{-S} - \boldsymbol{\Sigma}_{-S,S}\boldsymbol{\Sigma}_S^+\boldsymbol{\Sigma}_{S,-S}$ is diagonal and positive definite. First we show the forward direction. We know that $X_{-S} - \mu_{-S} = \boldsymbol{W}(X_S - \mu_S) + \epsilon$ where $\mathrm{Cov}(X_S, \epsilon) = \boldsymbol{0}$. Then Corollary 9 tells us that $X_{-S} - \mu_{-S} = \boldsymbol{\Sigma}_{-S,S}\boldsymbol{\Sigma}_S^+(X_S - \mu_S) + \epsilon$. Via this, we can compute that $\boldsymbol{\Sigma}_S = \boldsymbol{\Sigma}_{-S,S}\boldsymbol{\Sigma}_S^+\boldsymbol{\Sigma}_{S,-S} + \boldsymbol{D}$ where $\boldsymbol{D} \succ 0$ so $\boldsymbol{\Sigma}_S - \boldsymbol{\Sigma}_{-S,S}\boldsymbol{\Sigma}_S^+\boldsymbol{\Sigma}_{S,-S}$ is diagonal and positive definite. For the reverse direction Corollary 9 tells us that $X_{-S} - \mu_{-S} = \boldsymbol{\Sigma}_{-S,S}\boldsymbol{\Sigma}_S^+(X_S - \mu_S) + \epsilon$ where $\mathrm{Cov}(X_S, \epsilon) = 0$. A simple computation shows that ϵ must have diagonal, positive definite covariance and thus the distribution satisfies (A.3) with a size-k set S .

Now, consider $n > p$ samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution P that satisfies (A.3) for some size-k set S and suppose that P has population covariance $\mathbf{\Sigma}$ such that $\mathbf{\Sigma}_S \succ 0$. Then it must be the case that $\mathbf{\Sigma} \succ 0$. Thus, for every size-k subset U, $|\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U}| > 0$. Since $\hat{\mathbf{\Sigma}}$ converges almost surely to $\mathbf{\Sigma}$ as $n \to \infty$, we have that, for every size-k subset U,

$$\log\left(\frac{|\mathrm{Diag}(\hat{\mathbf{\Sigma}}_{-U} - \hat{\mathbf{\Sigma}}_{-U,U}\hat{\mathbf{\Sigma}}_U^+\hat{\mathbf{\Sigma}}_{U,-U})|}{|\hat{\mathbf{\Sigma}}_{-U} - \hat{\mathbf{\Sigma}}_{-U,U}\hat{\mathbf{\Sigma}}_U^+\hat{\mathbf{\Sigma}}_{U,-U}|}\right) \tag{A.15}$$

converges almost surely to

$$\log\left(\frac{|\mathrm{Diag}(\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^{-1}\mathbf{\Sigma}_{U,-U})|}{|\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^{-1}\mathbf{\Sigma}_{U,-U}|}\right) \tag{A.16}$$

As a consequence of [Petersen and Pedersen, 2008, 9.1.2], (A.16) is zero if and only if $\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^{-1}\mathbf{\Sigma}_{U,-U}$ is diagonal and otherwise it is strictly greater than zero. Thus, By the discussion above, it is zero when P satisfies (A.3) with U, and otherwise it is strictly greater than zero. Since $\hat{S}$ minimizes $T(U)$ it must also minimize (A.15). And since (A.15) converges almost surely to (A.16), eventually $\hat{S}$ will be one of the U for which (A.16) is zero, meaning P will eventually satisfy (A.3) with $\hat{S}$.

A.7.14 Proof of Lemma A.2.5

We adopt the notation from the proof of lemma A.2.1. Similarly we perform our analysis conditional on the $x^{(i)}$. Recall that

$$\hat{\mathbf{\Sigma}}_{-S} - \hat{\mathbf{\Sigma}}_{-S,S}\hat{\mathbf{\Sigma}}_S^+\hat{\mathbf{\Sigma}}_{S,-S} = \frac{1}{n}\boldsymbol{E}^\top(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}.$$

Since we have assumed that F has a density, we know that $\mathrm{rank}(\boldsymbol{H}) = k+1$. We can then similarly compute

$$\begin{aligned}
\widetilde{T}(S) &= n\log\left(\frac{(\mathrm{Tr}(\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E})/(p-k))^{p-k}}{|\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}|}\right) \\
&= n\log\left(\frac{(\mathrm{Tr}(\sigma^{-1}\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}\sigma^{-1})/(p-k))^{p-k}}{|\sigma^{-1}\boldsymbol{E}^T(\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{E}\sigma^{-1}|}\right) \\
&\overset{d}{=} n\log\left(\frac{(\mathrm{Tr}(W_{p-k}(I_{p-k}, n-k-1))/(p-k))^{p-k}}{|W_{p-k}(I_{p-k}, n-k-1)|}\right) \\
&\overset{d}{=} n\log\left(\left(\frac{\tilde{\chi}^2_{(p-k)(p-k-1)/2} + \sum_{j=1}^{p-k}\chi^2_{n-k-j}}{p-k}\right)^{p-k} \Big/ \left(\prod_{j=1}^{p-k}\chi^2_{n-k-j}\right)\right)
\end{aligned}$$

where $W_{p-k}(I_{p-k}, n-r)$ is a Wishart distribution, $\{\chi^2_\ell\}, \{\tilde{\chi}^2_\ell\}$ are mutually independent chi-squared random variables with degrees of freedom specified by their subscript, and again the final equality in distribution follows from the Bartlett decomposition of Wisharts [Anderson et al., 1958, Corollary 7.2.1]. Although these equalities in distribution are conditional, since the final distribution does not depend on the $x^{(i)}$, it also holds marginally. Note that the distribution is a point mass at zero when $k = p-1$.

Now we can easily establish the final claim. Recall that $\widetilde{Q}_{n,p,k}(1-\alpha)$ is the $(1-\alpha)$-quantile of

$$n \log \left(\left(\frac{\tilde{\chi}^2_{(p-k)(p-k-1)/2} + \sum_{j=1}^{p-k} \chi^2_{n-k-j}}{p-k} \right)^{p-k} \Big/ \left(\prod_{j=1}^{p-k} \chi^2_{n-k-j} \right) \right)$$

Since $\widetilde{T}_k \leq \widetilde{T}(S)$,

$$P(\widetilde{T}_k > \widetilde{Q}_{n,p,k}(1-\alpha)) \leq P(\widetilde{T}(S) > \widetilde{Q}_{n,p,k}(1-\alpha)) \leq \alpha.$$

A.7.15 Proof of Proposition A.2.1

Let k^* be the smallest k for which P satisfies the model (A.1) for some size-k subset S, and recall that $\hat{k}$ is the smallest k for which $\widetilde{T}_k \leq \widetilde{Q}_{n,p,k}(1-\alpha)$. It is possible that $k^* = p$, but then $P(\hat{k} > k) = 0$ if so. Thus, without loss of generality, we can suppose that $k^* < p$. Then, from lemma A.2.5,

$$P(\hat{k} > k^*) \leq P(\widetilde{T}_{k^*} > \widetilde{Q}_{n,p,k^*}(1-\alpha)) \leq \alpha.$$

A.7.16 Proof of Lemma A.2.6

First we claim that a distribution P satisfies (A.6) with a size-k set S if and only if, for the population covariance $\mathbf{\Sigma}$, $\mathbf{\Sigma}_{-S} - \mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^+\mathbf{\Sigma}_{S,-S}$ is isotropic, i.e., equal to $\eta \boldsymbol{I}_{p-k}$ for some $\eta > 0$. The argument is identical to that in the proof of Lemma A.2.4.

Now, consider $n > p$ samples $x^{(1)}, \ldots, x^{(n)} \in \mathbb{R}^p$ drawn from a distribution P that satisfies (A.6) for some size-k subset S. Further suppose that, at the population level no set of k variables perfectly linearly reconstruct the rest. This implies that $\mathrm{Tr}(\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U})$ is never zero for any size-k subset U. Then, since $\hat{\mathbf{\Sigma}}$ converges almost surely to $\mathbf{\Sigma}$ as $n \to \infty$, we can guarantee that

$$\log \left(\frac{(\mathrm{Tr}(\hat{\mathbf{\Sigma}}_{-U} - \hat{\mathbf{\Sigma}}_{-U,U}\hat{\mathbf{\Sigma}}_U^+\hat{\mathbf{\Sigma}}_{U,-U})/(p-k))^{(p-k)}}{|\hat{\mathbf{\Sigma}}_{-U} - \hat{\mathbf{\Sigma}}_{-U,U}\hat{\mathbf{\Sigma}}_U^+\hat{\mathbf{\Sigma}}_{U,-U}|} \right) \tag{A.17}$$

converges almost surely to

$$\log \left(\frac{(\mathrm{Tr}(\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U})/(p-k))^{(p-k)}}{|\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U}|} \right) \tag{A.18}$$

We claim that (A.18) is zero if and only if $\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U}$ is of the form $\eta \boldsymbol{I}_{p-k}$. To see this note that (A.18) is (resp. strictly) larger than zero if and only if

$$\frac{\mathrm{Tr}(\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U})/(p-k)}{|\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U}|^{1/(p-k)}} \tag{A.19}$$

is (resp. strictly) larger than one. But (A.19) is always larger than or equal to one because it is the ratio of the arithmetic mean to the geometric mean of the eigenvalues of $\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U}$. Equality is only attained when all the eigenvalues are equal, i.e., when $\mathbf{\Sigma}_{-U} - \mathbf{\Sigma}_{-U,U}\mathbf{\Sigma}_U^+\mathbf{\Sigma}_{U,-U}$ is of the form $\eta \boldsymbol{I}_{p-k}$, which completes the claim.

By the discussion above, this quantity will be zero when P satisfies (A.6) with U, and otherwise it will be

strictly greater than zero. Since $\widetilde{S}$ minimizes $\widetilde{T}(U)$, eventually it will be one of the U for which this quantity is zero, and thus eventually P will satisfy (A.3) with $\widetilde{S}$.

A.7.17 Correctness of subset search algorithms

We provide computations which justify the correctness algorithms in Section 2.3.1 as well as Appendix A.3.

Minimizing $T(\cdot)$: [Petersen and Pedersen, 2008, 9.1.2] and Lemma 1 jointly tell us that $\log|\mathbf{\Sigma}_{U+i}| = \log|\hat{\mathbf{\Sigma}}_U| + \log((\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii})$. Using Lemma 1, we can compute

$$\begin{aligned}
&\log|\hat{\mathbf{\Sigma}}_{U+i}| + \mathrm{Tr}(\log(\mathrm{Diag}(\hat{\mathbf{\Sigma}}_{R(X_{-(U+i)},X_{U+i})})))\\
&\quad= \log|\hat{\mathbf{\Sigma}}_U| + \log((\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii}) + \mathrm{Tr}(\log(\mathrm{Diag}((\hat{\mathbf{\Sigma}}_{R(X,X_{U+i})})_{-(U+i)})))\\
&\quad= \log|\hat{\mathbf{\Sigma}}_U| + \log((\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii})\\
&\qquad+ \mathrm{Tr}(\log(\mathrm{Diag}((\hat{\mathbf{\Sigma}}_{R(X,X_U)} - (\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{\bullet i}(\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{i\bullet}/(\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii}\cdot I_{(\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii}>0})_{-(U+i)})))\\
&\quad= \log|\hat{\mathbf{\Sigma}}_U| + \log((\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii}) + \sum_{j\notin U+i}\log((\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{jj} - (\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ij}^2/(\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii}\cdot I_{(\hat{\mathbf{\Sigma}}_{R(X,X_U)})_{ii}>0})
\end{aligned}$$

which is sufficient to imply that the $i \notin U$ we select minimizes the objective over subsets $U+i$.

Minimizing $\widetilde{T}(\cdot)$: Following similar reasoning to above, we can compute

$$\begin{aligned}
&\log|\mathbf{\Sigma}_{U+i}| + (p-k)\log(\mathrm{Tr}(\mathbf{\Sigma}_{R(X_{-(U+i)},X_{U+i})})/(p-k))\\
&\quad= \log|\mathbf{\Sigma}_U| + \log((\mathbf{\Sigma}_{R(X,X_U)})_{ii}) + (p-k)\log(\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_{U+i})})) - (p-k)\log(p-k)\\
&\quad= \log|\mathbf{\Sigma}_U| + \log((\mathbf{\Sigma}_{R(X,X_U)})_{ii})\\
&\qquad+ (p-k)\log(\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_U)} - (\mathbf{\Sigma}_{R(X,X_U)})_{\bullet i}(\mathbf{\Sigma}_{R(X,X_U)})_{i\bullet}/(\mathbf{\Sigma}_{R(X,X_U)})_{ii}\cdot I_{(\mathbf{\Sigma}_{R(X,X_U)})_{ii}>0})) - (p-k)\log(p-k)\\
&\quad= \log|\mathbf{\Sigma}_U| + \log((\mathbf{\Sigma}_{R(X,X_U)})_{ii}) + (p-k)\log(\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_U)}) - \|(\mathbf{\Sigma}_{R(X,X_U)})_{\bullet i}\|_2^2/(\mathbf{\Sigma}_{R(X,X_U)})_{ii}\cdot I_{(\mathbf{\Sigma}_{R(X,X_U)})_{ii}>0})\\
&\qquad- (p-k)\log(p-k)
\end{aligned}$$

which is sufficient to imply that the $i \notin U$ we select minimizes the objective over subsets $U+i$.

McCabe's first criterion: If there exists a k by k principal sub-matrix of $\mathbf{\Sigma}$ that is full rank then minimizing $|\mathbf{\Sigma}_{-S} - \mathbf{\Sigma}_{-S,S}\mathbf{\Sigma}_S^+\mathbf{\Sigma}_{S,-S}|$ is equivalent to maximizing $|\mathbf{\Sigma}_S|$. Consider having a currently selected subset U. We know from above that $|\mathbf{\Sigma}_{U+i}| = |\mathbf{\Sigma}_U|\cdot(\mathbf{\Sigma}_{R(X,U)})_{ii}$. This is sufficient to imply that the $i \notin U$ we select minimizes the objective over subsets $U+i$.

McCabe's second criterion: Consider having currently selected a subset U. For $i \notin U$ let $\beta = (\mathbf{\Sigma}_{R(X,X_U)})_{\bullet i}$. Then from Lemma 1:

$$\begin{aligned}
\mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_{U+i})}) &= \mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_U)} - \frac{\beta\beta^\top}{\beta_i}\cdot I_{\beta_i>0})\\
&= \mathrm{Tr}(\mathbf{\Sigma}_{R(X,X_U)}) - \frac{||\beta||_2^2}{\beta_i}\cdot I_{\beta_i>0}
\end{aligned}$$

which is sufficient to imply that the $i \notin U$ we select minimizes the objective over subsets $U + i$.

McCabe's third criterion: Consider having currently selected a subset U. For $i \notin U$ let $\beta = (\mathbf{\Sigma}_{R(X,X_U)})_{\bullet i}$. Then from Lemma 1:

$$\begin{aligned}\|\mathbf{\Sigma}_{R(X,X_{U+i})}\|_F^2 &= \|\mathbf{\Sigma}_{R(X,X_U)} - \frac{\beta\beta^\top}{\beta_i} \cdot I_{\beta_i>0}\|_F^2 \\ &= \mathrm{Tr}((\mathbf{\Sigma}_{R(X,X_U)} - \frac{\beta\beta^\top}{\beta_i} \cdot I_{\beta_i>0})^\top(\mathbf{\Sigma}_{R(X,X_U)} - \frac{\beta\beta^\top}{\beta_i} \cdot I_{\beta_i>0})) \\ &= \|\mathbf{\Sigma}_{R(X,X_U)}\|_F^2 + \left[\frac{\|\beta\|_2^4}{\beta_i^2} - \frac{2\beta^\top \mathbf{\Sigma}_{R(X,X_U)}\beta}{\beta_i}\right] \cdot I_{\beta_i>0}\end{aligned}$$

which is sufficient to imply that the $i \notin U$ we select minimizes the objective over subsets $U + i$.

McCabe's fourth criterion: McCabe's fourth criterion suggests finding a subset S that maximizes $\mathrm{Tr}(\mathbf{\Sigma}_S^+ \mathbf{\Sigma}_{S,-S}\mathbf{\Sigma}_{-S}^+\mathbf{\Sigma}_{-S,S})$.

We briefly justify that our generalized objective has the same interpretation. Fix a subset S. It is well known that $\mathrm{Tr}(\mathbf{\Sigma}_S^+ \mathbf{\Sigma}_{S,-S}\mathbf{\Sigma}_{-S}^+\mathbf{\Sigma}_{-S,S})$ is the sum of the squared canonical correlations when $\mathbf{\Sigma} \succ \mathbf{0}$ [McCabe, 1984]. When $\mathbf{\Sigma}$ is singular, suppose that the ranks of $\mathbf{\Sigma}_S$ and $\mathbf{\Sigma}_{-S}$ are r_1 and r_2 respectively. Then let Q_1 and Q_2 be rotations so that the last $k-r_1$ and $p-k-r_2$ entries of $Q_1 X_S$ and $Q_2 X_{-S}$ are zero. Let $Y_1 \in \mathbb{R}^{r_1}$ and $Y_2 \in \mathbb{R}^{r_2}$ be the first r_1 and r_2 entries of X_S and X_{-S}. By the definition of canonical correlations, the sum of the squared canonical correlations between X_S and X_{-S} and Y_1 and Y_2 are the same, and it is easy to verify that $\mathrm{Tr}(\mathbf{\Sigma}_{Y_1}^{-1}\mathbf{\Sigma}_{Y_1,Y_2}\mathbf{\Sigma}_{Y_2}^{-1}\mathbf{\Sigma}_{Y_2,Y_1})$ is equal to $\mathrm{Tr}(\mathbf{\Sigma}_S^+ \mathbf{\Sigma}_{S,-S}\mathbf{\Sigma}_{-S}^+\mathbf{\Sigma}_{-S,S})$.

Now, we compute that

$$\mathrm{Tr}(\mathbf{\Sigma}_S^+ \mathbf{\Sigma}_{S,-S}\mathbf{\Sigma}_{-S}^+\mathbf{\Sigma}_{-S,S}) = \mathrm{Tr}(\mathbf{\Sigma}_S^+(\mathbf{\Sigma}_S - \mathbf{\Sigma}_{R(X_S,X_{-S})})) = \mathrm{rank}(\mathbf{\Sigma}_S) - \mathrm{Tr}(\mathbf{\Sigma}_S^+ \mathbf{\Sigma}_{R(X_S,X_{-S})})$$

Consider having currently selected a subset U. Take a fixed $i \notin U$ and let $V = U + i$. Take j, h to be as described in the presentation of the modified algorithm. $\mathbf{\Sigma}_V$ will have rank larger than $\mathbf{\Sigma}_U$ by one if feature i cannot be exactly reconstructed by the features in U and the same rank as $\mathbf{\Sigma}_U$ otherwise. So, $\mathrm{rank}(\mathbf{\Sigma}_V) = \mathrm{rank}(\mathbf{\Sigma}_U) + I_{(\mathbf{\Sigma}_{R(X_{-U},X_U)})_{hh}>0}$. Taking $\tilde{\beta} = \mathbf{\Sigma}_{\bullet i} - \mathbf{\Sigma}_{\bullet,-V}\mathbf{\Sigma}_{-V}^+\mathbf{\Sigma}_{-V,i}$, we know from Lemma 1 that

$$\mathbf{\Sigma}_{R(X,X_{-V})} = \mathbf{\Sigma}_{R(X,X_{-U)})} + \frac{\tilde{\beta}\tilde{\beta}^\top}{\tilde{\beta}_i} \cdot I_{\tilde{\beta}_i>0}$$

From this, defining $\beta = \tilde{\beta}_V$, and recalling that $\beta_j = \tilde{\beta}_i$, we have

$$\mathbf{\Sigma}_{R(X_V,X_{-V})} = \mathbf{\Sigma}_{R(X_V,X_{-U)}} + \frac{\beta\beta^\top}{\beta_j} \cdot I_{\beta_j>0}$$

Since $i \in -U$, the jth row and column of $\mathbf{\Sigma}_{R(X_V,X_{-U)})}$ are zeros. Putting everything together

$$\begin{aligned}&\mathrm{rank}(\mathbf{\Sigma}_V) - \mathrm{Tr}(\mathbf{\Sigma}_V^+ \mathbf{\Sigma}_{R(X_V,X_{-V})}) \\ &= \mathrm{rank}(\mathbf{\Sigma}_U) + I_{(\mathbf{\Sigma}_{R(X_{-U},X_U)})_{hh}>0} - \mathrm{Tr}((\mathbf{\Sigma}_V^+)_{-j,-j}\mathbf{\Sigma}_{R(X_U,X_{-U)})}) - \frac{\beta^\top \mathbf{\Sigma}_V^+ \beta}{\beta_j} \cdot I_{\beta_j>0}\end{aligned}$$

which is sufficient to imply that the $i \notin U$ we select minimizes the objective over subsets $U + i$.

Now we consider the case where $\mathbf{\Sigma} \succ \mathbf{0}$. First, $\mathbf{\Sigma}_S^+ = \mathbf{\Sigma}_S^{-1}$ so our objective simplifies to minimizing $\mathrm{Tr}(\mathbf{\Sigma}_S^{-1}\mathbf{\Sigma}_{R(X_S,X_{-S})})$. In this setting, [Khan, 2008] tells us that $\mathbf{\Sigma}_{-V}^{-1} = (\mathbf{\Sigma}_{-U}^{-1})_{-h,-h} - (\mathbf{\Sigma}_{-U})_{-h,h}(\mathbf{\Sigma}_{-U})_{h,-h}/(\mathbf{\Sigma}_{-U})_{hh}$. Consider β from before. Since $\mathbf{\Sigma} \succ \mathbf{0}$, $\beta_j > 0$, so $\mathbf{\Sigma}_{R(X_V,X_{-V})} = \mathbf{\Sigma}_{R(X_V,X_{-U})} + \beta\beta^\top/\beta_j$. Also from [Khan, 2008], letting $\alpha = \mathbf{\Sigma}_U^{-1}\mathbf{\Sigma}_{Ui}$, $\delta = \mathbf{\Sigma}_{ii} - \mathbf{\Sigma}_{iU}\alpha$, we know that $(\mathbf{\Sigma}_V^{-1})_{-j,-j} = \mathbf{\Sigma}_U^{-1} - \alpha\alpha^\top/\delta$, $(\mathbf{\Sigma}_V^{-1})_{-j,j} = -\alpha/\delta$ and $(\mathbf{\Sigma}_V^{-1})_{j,j} = 1/\delta$. Again, recalling that the jth row and column of $\mathbf{\Sigma}_{R(X_V,X_{-U})}$ are zeros because $i \in -U$ we can compute

$$\begin{aligned}
\mathrm{Tr}(\mathbf{\Sigma}_V^{-1}\mathbf{\Sigma}_{R(X_V,X_{-V})}) &= \mathrm{Tr}((\mathbf{\Sigma}_U^{-1} - \alpha\alpha^\top/\delta)(\mathbf{\Sigma}_{R(X_U,X_{-U})} + \beta_{-j}\beta_{-j}^\top/\beta_j) - \alpha\beta_{-j}^\top/\delta) - \alpha^\top\beta_{-j}/\delta + \beta_j/\delta) \\
&= \mathrm{Tr}(\mathbf{\Sigma}_U^{-1}\mathbf{\Sigma}_{R(X_U,X_{-U})}) + \frac{\beta_{-j}^T\mathbf{\Sigma}_U^{-1}\beta_{-j}}{\beta_j} - \frac{(\alpha^T\beta_{-j})^2}{\delta\beta_j} - \frac{\alpha^\top\mathbf{\Sigma}_{R(X_U,X_{-U})}\alpha + 2\alpha^\top\beta_{-j} - \beta_j}{\delta}
\end{aligned}$$

which is sufficient to imply that the $i \notin U$ we select minimizes the objective over subsets $U + i$.

A.7.18 Proof of Proposition A.5.1

Fixing a size-k subset U and some variable $i \notin U$ we show finite sample L^2 norm concentration for the coefficients $\hat{\mathbf{\Sigma}}_{iU}\hat{\mathbf{\Sigma}}_U^+$ from the regression of X_i on X_U. The rest of the result then follows trivially from a union bound. We will use the same notation as in Appendix A.4 and refer to results from there as well. The one difference is that the c_i represent positive constants that are independent of n, p, and also k.

First we re-configure

$$\begin{aligned}
\|\hat{\mathbf{\Sigma}}_U^+\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_U^+\mathbf{\Sigma}_{Ui}\|_2 &= \|\hat{\mathbf{\Sigma}}_U^+\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_U^+\hat{\mathbf{\Sigma}}_{Ui} + \mathbf{\Sigma}_U^+\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_U^+\mathbf{\Sigma}_{Ui}\|_2 \\
&\le \|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2\|\hat{\mathbf{\Sigma}}_{Ui}\|_2 + \|\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_{Ui}\|_2\|\mathbf{\Sigma}_U^+\|_2 \\
&\le \|\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_{Ui}\|_2\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 + \|\mathbf{\Sigma}_{Ui}\|_2\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 + \|\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_{Ui}\|_2\|\mathbf{\Sigma}_U^+\|_2 \\
&\le \|\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_{Ui}\|_2\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 + B\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 + c_1\|\hat{\mathbf{\Sigma}}_{Ui} - \mathbf{\Sigma}_{Ui}\|_2
\end{aligned}$$

where in the last line we have used Assumption 5 to bound $\|\mathbf{\Sigma}_U^+\|_2 = \|\mathbf{\Sigma}_U^{-1}\|_2$ and bounded $\|\mathbf{\Sigma}_{Ui}\|_2$ by $B = \max_{j\in -U}\|\mathbf{\Sigma}_{Uj}\|_2$. Fixing $t \in (0,1)$, it suffices to get tail bounds for two things:

Tail bound for $\|\hat{\mathbf{\Sigma}}_{iU} - \mathbf{\Sigma}_{iU}\|_2$: Because of Assumption 4 we can apply our sub-exponential concentration result (A.8) and get a tail bound for $\|\hat{\mathbf{\Sigma}}_{iU} - \mathbf{\Sigma}_{iU}\|_2$.

$$\begin{aligned}
P(\|\hat{\mathbf{\Sigma}}_{iU} - \mathbf{\Sigma}_{iU}\|_2 > t) &= P(\sum_{j\in U}(\hat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij})^2 > t^2) \\
&\le \sum_{j\in U} P(|\hat{\mathbf{\Sigma}}_{ij} - \mathbf{\Sigma}_{ij}| > t/k^{1/2}) && \text{(union bound)} \\
&\le k\exp(-c_2\min(t/k^{1/2}, t^2/k)n) && \text{(sub-exponential concentration)} \\
&\le k\exp\left(\frac{-c_2t^2n}{k}\right) && (t < 1)
\end{aligned}$$

Tail bound for $\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2$: To get concentration for $\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2$ we note that, on the event that

$\hat{\mathbf{\Sigma}}_U^+$ is invertible, we can use Assumption 2 to see that

$$\begin{aligned}
\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 &= \|\hat{\mathbf{\Sigma}}_U^{-1} - \mathbf{\Sigma}_U^{-1}\|_2 \\
&= \|\mathbf{\Sigma}_U^{-1/2}(\mathbf{\Sigma}_U^{1/2}\hat{\mathbf{\Sigma}}_U^{-1}\mathbf{\Sigma}_U^{1/2} - \boldsymbol{I}_k)\mathbf{\Sigma}_U^{-1/2}\| \\
&\leq \|\mathbf{\Sigma}_U^{-1}\|_2\|\mathbf{\Sigma}_U^{1/2}\hat{\mathbf{\Sigma}}_U^{-1}\mathbf{\Sigma}_U^{1/2} - \boldsymbol{I}_k\|_2 \\
&\leq c_3\|\mathbf{\Sigma}_U^{1/2}\hat{\mathbf{\Sigma}}_U^{-1}\mathbf{\Sigma}_U^{1/2} - \boldsymbol{I}_k\|_2 \\
&= c_3 \max_{j\in[k]} |s_j^{-2}(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1|
\end{aligned}$$

We recall the concentration result [Vershynin, 2018, Theorem 4.6.1]. Let M be the sub-Gaussian norm of X_U. Then for any $x \geq 0$

$$P(\max_{j\in[k]} |s_i(\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - n^{1/2}| > cM^2(k^{1/2} + x)) \leq 2\exp(-x^2),$$

which means more generally for any x that

$$P(\max_{j\in[k]} |s_i(\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - n^{1/2}| > x) \leq 2\exp(-(x/(cM^2) - k^{1/2})_+^2),$$

Fix some $c_4 > 0$. Then we can apply this result to see that

$$\begin{aligned}
&P(\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 > t) \\
&\leq P\left(\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 > t, \max_{i\in[k]} |s_j(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| < c_4\right) \\
&\quad + P\left(\max_{j\in[k]} |s_j(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| \geq c_4\right) \\
&\leq P\left(\max_{i\in[k]} |s_j^{-2}(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| > c_5 t, \max_{i\in[k]} |s_j(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| < c_4\right) \\
&\quad + P\left(\max_{j\in[k]} |s_j(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| \geq c_4\right) \\
&\qquad\qquad (\hat{\mathbf{\Sigma}}_U \text{ is invertible on the event}) \\
&\leq P\left(\max_{j\in[k]} |s_i(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| > c_6 t, \max_{j\in[k]} |s_i(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| < c_4\right) \\
&\quad + P\left(\max_{j\in[k]} |s_i(n^{-1/2}\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - 1| \geq c_4\right) \\
&\qquad\qquad (\text{Lemma A.4.1 with } a = c_4) \\
&\leq P\left(\max_{j\in[k]} |s_i(\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - n^{1/2}| > c_6 n^{1/2} t\right) + P\left(\max_{j\in[k]} |s_i(\boldsymbol{X}_{\bullet U}\mathbf{\Sigma}_U^{-1/2}) - n^{1/2}| \geq c_4 n^{1/2}\right) \\
&\leq 2\exp(-(c_6 n^{1/2} t/(cM^2) - k^{1/2})_+^2) + 2\exp(-(c_4 n^{1/2}/(cM^2) - k^{1/2})_+^2) \\
&\qquad\qquad (\text{above singular value concentration}) \\
&\leq 4\exp(-(c_7 n^{1/2} t/M^2 - k^{1/2})_+^2) \qquad\qquad (t < 1)
\end{aligned}$$

Putting everything together, we can use a union bound to see that

$$
\begin{aligned}
&P(\|\hat{\mathbf{\Sigma}}_{iU}\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_{iU}\mathbf{\Sigma}_U^+\|_2 > t) \\
&\leq P(\|\hat{\mathbf{\Sigma}}_{iU} - \mathbf{\Sigma}_{iU}\|_2 > \frac{t^{1/2}}{\sqrt{3}}) + P(\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 > \frac{t^{1/2}}{\sqrt{3}}) \\
&\quad + P(\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 > \frac{t}{3B}) + P(\|\hat{\mathbf{\Sigma}}_{iU} - \mathbf{\Sigma}_{iU}\|_2 > \frac{t}{3c_1}) \\
&\leq 2P(\|\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_U^+\|_2 > c_9 t/\max(B,1)) + 2P(\|\hat{\mathbf{\Sigma}}_{iU} - \mathbf{\Sigma}_{iU}\|_2 > c_{10}t) \\
&= 8\exp(-(c_{11}n^{1/2}t/(\max(BM^2, M^2)) - k^{1/2})_+^2) + 2k\exp\left(-(c_{12}n^{1/2}t/k^{1/2})^2\right) \\
&= 8\exp(-k(c_{11}n^{1/2}t/(k^{1/2}\max(BM^2, M^2)) - 1)_+^2) + 2k\exp\left(-(c_{12}n^{1/2}t/k^{1/2} - 1)_+^2\right) \\
&\leq c_{13}k\exp(-(c_{14}n^{1/2}t/(k^{1/2}\max(BM^2, M^2, 1)) - 1)_+^2) \\
&= c_{13}k\exp\left(-\left(\frac{c_{14}tn^{1/2}}{k^{1/2}\max(BM^2, M^2, 1)} - 1\right)_+^2\right)
\end{aligned}
$$

Since the final bound has no dependence on i, we have by a union bound over $i \notin U$ that

$$
P(\|\hat{\mathbf{\Sigma}}_{\bullet U}\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_{\bullet U}\mathbf{\Sigma}_U^+\|_{2\to\infty} > t) \leq c_{13}pk\exp\left(-\left(\frac{c_{14}tn^{1/2}}{k^{1/2}\max(BM^2, M^2, 1)} - 1\right)_+^2\right)
$$

Since one can easily show using Assumption 4 that $B \leq c_{14}\sqrt{k}$ and $M \leq c_{15}\sqrt{k}$, we also have the bound

$$
P(\|\hat{\mathbf{\Sigma}}_{\bullet U}\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_{\bullet U}\mathbf{\Sigma}_U^+\|_{2\to\infty} > t) \leq c_{13}pk\exp\left(-\left(\frac{c_{16}tn^{1/2}}{k^2} - 1\right)_+^2\right),
$$

where the right hand side has no dependence on the subset U. By union bounding over all subsets U, we get

$$
P(\max_{U\subseteq[p]:|U|=k}\|\hat{\mathbf{\Sigma}}_{\bullet U}\hat{\mathbf{\Sigma}}_U^+ - \mathbf{\Sigma}_{\bullet U}\mathbf{\Sigma}_U^+\|_{2\to\infty} > t) \leq c_{13}p^{k+1}k\exp\left(-\left(\frac{c_{16}tn^{1/2}}{k^2} - 1\right)_+^2\right).
$$

A.7.19 Proofs of finding best subset

We give proofs that our greedy (Algorithm 1) and swapping (Algorithm 2) algorithms find the best subset in certain settings. Our language suggests that there is one best subset, but all the arguments hold if multiple subsets are tied for having the lowest CSS objective value.

Diagonal covariance

If $\mathbf{\Sigma}$ is a diagonal matrix, then the CSS objective for a subset S is simply $\sum_{j\notin S}\mathbf{\Sigma}_{jj}$. It is clear this is minimized when S is the set of variables with the largest variances.

Greedy: Lemma 1 implies that our greedy algorithm will select the remaining variable with highest variance on each iteration.

Swapping: Lemma 1 implies that our swapping algorithm can always improve the objective by swapping out a variable in the selected set that does not have one of the top k variances for one that does.

Perfect reconstruction

If $\mathbf{\Sigma}$ is such that there exists a set of k variables that perfectly linearly reconstruct the remaining, then all the variables are in the span of these k variables. This means all the variables live in a q-dimensional linear subspace where $q \leq k$.

Greedy: On the i-th iteration of our greedy algorithm, if the dimension of the subspace spanned by the previous selections is $< q$, then there must be at least one remaining variable that is not yet perfectly linearly reconstructed by our selections so far (otherwise the variables would live in a linear subspace of dimension $< q$). By Lemma 1 our greedy algorithm will select one such variable on the ith iteration, and therefore the dimension of the subspace spanned by our selections must increase by one (because we've added something linearly independent to our currently selected subset). Thus, by the end of the kth iteration we must have selected a subset that spans a q-dimensional subspace. Our selected variables must necessarily act as a basis of the q-dimensional subspace that all the variables live in, and the greedy algorithm will thus achieve an objective value of zero.

Swapping: For our swapping algorithm, if there is a size-k subset S that spans a subspace of dimension $< q \leq k$, then there must be (1) a variable j_1 in that size-k subset that is perfectly linearly reconstructed by the remaining $k-1$ variables and (2) some variable j_2 not in the subset that we cannot perfectly linearly reconstruct with variables in the subset (otherwise all our variables would live in a $< q$ dimensional subspace). Lemma 1 tells us that swapping variable j_1 out of S and j_2 into S improves the CSS objective. It also increases the dimension of the subspace spanned by the variables in S by one. Thus, our swapping algorithm will swap variables out until the dimension of the subspace spanned by variables in the selected subset is q, and by the same reasoning as for our greedy algorithm, the swapping algorithm will eventually achieve a CSS objective of zero.

Block diagonal

If $\mathbf{\Sigma}$ is a block diagonal correlation matrix with blocks $M_i \in \mathbb{R}^{q \times q}$, and each block has a row/column with square L^2 norm $> q/2$, then the best subset is one that consists of the variable from each block that has the largest row/column norm in the block. To see why, consider a subset S that does not have a representative from each block. In particular, it must be missing a representative from some block M_{i_1} and have at least two representatives from some block M_{i_2}. If we remove all the representatives from block M_{i_2} then Lemma 1 tells us that the CSS objective can increase by at most q (because only q unit variance variables are in the block, and those variables are independent of all the variables in the other blocks). If we then add back in the two variables with largest row/column norm from block M_{i_1} and M_{i_2} then Lemma 1 tells us that the CSS objective will decrease by strictly more than q. This new subset has less variables than our original subset and a lower CSS objective. Thus, the best subset must have exactly one representative from each block. Lemma 1 implies it should be the variable with the largest row/column norm from each block.

Greedy: We can prove the greedy algorithm finds the optimal subset by induction. On the first iteration, it is trivially true that the greedy algorithm will select a variable from a block it hasn't selected from yet. Lemma 1 tells us that, for whatever block it selects from, it will select the variable with largest row/column norm. Suppose now that up to the $(i-1)$st iteration, our greedy algorithm has never selected two variables from the same block and has always has selected the variable in each block with the largest row/column norm. Then on the ith iteration, if the greedy algorithm selects a variable from a new block with largest row/column norm, Lemma 1 tells us that it will reduce the CSS objective by strictly more than $q/2$. If it selects a variable from a block it has previously selected from however, it can reduce the CSS objective by only strictly less than $q/2$, because a variable can only further reduce the reconstruction error for variables in its own block, and Lemma 1 tells us that the cumulative reconstruction error for a block that already has a selected representative is strictly less than $q/2$. Thus on the ith iteration our greedy algorithm will select a variable from a new block, and Lemma 1 guarantees that it will select a variable that has largest row/column norm in said block. This is sufficient to imply that the greedy algorithm will find the best subset.

Appendix B

Supplementary material relating to selective inference

B.1 Permutation test p-value

The randomized permutation test from Proposition 3 of Hemerik and Goeman [2018] uses the p-value

$$p = \frac{\#\{1 \leq j \leq w : T(g_j(X)) > T(X)\}}{w} + U_{aux}\frac{\#\{1 \leq j \leq w : T(g_j(X)) = T(X)\}}{w},$$

where $U_{aux} \sim \text{Unif}([0,1])$ adds auxiliary randomness that is independent of X.

B.2 F-test p-value

For the matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, let $\boldsymbol{X}_q \in \mathbb{R}^{n\times q}$ denote the matrix that just contains the first q columns of $\boldsymbol{X}$. Then define

$$\text{RSS}_f = \|Y - \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top Y\|_2^2$$

$$\text{RSS}_r = \|Y - \boldsymbol{X}_q(\boldsymbol{X}_q^\top\boldsymbol{X}_q)^{-1}\boldsymbol{X}_q^\top Y\|_2^2$$

Under the null $H_0 : \beta_{q+1} = \cdots = \beta_d = 0$, the F statistic

$$F = \frac{(\text{RSS}_r - \text{RSS}_f)/(p-q)}{\text{RSS}_f/(n-p-1)}$$

has an exact $F_{d-q,n-q-1}$ distribution conditional on $\boldsymbol{X}$. Let G denote the CDF of the $F_{d-q,n-q-1}$ distribution. Under H_0, the p-value $p = 1 - G(p)$ for this test has an exact $\text{Unif}([0,1])$ distribution conditional on $\boldsymbol{X}$.

B.3 Selective dominance and independence

Consider a p-value p for the null H_0 which is selectively dominant. Suppose that under P_{H_0} the p-value p is always independent of some random variable Z. We show that p is then selectively dominant given Z.

Fix a distribution P in the null H_0 and let $f(x)$ be PDF of p. Considering a selection function $s(x)$ such that $\int_0^1 s(x)f(x) > 0$, we have that

$$\frac{\int_0^t s(x)f(x)dx}{\int_0^1 s(x)f(x)dx} \leq \frac{\int_0^t s(x)dx}{\int_0^1 s(x)dx} \text{ for all } t \in [0,1]$$

from the selective dominance (4.3) of p.

Now, consider a new selection function $s(x,z)$. Note that due to independence, $f(x)$ is still the conditional PDF of p given $Z = z$. For any z such that $\int_0^1 s(x,z)f(x) > 0$, it follows from above that

$$\frac{\int_0^t s(x,z)f(x)dx}{\int_0^1 s(x,z)f(x)dx} \leq \frac{\int_0^t s(x,z)dx}{\int_0^1 s(x,z)dx} \text{ for all } t \in [0,1]$$

If under $s(x,z)$, we have $P(S=1) = 0$, then the selective dominance condition (4.3) trivially holds. On the other hand, if $P(S=1) > 0$, then we have shown in Appendix B.19.1 that $\int_0^1 s(x,Z)f(x) > 0$ a.e. under $P(\cdot|S=1)$. Thus the above guarantees that a.e. under $P(\cdot|S=1)$,

$$\frac{\int_0^t s(x,Z)f(x)dx}{\int_0^1 s(x,Z)f(x)dx} \leq \frac{\int_0^t s(x,Z)dx}{\int_0^1 s(x,Z)dx} \text{ for all } t \in [0,1].$$

By arguments from Appendix B.19.1, the left-hand side is equal to the left-hand side of (4.3) a.e. under $P(\cdot|S=1)$, and the right-hand side equals the right-hand side of (4.3) a.e. under $P(\cdot|S=1)$, so we have established the claim.

B.4 Data carving for the file-drawer problem

We have two data samples $X_1 \sim N(\mu, 2)$ and $X_2 \sim N(\mu, 2)$ that are independent and want to test $H_0 : \mu \leq 0$. Suppose we only do inference because we observed that $X_1 > t$ for some threshold t. If we consider the p-value $p_{full} = 1 - \Phi((X_1 + X_2)/2)$, then our selection function is given by

$$\begin{aligned} s(x) &= P(X_1 > t | p_{full} = x) \\ &= P(X_1 > t | \frac{X_1 + X_2}{2} = \Phi^{-1}(1-x)) \\ &= 1 - \Phi(t - \Phi^{-1}(1-x)) \end{aligned}$$

where we have used that

$$\begin{bmatrix} X_1 \\ \frac{X_1+X_2}{2} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}, \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \right)$$

so

$$X_1 | \frac{X_1 + X_2}{2} = y \sim N(y, 1)$$

Thus our corrected p-value is given by

$$p_{carve} = \frac{\int_0^{p_{full}} 1 - \Phi(t - \Phi^{-1}(1-x))dx}{\int_0^1 1 - \Phi(t - \Phi^{-1}(1-x))dx}$$

$$p_{carve} = \frac{\int_{\bar{X}}^{\infty} \phi(z)(1 - \Phi(t-z))dz}{\int_{-\infty}^{\infty} \phi(z)(1 - \Phi(t-z))dz} = \frac{\int_{\bar{X}}^{\infty} \phi(z)(1 - \Phi(t-z))dz}{1 - \Phi(t/\sqrt{2})}$$

We now show that p_{carve} is monotone non-decreasing in t. Letting Z and Y be independent standard normal random variables and fixing some constant a, the selective p-value is given by

$$p_{carve} = \frac{P(Z + Y > t, Z > a)}{P(Z + Y > t)} = P(Z > a | Z + Y > t)$$

for $a = \bar{X}$. Letting $W = Z + Y$ we can write $Z = \frac{1}{2}W + \epsilon$ where ϵ is independent of W. This gives us

$$p_{carve} = P(\frac{1}{2}W + \epsilon > a | W > t) = E[P(W > 2(a-\epsilon)|W > t, \epsilon)|W > t] = E[P(W > 2(a-\epsilon)|W > t, \epsilon)]$$

Then the fact that p_{carve} is monotone non-decreasing in t follows from the fact that $P(W > c|W > t)$ is monotone non-decreasing in t for every constant c:

$$P(W > c|W > t) = \begin{cases} \frac{P(W>c)}{P(W>t)} & \text{if } t \leq c, \\ 1 & \text{if } t > c. \end{cases}$$

B.5 More info on data splitting, carving, fission, and thinning

Crucially, in Example 10, the conditional distribution of the random variable X_1 used for selection given the p-value p_{full} did not depend on the unknown parameter μ. Hence, the selection function $s(x)$ had no dependence on μ, and we were able to construct our selective p-value without any issues. This did not happen by accident, and is actually a consequence of more general and interesting fact regarding the relationship between between data splitting, data carving, data fission [Leiner et al., 2023], and data thinning [Dharamshi et al., 2024, Neufeld et al., 2024].

In the most basic version of data fission, we add and subtract independent normal noise $Z \sim N(0,1)$ to a normal sample $X \sim N(\mu, 1)$ to get two independent samples $X_1, X_2 \sim N(\mu, 2)$ centered at the same mean. This is meant to mimic data splitting: the first sample can be used for selection and the second for inference. Data thinning generalizes this idea by considering a random vector X from a parametric family and adding noise to make k new random vectors $X_1, \ldots, X_k$ that (1) are independent and (2) can be used to recover X via a deterministic function $X = T(X_1, \ldots, X_k)$. Vanilla data thinning involves using some of the X_i to perform selection and then the rest to do inference. Data carving, however, suggests using a p-value $p = p(T(X_1, \ldots, X_n))$ that leverages all of the data, despite some of the X_i having been used for selection. Because the noise we add to X to get the X_i has no dependence on the unknown parameter, the joint distribution of the X_i given p also has no dependence on the unknown parameter. Therefore, contrary to the what Leiner et al. [2023] suggest, the selection function $s(x)$ is always known, and we can always apply our framework to data carve and get more power. If the selection process is highly complicated, it is true that $s(x)$ may be very difficult to compute, but in theory it is always accessible to us via extensive

simulations.

B.6 Post selection inference for the LASSO

In the below example, we freely refer to results from Lee et al. [2016b].

Example 11 (Post selection inference for the LASSO)**.** *Suppose that p is a valid p-value for testing the null H_0 and it is selectively dominant given Z. Imagine there are known functions $q^+(z) \leq q^-(z)$ and $v(z)$, and we only choose to test H_0 with p when $x(Z) > 0$ and $p \in [q^+(Z), q^-(Z)]$. Applying our framework with the selection function $s(x) = I(p \in [q^+(Z), q^-(Z)])I(v(Z) > 0)$, Theorem 5 tells us that we control selective Type I error if we reject according to the selective p-value $p_{sel} = \frac{p-q^+(Z)}{q^-(Z)-q^+(Z)}$:*

$$P_{H_0}\left(\frac{p - q^+(Z)}{q^-(Z) - q^+(Z)} \leq \alpha | S = 1\right) \leq \alpha. \tag{B.1}$$

Now, suppose we observe n-dimensional data $Y \in N(\mu, \sigma^2)$ and a fixed matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. For a fixed regularization strength λ, the LASSO solution is composed of a fitted coefficient vector $\hat{\beta} \in \mathbb{R}^p$ and a sign vector $\hat{s} \in \mathbb{R}^p$ that satisfy the following Karush-Kuhn-Tucker conditions:

$$\begin{aligned} &X^\top(X\hat{\beta} - Y) + \lambda \hat{s} = 0, && \\ &\hat{s}_i = sign(\hat{\beta}_j) && \text{if } \hat{\beta}_j \neq 0, \\ &\hat{s}_i \in [-1, 1] && \text{if } \hat{\beta}_j = 0. \end{aligned}$$

We define the fitted model from the LASSO to be the set

$$\hat{M} = \{i \in \{1, \ldots, p\} : |\hat{s}_i| = 1\}$$

For almost every λ, this is equal to the set of $i \in [p]$ for which $\hat{\beta}_i$ is non-zero.

Now we establish our inferential target. For a subset $M \subseteq [p]$, define $\boldsymbol{X}_M \in \mathbb{R}^{n \times |M|}$ to be the matrix that has the columns of $\boldsymbol{X}$ indexed by M. Our goal is to do inference on the population parameters for the model M,

$$\beta^M = \underset{b \in R^{|M|}}{\operatorname{argmin}} E[\|Y - \boldsymbol{X}_M b\|_2^2] = \boldsymbol{X}_M^\dagger \mu.$$

Particularly, we focus on doing inference on the jth index of this parameter vector $\beta_j^M = e_j^\top \boldsymbol{X}_M^\dagger \mu$ by testing the null $H_{0,j}^M : \beta_j^M \leq 0$. Letting $\Omega^M = \boldsymbol{X}_M^\dagger (\boldsymbol{X}_M^\dagger)^\top$, we see that $e_j^\top \boldsymbol{X}_M^\dagger y \sim N(\beta_j^M, \sigma^2 \Omega_{jj}^M)$. Thus we can test this null using the p-value

$$p_j^M = 1 - \Phi\left(\frac{e_j^\top \boldsymbol{X}_M^\dagger y}{\sigma\sqrt{\Omega_{jj}^M}}\right),$$

which is selectively dominant by Example 3. However, we only use this p-value to test the null when we observe that $\hat{M} = M$.

Define $Z = (I_n - (\boldsymbol{X}_M^\dagger)^\top e_j e_j^\top \boldsymbol{X}_M^\dagger / \Omega_{jj}^M) Y$ so that Y and Z are independent. Lemma 5.1 of Lee et al.

[2016b], often dubbed the polyhedral lemma, precisely tells us that there are functions

$$q_j^{M,s,+}(z) = 1 - \Phi\left(\frac{\mathcal{V}^{M,s,+}(z)}{\sigma\sqrt{\Omega_{jj}^M}}\right)$$
$$q_j^{M,s,-}(z) = 1 - \Phi\left(\frac{\mathcal{V}^{M,s,-}(z)}{\sigma\sqrt{\Omega_{jj}^M}}\right)$$
$$\mathcal{V}^{M,s,0}(z)$$

of z, such that

$$\{\hat{M} = M, \hat{s}_M = s\} = \{p_j^M \in [q_j^{M,s,+}(Z), q_j^{M,s,-}(Z)], \mathcal{V}_{M,s}^0(Z) \geq 0\}$$

The functions $\mathcal{V}^{M,s,-}(z)$, $\mathcal{V}^{M,s,+}(z)$, $\mathcal{V}^{M,s,0}(z)$ match those given in Lee et al. [2016b].

Now it is easy to see that rejecting the data-dependent null $H_{0,j}^{\hat{M}}$ when $(p_j^{\hat{M}} - q_j^{\hat{M},\hat{s},+}(Z))/(q_j^{\hat{M},\hat{s},-}(Z) - q_j^{\hat{M},\hat{s},+}(Z))$ is at most α controls Type I error conditionally on $\hat{M}$, and therefore also marginally. If $H_{0,j}^M$ is false, then trivially $P(\text{falsely reject } H_{0,j}^{\hat{M}} | \hat{M} = M, \hat{s}_M = s) = 0 \leq \alpha$. For the case that $H_{0,j}^M$ is true, the selection event $\hat{M} = M$ and $\hat{s}_M = s$ is the same as selecting p_j^M for inference in (B.1), so

$$\begin{aligned}
P(\text{falsely reject } H_{0,j}^{\hat{M}} | \hat{M} = M, \hat{s}_M = s) &= P\left(\frac{p_j^{\hat{M}} - q_j^{\hat{M},\hat{s},+}(Z)}{q_j^{\hat{M},\hat{s},-}(Z) - q_j^{\hat{M},\hat{s},+}(Z)} \leq \alpha \,\middle|\, \hat{M} = M, \hat{s}_M = s\right) \\
&= P\left(\frac{p_j^M - q_j^{M,s,+}(Z)}{q_j^{M,s,-}(Z) - q_j^{M,s,+}(Z)} \leq \alpha \,\middle|\, \hat{M} = M, \hat{s}_M = s\right) \\
&\leq \alpha
\end{aligned}$$

Conditional error control on $\hat{M}$ then follows from the law of total probability,

$$\begin{aligned}
P(\text{falsely reject } H_{0,j}^{\hat{M}} | \hat{M} = M) &= \sum_{s \in \{-1,1\}^{|M|}} P(\hat{s}_M = s | \hat{M} = M) P(\text{falsely reject } H_{0,j}^{\hat{M}} | \hat{M} = M, \hat{s}_M = s) \\
&\leq \alpha \sum_{s \in \{-1,1\}^{|M|}} P(\hat{s}_M = s | \hat{M} = M) \\
&= \alpha,
\end{aligned}$$

as does marginal error control,

$$\begin{aligned}
P(\text{falsely reject } H_{0,j}^{\hat{M}}) &= \sum_{M \subseteq [p]} P(\hat{M} = M) P(\text{falsely reject } H_{0,j}^{\hat{M}} | \hat{M} = M) \\
&\leq \alpha \sum_{M \subseteq [p]} P(\hat{M} = M) \\
&= \alpha.
\end{aligned}$$

B.7 Comparing conditional and simultaneous inference

We consider a setting where we have n independent and selectively dominant p-values $p_1, \ldots, p_n$ that are all anti-conservative, i.e., $p_j \preceq \text{Unif}([0,1])$. At worst, these p-values are exact uniforms (e.g., they come from the boundary of the null).

We will show that, on an event with probability at least $1-\epsilon$, the conditional procedure, which rejects when $p_{(1)} \leq \alpha p_{(2)}$, can only reject if $p_{(1)} \leq C_\epsilon/n$ for some constant $C_\epsilon > 0$. Hence, without conservative nulls, the conditional approach behaves roughly on the same order as the classical approach (Sidak).

Letting $U_1, \ldots, U_n$ be independent Unif($[0,1]$) random variables, two facts are clear. First that $U_{(2)} \sim \text{Beta}(2, n-1)$ has mean $\frac{2}{n+1} < \frac{2}{n}$ and standard deviation $\sqrt{\frac{2(n-1)}{(n+1)^2(n+2)}} \leq \frac{2}{n}$, and second that $p_{(2)} \preceq U_{(2)}$.

Fix any $\epsilon > 0$. We have by Chebyshev's inequality that

$$\begin{aligned} P\left(p_{(2)} \leq \frac{2}{n}(1+\epsilon^{-\frac{1}{2}})\right) &\geq P(p_{(2)} \leq E[U_{(2)}] + \sqrt{\text{Var}(U_{(2)})}/\sqrt{\epsilon}) \\ &\geq P(U_{(2)} \leq E[U_{(2)}] + \sqrt{\text{Var}(U_{(2)})}/\sqrt{\epsilon}) \\ &\geq P(|U_{(2)} - E[U_{(2)}]| < \sqrt{\text{Var}(U_{(2)})}/\sqrt{\epsilon}) \\ &\geq 1-\epsilon \end{aligned}$$

Define $C_\epsilon = 2(1+\epsilon^{-\frac{1}{2}})\alpha$ and the event $A_\epsilon = \{p_{(2)} \leq \frac{2}{n}(1+\epsilon^{-\frac{1}{2}})\}$. The event A_ϵ has probability at least $1-\epsilon$. Furthermore, on this event, the conditional procedure never rejects when $p_{(1)} > C_\epsilon/n$, completing our claim.

In other words, the conditional procedure can only reject when $p_{(1)} \leq C_\epsilon/n$ (aside for on a small probability event), and hence suffers the same curse of dimensionality as the classical method:

$$P(p_{(1)} \leq \alpha p_{(2)} \text{ and } p_{(1)} > C_\epsilon/n) \leq P(A_\epsilon^c) \leq \epsilon \text{ for some constant } C_\epsilon \text{ independent of } n.$$

B.8 Conditional inference on winners

We first provide the standard derivation for the conditional LCB and CI for the winning mean in the setting of independent Gaussian data. Then we show that it matches our p-value viewpoint. We then argue that the selective p-value we use for this winner problem is monotone in the null parameter, which allows us to prove the some facts about the more general conditional LCB and CI that applies for all MLR families. Finally, we use our p-value viewpoint to argue that, in the independent Gaussian case, the standardized distance between the winner X_W and conditional LCB is purely a function of the standardized distance between the winner X_W and runner-up X_R.

B.8.1 Standard derivation

Suppose we observe independent Gaussian data $X \sim N(\mu, I_n)$ and want to make a LCB for the mean μ_W of the winner $W = \text{argmax}_{i \in [n]} X_i$.

To construct the conditional LCB, we follow the framework of Fithian et al. [2017]. Letting R be the index of the runner-up (second largest observation), we note that the deviation of X_W from μ_W has a truncated

normal distribution once we condition on W and the nuisance statistics X_{-W}:

$$X_W - \mu_W \mid W, X_{-W} \sim TN(0, 1, X_R - \mu_W, \infty). \tag{B.2}$$

Let

$$q_{1-\alpha}(x_r, \mu_w) = \text{Quantile}_\mu(1-\alpha, X_W - \mu_W \mid W = w, X_{-w} = x_{-w}) \tag{B.3}$$

denote the $1-\alpha$ quantile of this conditional distribution (B.2), which is a function of largest entry x_r of x_{-w} and the mean μ_w at the winning index. It is straightforward to show that

$$LCB_{cond}(X) = \{\mu_0 : \mu_0 > X_W - q_{1-\alpha}(X_R, \mu_0)\} \tag{B.4}$$

is a $1-\alpha$ confidence region for μ_W that has exact coverage conditional on W and X_{-W}:

$$P_\mu(\mu_W \in LCB_{cond}(X)|W, X_{-W}) = P_\mu(X_W - \mu_W < q_{1-\alpha}(X_R, \mu_W)|W, X_{-W}) = 1-\alpha.$$

Later, via our p-value viewpoint, we will argue that this confidence region is indeed an LCB.

Fixing some $0 < \alpha_1, \alpha_2 < \alpha$ such that $\alpha_1 + \alpha_2 = \alpha$, we also see that

$$CI_{cond}(X) = \{\mu_0 : X_W - q_{\alpha_2}(X_R, \mu_0) > \mu_0 > X_W - q_{1-\alpha_1}(X_R, \mu_0)\} \tag{B.5}$$

is a $1-\alpha$ confidence region that has exact coverage conditional on W and X_{-W}. Again, later via our p-value viewpoint we will argue that this region is a CI.

B.8.2 The p-value viewpoint

For the same setting as above, we want to use the p-values $p_i^{\mu_0} = 1 - \Phi(X_i - \mu_0)$ to characterize when the conditional LCB does not include $\mu_0 \in \mathbb{R}$. This happens exactly when μ_0 is not included in the set (B.4). Examining (B.2), (B.3), and (B.4), this happens when $X_W - \mu_0$ is at least as large as the $1-\alpha$ quantile Q of a standard normal truncated to be larger than $X_R - \mu_0$. This quantile satisfies

$$\alpha = \frac{1 - \Phi(Q)}{1 - \Phi(X_R - \mu_0)}.$$

Solving for Q gives $Q = \Phi^{-1}(1 - \alpha(1 - \Phi(X_R - \mu_0)))$, meaning we reject exactly when

$$\begin{aligned} X_W - \mu_0 \geq \Phi^{-1}(1 - \alpha(1 - \Phi(X_R - \mu_0))) &\iff 1 - \Phi(X_W - \mu_0) \leq \alpha(1 - \Phi(X_R - \mu_0)) \\ &\iff p_{(1)}^{\mu_0} \leq \alpha p_{(2)}^{\mu_0}. \end{aligned}$$

Now we do the same for the conditional CI (B.5). We want to characterize when μ_0 is not included in the set (B.5). Examining (B.2), (B.3), and (B.5), this happens either when $X_W - \mu_0$ is at least as large as the $1-\alpha_1$ quantile or at most as small as the α_2 quantile of the same truncated normal distribution. That is, either

$$X_W - \mu_0 \geq \Phi^{-1}(1 - \alpha_1(1 - \Phi(X_R - \mu_0))) \iff p_{(1)}^{\mu_0} \leq \alpha_1 p_{(2)}^{\mu_0},$$

or

$$X_W - \mu_0 \leq \Phi^{-1}(1 - (1-\alpha_1)(1 - \Phi(X_R - \mu_0))) \iff p_{(1)}^{\mu_0} \geq (1-\alpha_2) p_{(2)}^{\mu_0},$$

This will match the conditional CI we give later, which we will write in terms of p-values.

B.8.3 Monotonicity of the selective p-value

Recall the setting where $X_i \sim P_{\theta_i}$ are independent samples from some parametric family P_θ with MLR in $T(x)$, and let $p_i^{\theta_0}$ be the UMP p-values for testing $H_0 : \theta \leq \theta_0$.

We want to show that the selective p-value $p_j^{\theta_0} / \min_{i \neq j} p_j^{\theta_0}$ is monotone non-decreasing in θ_0. This will imply that $p_{(1)}^{\theta_0} / p_{(2)}^{\theta_0}$ is monotone non-decreasing in θ_0. So long as we can write our selection function in terms of the data with no dependence on θ_0, Appendix B.15.3 guarantees that this will be the case. Recall that each p-value $p_i^{\theta_0}$ is a function of $T(X_i)$ and auxiliary uniform random variables $U_{i,aux}$ that are independent of the data and each other. Imagining using our framework with $p = p_j^{\theta_0}$ and $Z = (T(X_{-j}), U_{-j,aux})$, we can write our selection function in terms of the data $T(X_i)$ and $U_{i,aux}$ as

$$\tilde{s}(t_j, u_j, t_{-j}, u_{-j}) = \begin{cases} 1, & \text{if } t_j > \max_{i \neq j} t_i, \\ 1, & \text{if } t_j = \max_{i \neq j} t_i \text{ and } u_j \leq \max_{i \neq j : t_i = t_j} u_i, \\ 0, & \text{otherwise.} \end{cases}$$

where t_{-j} and u_{-j} jointly represent z.

Because this selection function does not depend on θ_0, Appendix B.15.3 guarantees that the selective p-value resulting from it will be monotone. Note that, to apply Appendix B.15.3, we also needed to ensure that Z did not depend on θ_0, which is was not the case in Example 8's original treatment, but is the case in our current treatment.

B.8.4 The conditional LCB and CI for MLR families

Again suppose that $X_i \sim P_{\theta_i}$ are independent samples from some MLR family P_θ, and let $p_i^{\theta_0}$ be the UMP p-values for testing $H_0 : \theta \leq \theta_0$. Let $W = \operatorname{argmin}_{i \in [n]} p_i^{\theta_0}$ be the index of the smallest p-value, and define the parameter vector $\Theta = (\theta_1, \ldots, \theta_n)$.

First we claim that

$$\{\theta_0 : p_{(1)}^{\theta_0} / p_{(2)}^{\theta_0} > \alpha\}$$

is a LCB for the winning parameter θ_W that has exact $1-\alpha$ coverage conditional on W. The region being an LCB follows from what we showed earlier: $p_{(1)}^{\theta_0} / p_{(2)}^{\theta_0}$ is monotone non-decreasing in θ_0. We get exact $1-\alpha$ coverage because $p_j^{\theta_j}$ has an exact uniform distribution given $p_{-j}^{\theta_j}$ under P_Θ (see Lemma B.15.3). Thus conditional on $W = j$ and $p_{-j}^{\theta_j}$ (i.e., conditional on selection and Z), Theorem 5 tells us that the selective p-value $p_j^{\theta_j} / \min_{i \neq j} p_i^{\theta_j}$ has an exact uniform distribution conditional on selection:

$$P_\Theta\left(\theta_W \in \left\{\theta_0 : \frac{p_{(1)}^{\theta_0}}{p_{(2)}^{\theta_0}} > \alpha\right\} | W = j\right) = P_\Theta\left(\theta_j \in \left\{\theta_0 : \frac{p_j^{\theta_0}}{\min_{i\neq j} p_i^{\theta_0}} > \alpha\right\} | W = j\right)$$
$$= P_\Theta\left(\frac{p_j^{\theta_j}}{\min_{i\neq j} p_i^{\theta_j}} > \alpha | W = j\right)$$
$$= 1 - \alpha,$$

Likewise, we see that

$$\{\theta_0 : \alpha_1 < p_{(1)}^{\theta_0}/p_{(2)}^{\theta_0} < 1 - \alpha_2\}$$

is a CI for θ_W that has exact $1-\alpha$ coverage conditional on W. The fact that it is a CI again follows from the monotonicity of $p_{(1)}^{\theta_0}/p_{(2)}^{\theta_0}$, and exact coverage follows from an identical argument to the one above.

B.8.5 Distance between winner and conditional LCB

Consider observing Gaussian data $X \sim N(\mu, \sigma^2 I_n)$ and let W and R be the indices of the winner and runner up respectively. We will use our p-value viewpoint to show that that the standardized distance $D = (X_W - \hat{\mu})/\sigma$ between the winner X_W and the conditional LCB $\hat{\mu}$ for μ_W depends only on the standardized gap $(X_W - X_R)/\sigma$ between the winner and runner-up. Per our earlier discussions, the conditional LCB $\hat{\mu}$ satisfies

$$\frac{p^{\hat{\mu}}(X_W)}{p^{\hat{\mu}}(X_R)} = \alpha \iff \frac{1 - \Phi((X_W - \hat{\mu})/\sigma)}{1 - \Phi((X_R - \hat{\mu})/\sigma)} = \alpha \iff \frac{1 - \Phi(D)}{1 - \Phi(D - (X_W - X_R)/\sigma)} = \alpha.$$

Clearly then, D is a function of $(X_W - X_R)/\sigma$ and does not at all depend on the problem dimension n.

B.9 Conditional inference on winners for exponentials

Recall the exponential distribution $X \sim \text{Exp}(\lambda_i)$ which has PDF

$$f_\lambda(x) = \begin{cases} \lambda e^{-\lambda x} & x > 0, \\ 0 & x \leq 0 \end{cases}.$$

Defining $T(x) = 1/x$, we see for $x > 0$ and $\lambda_2 \geq \lambda_1$, the ratio

$$f_{\lambda_2}(x)/f_{\lambda_1}(x) = \frac{\lambda_2}{\lambda_1} \exp\left(-\frac{\lambda_2 - \lambda_1}{T(x)}\right)$$

is monotone non-decreasing in $T(x)$. Thus this family has an MLR in $T(x)$, and the UMP test for $H_0 : \lambda \leq \lambda_0$ thus rejects when $T(X)$ is large, or correspondingly when X is small. In particular, noting that the CDF of X is given by

$$F_X(x) = \begin{cases} 1 - e^{-\lambda x} & x > 0, \\ 0 & x \leq 0 \end{cases}$$

it rejects according to the p-value $p^{\lambda^0} = 1 - e^{-\lambda_0 X}$ (see Appendix B.15.1 for details about UMP testing in MLR families).

Now consider observing some independent data $X_i \sim \text{Exp}(\lambda_i)$. We know from Appendix B.8 that the selective p-value $p^{\lambda_0}_{(1)}/p^{\lambda_0}_{(2)}$ is monotone non-decreasing in λ_0. Using L'Hopital's rule, we can compute that

$$\begin{aligned}\lim_{\lambda_0 \downarrow 0} \frac{p^{\lambda_0}_{(1)}}{p^{\lambda_0}_{(2)}} &= \lim_{\lambda_0 \downarrow 0} \frac{1 - e^{-\lambda_0 X_{(1)}}}{1 - e^{-\lambda_0 X_{(2)}}} \\ &= \lim_{\lambda_0 \downarrow 0} \frac{X_{(1)} e^{-\lambda_0 X_{(1)}}}{X_{(2)} e^{-\lambda_0 X_{(2)}}} \\ &= \frac{X_{(1)}}{X_{(2)}},\end{aligned}$$

which is sufficient to imply our claim in the main text.

B.10 Hybrid inference on winners

We first provide the standard derivation for the hybrid LCB and CI for the winning mean in the setting of independent Gaussian data. Then we show that it matches our p-value viewpoint. We then provide more general hybrid confidence regions that apply for all MLR families, and show that they are indeed an LCB and CI. Finally, we use our p-value viewpoint to argue that, in the independent Gaussian case, the standardized distance between the winner X_W and conditional LCB is purely a function of the standardized distance between the winner X_W and runner-up X_R as well as the problem dimension n.

B.10.1 Standard derivation

We want to make a hybrid LCB for the mean μ_W of the winner $W = \text{argmax}_{i \in [n]} X_i$ in the case of independent Gaussian data $X \sim N(\mu, I_n)$ with unknown mean $\mu \in \mathbb{R}^n$.

The core idea behind hybrid inference is giving a confidence region $C_{hyb}(X)$ that has a very high probability of containing μ_W on a "good" event G_μ. Oddly, this good event depends on the unknown parameter. For some $\beta < \alpha$, we need G_μ to happen with probability at least $1 - \beta$. Then, if we ensure that $C_{hyb}(X)$ has at least $(1 - \alpha)/(1 - \beta)$ coverage on the G_μ, it will achieve $1 - \alpha$ coverage overall:

$$\begin{aligned}P_\mu(\mu_W \in C_{hyb}(X)) &= P_\mu(G_\mu) P_\mu(\mu_W \in C_{hyb}(X) | G_\mu) + P_\mu(G_\mu^c) P_\mu(\mu_W \in C_{hyb}(X) | G_\mu^c) \\ &\geq (1 - \beta) \left(\frac{1 - \alpha}{1 - \beta} \right) \\ &= 1 - \alpha.\end{aligned}$$

Considering some $\beta < \alpha$ and defining $\beta_n = 1 - (1 - \beta)^{1/n}$, our good event is that the confidence lower bounds $X_i - z_{1-\beta_n}$ for the means μ_i all simultaneously hold:

$$G_\mu = \{X_i < \mu_i + z_{1-\beta_n} \text{ for all } i \in [n]\}.$$

It is not hard to show that this good event happens with probability exactly $1 - \beta$.

We note that our treatment differs slightly from Andrews et al. [2023]. Andrews et al. [2023] focus on making confidence intervals, and rheir good event is that there are CIs (not LCBs) that simultaneously cover all the μ_i. The purpose of restricting to this good event is that on this good event, we know all the X_i are not too far from their μ_i, and therefore our inferences on this event should not result in exploding intervals (even if they are of conditional flavor). Because the explosion only happens for the lower endpoint, however, it is most important to ensure that the X_i are not too far above their means. Hence we put all our β error budget on the good event into making simultaneous LCBs, rather than two-sided CIs like Andrews et al. [2023].

Now, we can make a confidence region that contains the mean with probability at least $(1-\alpha)/(1-\beta)$ on this good event. If we condition on G_μ along with W and X_{-W}, the deviation of X_W from μ_W has a truncated normal distribution like (B.2) that is further truncated from above:

$$X_W - \mu_W \mid W, X_{-W}, G_\mu \sim TN(0, 1, X_R - \mu_W, z_{1-\beta_n}). \tag{B.6}$$

On the good event G_μ, we always have that $X_R - \mu_W < X_W - \mu_W < z_{1-\beta_n}$, so the lower truncation is indeed below the upper one. Let

$$q^h_{\frac{1-\alpha}{1-\beta}}(x_r, \mu_w) = \text{Quantile}_\mu\left(\frac{1-\alpha}{1-\beta}, X_W - \mu_W \mid W = w, X_{-w} = x_{-w}, G_\mu\right) \tag{B.7}$$

denote the $(1-\alpha)/(1-\beta)$ quantile of the conditional distribution (B.6), which is a function of the largest entry x_r of x_{-w} and the winning mean μ_w. Per the prior discussion, the function (B.7) only makes sense if $x_r - \mu_0$ at most $z_{1-\beta_n}$, and we will take the quantile (B.7) to be $-\infty$ if it is not. It is then straightforward to show that

$$LCB_{hyb}(X) = \{\mu_0 : \mu_0 > X_W - q^h_{\frac{1-\alpha}{1-\beta}}(X_R, \mu_0)\} \tag{B.8}$$

contains μ_W with high probability conditional on W, X_{-W}, and the event G_μ:

$$P_\mu(\mu_W \in LCB_{hyb}(X) | W, X_{-W}, G_\mu) = P_\mu(X_W - \mu_W < q^h_{\frac{1-\alpha}{1-\beta}}(X_R, \mu_W) | W, X_{-W}, G_\mu) = \frac{1-\alpha}{1-\beta}.$$

Based on our earlier discussions, this is sufficient to imply that $LCB_{hyb}(X)$ from (B.8) will contain μ_W with probability at least $1-\alpha$. We argue later (via our p-value viewpoint) that (B.8) is an LCB.

Similarly, fixing some $\gamma_1, \gamma_2 > 0$ such that $\gamma_1 + \gamma_2 = 1 = (1-\alpha)/(1-\beta) = \frac{\alpha-\beta}{1-\beta}$, we can argue that

$$CI_{hyb}(X) = \{\mu_0 : X_W - q^h_{\gamma_2}(X_R, \mu_0) > \mu_0 > X_W - q^h_{1-\gamma_1}(X_R, \mu_0)\} \tag{B.9}$$

will contain μ_W with probability at least $1-\alpha$. Again, we will show via our p-value viewpoint that (B.9) is a CI.

B.10.2 The p-value viewpoint

For the same setting as above, we want to use the p-values $p_i^{\mu_0} = 1 - \Phi(X_i - \mu_0)$ to characterize when the hybrid LCB (B.8) does not contain $\mu_0 \in \mathbb{R}$. Examining (B.6), (B.7), and (B.8), we can consider two cases to figure out when this happens.

Case One - $X_R - \mu_0 \geq z_{1-\beta_n}$: If $X_R - \mu_0 \geq z_{1-\beta_n}$, then $q^h_{\frac{1-\alpha}{1-\beta}}(W, X_{-W}, \mu_0) = -\infty$, so μ_0 cannot be in (B.8). This case happens precisely when

$$X_R - \mu_0 \geq z_{1-\beta_n} \iff 1 - \Phi(X_R - \mu_0) \leq 1 - \Phi(z_{1-\beta_n}) \iff p^{\mu_0}_{(2)} \leq \beta_n.$$

Case Two - $X_R - \mu_0 < z_{1-\beta_n}$: If $X_R - \mu_0 < z_{1-\beta_n}$, then μ_0 is not in (B.8) exactly when $X_W - \mu_0$ is at least as large as the $\frac{1-\alpha}{1-\beta}$ quantile Q of a standard normal truncated to be larger than $X_R - \mu_0$ but smaller than $z_{1-\beta_n}$. This quantile satisfies

$$\frac{\alpha - \beta}{1 - \beta} = \frac{\Phi(z_{1-\beta_n}) - \Phi(Q)}{\Phi(z_{1-\beta_n}) - \Phi(X_R - \mu_0)} = \frac{1 - \beta_n - \Phi(Q)}{1 - \beta_n - \Phi(X_R - \mu_0)}$$

Solving for Q gives

$$Q = \Phi^{-1}\left(1 - \beta_n - \frac{\alpha - \beta}{1 - \beta}\left(1 - \beta_n - \Phi(X_R - \mu_0)\right)\right) = \Phi^{-1}\left(\left(1 - \frac{\alpha - \beta}{1 - \beta}\right)(1 - \beta_n) + \frac{\alpha - \beta}{1 - \beta}\Phi(X_R - \mu_0)\right),$$

so μ_0 is not in (B.8) exactly when

$$\begin{aligned} X_W - \mu_0 &\geq \Phi^{-1}\left(\left(1 - \frac{\alpha - \beta}{1 - \beta}\right)(1 - \beta_n) + \frac{\alpha - \beta}{1 - \beta}\Phi(X_R - \mu_0)\right) \\ \iff 1 - \Phi(X_W - \mu_0) &\leq \left(1 - \frac{\alpha - \beta}{1 - \beta}\right)\beta_n + \frac{\alpha - \beta}{1 - \beta}(1 - \Phi(X_R - \mu_0)) \\ &\iff p^{\mu_0}_{(1)} \leq \frac{\alpha - \beta}{1 - \beta}p^{\mu_0}_{(2)} + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right)\beta_n \end{aligned}$$

It turns out we can combine these two cases. Because $p^{\mu_0}_{(1)} \leq p^{\mu_0}_{(2)}$, the fact that $p^{\mu_0}_{(2)} \leq \beta_n$ in Case One implies that $p^{\mu_0}_{(1)} \leq \beta_n$ also. Therefore in Case One, $p^{\mu_0}_{(1)}$ must be strictly smaller than a mixture of $p^{\mu_0}_{(2)}$ and β_n:

$$p^{\mu_0}_{(1)} \leq \frac{\alpha - \beta}{1 - \beta}p^{\mu_0}_{(2)} + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right)\beta_n. \tag{B.10}$$

Therefore, if say μ_0 is not included whenever (B.10), we will always say it is not included in Case One, and we will say it is not included at the appropriate times in Case Two.

Now, we do the identical analysis for the hybrid CI (B.9). Case One is identical. We redo Case Two below.

Case Two - $X_R - \mu_0 < z_{1-\beta_n}$: If $X_R - \mu_0 < z_{1-\beta_n}$, then μ_0 is not in (B.9) exactly when $X_W - \mu_0$ is at least as large as the $1 - \gamma_1$ quantile of the same truncated normal, and at most as small as that truncated normal's γ_2 quantile. That is, when either

$$\begin{aligned} X_W - \mu_0 &\geq \Phi^{-1}\left(\gamma_1(1 - \beta_n) + (1 - \gamma_1)\Phi(X_R - \mu_0)\right) \\ &\iff p^{\mu_0}_{(1)} \leq \gamma_1 p^{\mu_0}_{(2)} + (1 - \gamma_1)\beta_n \end{aligned}$$

or

$$X_W - \mu_0 \leq \Phi^{-1}\left((1-\gamma_2)(1-\beta_n) + \gamma_2\Phi(X_R - \mu_0)\right)$$
$$\iff p_{(1)}^{\mu_0} \geq (1-\gamma_2)p_{(2)}^{\mu_0} + \gamma_2\beta_n$$

By the same reasoning as above, we can combine the two cases, and say that we do not include μ_0 whenever

$$p_{(1)}^{\mu_0} \leq \gamma_1 p_{(2)}^{\mu_0} + (1-\gamma_1)\beta_n \text{ or } p_{(1)}^{\mu_0} \geq (1-\gamma_2)p_{(2)}^{\mu_0} + \gamma_2\beta_n.$$

B.10.3 The hybrid LCB and CI for MLR families

Recall the setting where $X_i \sim P_{\theta_i}$ are independent samples from some parametric family P_θ with MLR in $T(x)$, and let $p_i^{\theta_0}$ be the UMP p-values for testing $H_0 : \theta \leq \theta_0$. Let $\Theta = (\theta_1, \ldots, \theta_n)$ be the parameter vector and let W $\text{argmin}_{i\in[n]}\, p_i^{\theta_0}$ be the index of the smallest p-value.

We will present a hybrid LCB and CI for θ_W and directly prove their validity. Importantly, note that the event $G_\Theta = \{p_i^{\theta_i} \geq \beta_n \text{ for all } i \in [n]\}$ happens with probability $1-\beta$ under P_Θ. Thus it suffices to show that our confidence regions have $(1-\alpha)/(1-\beta)$ coverage on this event, per our earlier discussion. Also importantly, thanks to $p_j^{\theta_j}$ having a uniform distribution under P_{θ_j} (see Lemma B.15.3), the conditional distribution of $p_j^{\theta_j}$ given $W = j$, $p_{-j}^{\theta_j}$, and G_Θ is a uniform distribution on $[\beta_n, \min_{i\neq j} p_i^{\theta_j}]$. Using this, we can see that

$$\left\{\theta_0 : p_{(1)}^{\theta_0} > \frac{\alpha-\beta}{1-\beta}p_{(2)}^{\theta_0} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n\right\}$$

is a $1-\alpha$ confidence region for θ_W:

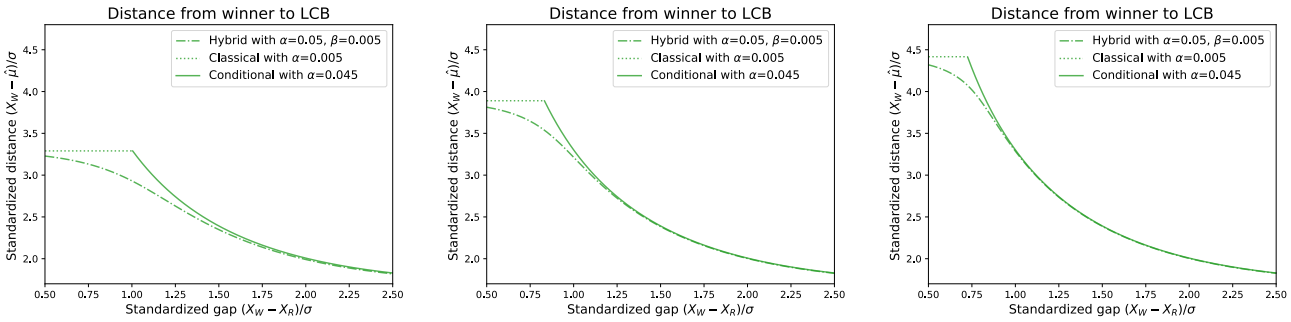
$$\begin{aligned}
&P_\Theta\left(\theta_W \in \left\{\theta_0 : p_{(1)}^{\theta_0} > \frac{\alpha-\beta}{1-\beta}p_{(2)}^{\theta_0} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n\right\} | W = j, p_{-j}^{\theta_j}, B_\Theta\right)\\
&= P_\Theta\left(\theta_j \in \left\{\theta_0 : \frac{p_j^{\theta_0} - \beta_n}{\min_{i\neq j} p_i^{\theta_0} - \beta_n} > \frac{\alpha-\beta}{1-\beta}\right\} | W = j, p_{-j}^{\theta_j}, B_\Theta\right)\\
&= P_\Theta\left(\frac{p_j^{\theta_0} - \beta_n}{\min_{i\neq j} p_i^{\theta_0} - \beta_n} > \frac{\alpha-\beta}{1-\beta} | W = j, p_{-j}^{\theta_j}, B_\Theta\right)\\
&= \frac{1-\alpha}{1-\beta},
\end{aligned}$$

To see that the region is actually an LCB, we rewrite it as

$$\left\{\theta_0 : \frac{p_{(1)}^{\theta_0}}{p_{(2)}^{\theta_0}} > \frac{\alpha-\beta}{1-\beta} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n \Big/ p_{(2)}^{\theta_0}\right\} \tag{B.11}$$

because we know that $p_{(2)}^{\theta_0}$ is monotone non-decreasing in θ_0 and also from Appendix B.8 that $p_{(1)}^{\theta_0}/p_{(2)}^{\theta_0}$ is monotone non-decreasing in θ_0, once some θ_0 belongs to (B.11) all larger θ_0 will as well. Hence, it is an LCB.

Considering $\gamma_1, \gamma_2 > 0$ such that $\gamma_1 + \gamma_2 = (\alpha-\beta)/(1-\beta)$, we can do the exact same analysis for the



(a) $n = 10$ (b) $n = 100$ (c) $n = 1000$

Figure B.1: For $n = 10$ (left), $n = 100$ (middle), and $n = 1000$ (right) the distance between the hybrid LCB to the winner (dash-dot line) and union bound LCB to the winner (dotted and solid line) with $\alpha = 0.05$ and $\beta = 0.005$ plotted as a function of the gap between the winning and runner-up observation.

hybrid CI

$$\left\{\theta_0 : \gamma_1 p^{\theta_0}_{(2)} + (1-\gamma_1)\beta_n < p^{\theta_0}_{(1)} < (1-\gamma_2)p^{\theta_0}_{(2)} + \gamma_2\beta_n\right\}$$

to show that it both covers θ_W with probability at least $1-\alpha$ and that it is indeed a CI.

Note that, when put in our selective dominance framework, the hybrid selection event actually depends on the null parameter θ_0. Hence we cannot argue monotonicity of the selective p-value via Appendix B.15.3 as we did in the conditional case, and instead had to use different arguments to justify that our confidence regions were actually an LCB and CI.

B.10.4 Distance between winner and hybrid LCB

Consider observing Gaussian data $X \sim N(\mu, \sigma^2 I_n)$ and let W and R be the indices of the winner and runner up respectively. We will use our p-value viewpoint to show that the standardized distance $D = (X_W - X_R)/\sigma$ between the winner X_W and the hybrid LCB $\hat{\mu}$ for μ_W depends only on the standardized gap $(X_W - X_R)/\sigma$ between the winner and runner-up. Per our earlier discussion the hybrid LCB $\hat{\mu}$ satisfies

$$\begin{aligned}
p^{\hat{\mu}}(X_W) &= \frac{\alpha-\beta}{1-\beta}p^{\hat{\mu}}(X_R) + \left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n \\
1 - \Phi\left(\frac{X_W - \hat{\mu}}{\sigma}\right) &= \frac{\alpha-\beta}{1-\beta}\left(1 - \Phi\left(\frac{X_R - \hat{\mu}}{\sigma}\right)\right) + \left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n \\
\iff 1 - \Phi(D) &= \frac{\alpha-\beta}{1-\beta}\left(1 - \Phi\left(D - \frac{X_W - X_R}{\sigma}\right)\right) + \left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n.
\end{aligned}$$

Clearly then, D is a function of $(X_W - X_R)/\sigma$ and n.

B.11 Comparing hybrid inference to the union bound

As discussed earlier the hybrid cutoff (5.9) is strictly larger than the union bound cutoff (5.11) when $p_{(1)} > \beta_n$. Thus, both procedures reject when $p_{(1)} \leq \beta_n$. When $p_{(1)} > \beta_n$, the hybrid procedure rejects but

the union bound does not whenever

$$p_{(1)} \in \left((\alpha-\beta)p_{(2)}, \frac{\alpha-\beta}{1-\beta}p_{(2)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right)\beta_n\right]$$

When $p_{(2)} = 1$ and $n = 1$, the length of the interval is β, which is the largest it can possible be. Thus, in the case where we can have additional rejections, the hybrid cutoff is never more than β plus the union bound cutoff. Andrews et al. [2023] suggests taking $\beta = \alpha/10$, so when $\alpha = 0.05$, for example, $\beta = 0.005$ is quite small.

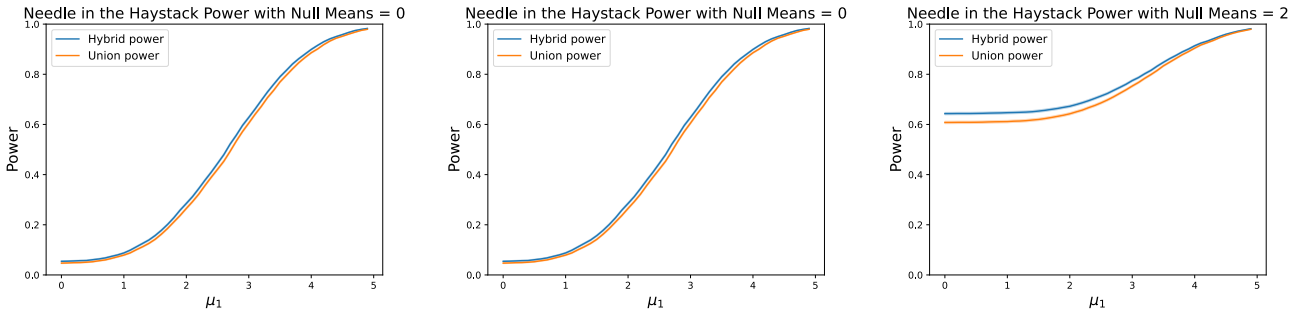
Still, this is not a precise statement about power gain. The computations required to compute the power gain analytically are messy, so instead we gauge the power gain via simulation. We sample data $X \sim N(\mu, I_n)$ for $n = 10$ and attempt to reject the winning null $H_W : \mu_W \leq 0$ where $W = \text{argmin}_{i\in[n]} 1 - \Phi(X_i)$ is the index of the smallest p-value $p_i = 1 - \Phi(X_i)$.

We choose $n = 10$ because it is a reasonably small dimension size where one may apply hybrid inference (e.g., the main example from Andrews et al. [2023] has $n = 13$). Let R denote the index of the runner-up (second smallest p-value). For the dimensions $n = 10, 100, 1000$, Figure B.1 compares the distance from the winner X_W the hybrid LCB and the union bound LCB. As illustrated in the plot, the benefit of hybrid inference dissipates as the dimension of the problem increases. This is because conditioning on the "good event" has less and less of an effect as n grows.

Our simulation results indicate that hybrid inference typically results in a fairly small power gain. We consider two simulated settings:

Needle in a haystack: First, we consider a needle-in-the-haystack setting where $\mu_1 > 0$ and all the other μ_i for $i \neq 1$ are set to μ_2. We try $\mu_2 = -2, 0, 2$. The power comparison is ploted in Figure B.2. Whenever we truly have a needle-in-the-haystack problem, i.e., $\mu_1 > \mu_2$, hybrid inference results in essentially no power gain. The only setting where we see some gain (up to around 0.05) is when $\mu_2 > \mu_1$. In this setting we actually have a dense alternative (many small signals). We expect the top two p-values to be close to each other, so conditional methods should perform poorly. The union bound approach indeed performs essentially identically to the level β classical test (not pictured). Hybrid, however, manages to eke out some additional power. Both methods pale in comparison to the level α classical test however, which would achieve power > 0.95 throughout the whole plot (not pictured). For various values of σ_1 and σ_2, which we assume are known, we also tried re-running the experiments with $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_i \sim N(\mu_2, \sigma_2^2)$ when $i > 1$. The results were not appreciably different.

Two possible signals: Seeing as the hybrid and union bound approaches both reject based on the winning and running-up p-value, we ran a simulation for all pairs $\mu_1, \mu_2 \in \{-3, -2.9, \ldots, 2.9, 3\}$ with $\mu_1 > \mu_2$ and $\mu_1 > 0$. We forcibly set $\mu_i = -\infty$ for $i > 2$. For each setting we ran $N = 10000$ to get an empirical estimate of power for each method. Across the 1492 simulated settings, the median empirical power increase from hybrid was ≈ 0.003, the 90th percentile empirical power increase was ≈ 0.023 and the maximum empirical power increase was ≈ 0.042. As the results indicate, the power increases from hybrid were negligible for most settings. We also re-ran the same simulations but sampled $X_1 \sim N(\mu, \sigma_1^2)$ and $X_2 \sim N(\mu, \sigma_2^2)$ for various values of σ_1 and σ_2, which we assume are known. The results were not appreciably different, and, if anything, the difference in power was notably smaller for some values of σ_1 and σ_2.



(a) $\mu_2 = -2$ (b) $\mu_2 = 0$ (c) $\mu_2 = 2$

Figure B.2: For $\mu_2 = 0$ (left), $\mu_2 = -0.5$ (middle), and $\mu_2 = -1$ (right) the empirical power over $N = 10^4$ trials of the hybrid inference approach versus the union bound approach for the needle-in-the-haystack alternative. One standard error bands are also plotted. For the most part, they are so small that they are hardly visible.

B.12 Number of ties in rank verification

Considering a random vector $X \in \mathbb{R}^n$ let

$$Y_i = \frac{X_i - X_j}{2}, \qquad Y_j = \frac{X_i + X_j}{2}, \qquad Y_\ell = X_\ell \text{ for } \ell \neq i, j, \tag{B.12}$$

Suppose that $X_i \geq X_k$ for all $k \neq i$, and $X_i = X_\ell$ for some $\ell \neq i$ (i.e., there is at least one tie). We count the number of ties in for the winner in three different cases.

- Suppose $Y_j > \max_{k \neq i,j} Y_k$. If any of the X_k for $k \neq i, j$ were equal to X_i, then we would have

$$Y_k = X_k = \frac{X_i + X_k}{2} \geq \frac{X_i + X_j}{2} = Y_j$$

  which would be a contradiction. Thus, if there is a tie, the only possible tie is X_j. Since,

$$1 + |\{\ell \neq i : Y_\ell = \max_{k \neq i} Y_k\}| = 2$$

  in this case, we are done.

- Suppose that $Y_j < \max_{k \neq i,j} Y_k$. Then some X_k for $k \neq i, j$ must be strictly than larger X_j. Otherwise we would have for every $k \neq i, j$ that

$$Y_j = \frac{X_i + X_j}{2} \geq \frac{X_i + X_k}{2} \geq X_k = Y_k,$$

  which is a contradiction. Thus, if there is a tie, the number of ties is one plus the number of X_k for $k \neq i, j$ that are equal to each other. In this case, this matches

$$1 + |\{\ell \neq i : Y_\ell = \max_{k \neq i} Y_k\}|,$$

  so we are done.

- Now suppose that $Y_j = \max_{k \neq i,j} Y_k$. In this case we must have $X_j = X_i$. If not, then $X_j < X_i$ and

there must be some $\ell \neq i, j$ such that $X_i = X_\ell$, so

$$Y_j = \frac{X_i + X_j}{2} < \frac{X_i + X_\ell}{2} = X_\ell = Y_\ell,$$

which is a contradiction. Thus $Y_j = X_i$ in this case, and the number of ties is therefore clearly

$$1 + |\{\ell \neq i : Y_\ell = \max_{k \neq i} Y_k\}|.$$

B.13 Selective p-value for rank verification in exponential families

Suppose p is a p-value for the null H_0 that is selectively dominant given Z and we select p to use for inference according to the selection function

$$s(x,z) = \begin{cases} 1 & \text{if } x < q^+(z), \\ \frac{1}{N(z)} & \text{if } x \in [q^+(z), q(z)], \\ 0 & \text{otherwise,} \end{cases},$$

Then the adjusted p-value (4.4) is given by

$$p_{sel} = \frac{\int_0^p s(x,Z)dx}{\int_0^1 s(x,Z)dx} = \begin{cases} \frac{p}{q^+(Z) + \frac{1}{N(Z)}(q(Z) - q^+(Z))} & \text{if } p < q^+(Z), \\ \frac{q^+(Z) + \frac{1}{N(Z)}(p - q^+(Z))}{q^+(Z) + \frac{1}{N(Z)}(q(Z) - q^+(Z))} & \text{if } p \in [q^+(Z), q(Z)], \end{cases},$$

which can be re-written as

$$p_{sel} = \frac{p - (1 - \frac{1}{N(Z)})(p - q^+(Z))_+}{q^+(Z) + \frac{1}{N(Z)}(q^+(Z) - q(Z))}.$$

This is sufficient to imply the claim from the main text.

B.14 Rank verification additional details

Example 12 (Rank verification in exponential families)**.** *Suppose p is a p-value for the null H_0 that is selectively dominant given Z. If we select p to use for inference according to the selection function*

$$s(x,z) = \begin{cases} 1 & \textit{if } x < q^+(z), \\ \frac{1}{N(z)} & \textit{if } x \in [q^+(z), q(z)], \\ 0 & \textit{otherwise,} \end{cases},$$

where $N(z) > 1$ and $0 \leq q^+(z) \leq q(z) \leq 1$ are known functions of z, then the selective p-value from (4.4) turns out to be (see Appendix B.13 for computations)

$$p_{sel} = \frac{p - (1 - \frac{1}{N(Z)})(p - q^+(Z))_+}{q^+(Z) + \frac{1}{N(Z)}(q(Z) - q^+(Z))}. \tag{B.13}$$

Therefore, Theorem 5 tells us that rejecting when (B.13) is at most α is a selective Type I error controlling procedure:

$$P_{H_0}\left(\frac{p-(1-\frac{1}{N(Z)})(p-q^+(Z))_+}{q^+(Z)+\frac{1}{N(Z)}(q(Z)-q^+(Z))}.\leq\alpha \mid S=1\right)\leq\alpha. \tag{B.14}$$

Now, suppose we observe X drawn from the exponential family (6.7) and let W be the index of the largest X_i (with ties broken randomly). For $i\neq j$, Example 4 tells us that the UMPU p-value $p=p_{ij}$ from (6.11) for the null $H_{0,ij}:\theta_i\leq\theta_j$ is selectively dominant given the transformed nuisance statistics Y_{-i} from (6.8). Taking $q^+_{ij}(Y_{-i})$, $q_{ij}(Y_{-i})$, $N_i(Y_{-i})$, and f from (6.12), (6.13), (6.14), it is now easy to show that rejecting the data-dependent null $H_{0,Wj}$ for $j\neq W$ when

$$\frac{p_{Wj}-(1-\frac{1}{N_W})(p_{Wj}-q^+_{Wj})_+}{q^+_{Wj}+\frac{1}{N_W}(q_{Wj}-q^+_{Wj})}\leq\alpha \tag{B.15}$$

controls Type I error conditionally on W.

For $i\neq j$, if $H_{0,ij}$ is false, then trivially $P(\text{falsely reject } H_{0,Wj}|W=i)$. For the case that $H_{0,ij}$ is true, the event $W=i$ is the same event as selecting p_{ij} for inference in (B.13), so

$$\begin{aligned}
P(\text{falsely reject } H_{0,Wj}|W=i) &= P\left(\frac{p_{Wj}-(1-\frac{1}{N_W})(p_{Wj}-q^+_{Wj})_+}{q^+_{Wj}+\frac{1}{N_W}(q_{Wj}-q^+_{Wj})}\leq\alpha \,\middle|\, W=i\right)\\
&= P\left(\frac{p_{ij}-(1-\frac{1}{N_i})(p_{ij}-q^+_{ij})_+}{q^+_{ij}+\frac{1}{N_i}(q_{ij}-q^+_{ij})}\leq\alpha \,\middle|\, W=i\right),\\
&\leq\alpha.
\end{aligned}$$

which is sufficient to imply conditional error control.

Now suppose we reject the data-dependent null $\cup_{j\neq W}H_{0,Wj}$ when we reject all of the individual nulls $H_{0,Wj}$ for $j\neq W$. It is straightforward to see that this will control Type I error both conditionally on W and marginally. If i is such that $\theta_i>\theta_j$ for all $j\neq i$, then trivially $P(\text{falsely reject } \cup_{j\neq W}H_{0,Wj}|W=i)=0$. If this is not true, then there exists some $k\neq i$ such that $\theta_k\geq\theta_i$, so

$$P(\text{falsely reject } \cup_{j\neq W}H_{0,Wj}|W=i)\leq P(\text{falsely reject } H_{0,Wk}|W=i)\leq\alpha.$$

Again, marginal error control follows from the usual law of total probability argument:

$$\begin{aligned}
P(\text{falsely reject } \cup_{j\neq W}H_{0,Wj}) &= \sum_{i=1}^n P(W=i)P(\text{falsely reject } \cup_{j\neq W}H_{0,Wj}|W=i)\\
&\leq\alpha\sum_{i=1}^n P(W=i)\\
&=\alpha.
\end{aligned}$$

B.15 Selective dominance and one-sided testing

In this appendix, we establish the selective dominance property for UMP p-values in MLR families and UMPU p-values in exponential families. We also show that in these cases, the adjusted p-value from Theorem 5 is monotone in the parameter, under suitable conditions. Throughout the appendix, we draw from the discussion and proof strategy in Appendix B.1 of Lei and Fithian [2018].

B.15.1 MLR Families

We consider a parametric family P_θ parameterized by a real parameter $\theta \in R$ such that each P_θ has density $p_\theta(x)$ with respect to some carrier measure μ. We will suppose that these densities share support (i.e., for any two θ and θ' we have $p_\theta(x) > 0 \iff p_{\theta'}(x) > 0$). Further, we suppose that for any $\theta \leq \theta'$, the likelihood ratio $p_{\theta'}(x)/p_\theta(x)$ is a monotone non-decreasing function of some real-valued function $T(x)$ on this support (i.e., for any $x_1 \leq x_2$ with $p_\theta(x_1), p_\theta(x_2) > 0$, we have $p_{\theta'}(x_1)/p_\theta(x_1) \leq p_{\theta'}(x_2)/p_\theta(x_2)$).

Recall that for a testing problem, the critical function $\phi(x)$ (see [Lehmann et al., 1986, Section 3.1]) tells us the probability of rejecting the null having observed data x, so $\phi(X) = P(\text{reject } H_0|X)$. From Theorem 3.4.1 of Lehmann et al. [1986] we know the test governed by the critical function

$$\phi(x) = \begin{cases} 1 & \text{if } T(x) > C \\ \gamma & \text{if } T(x) = C \\ 0 & \text{otherwise} \end{cases} \tag{B.16}$$

is UMP for testing $H_0 : \theta \leq \theta_0$ against the alternatives $H_a : \theta > \theta_0$ so long C and γ satisfy

$$P_{\theta_0}(T(X) > C) + \gamma P_{\theta_0}(T(X) = C) = \alpha. \tag{B.17}$$

Denote the left-continuous survival function of $T(X)$ and its right-hand limit under P_{θ_0} as

$$G(t) = P_{\theta_0}(T(X) \geq t) \qquad G^+(t) = \lim_{u \downarrow t} G(u).$$

Since $G(t)$ is a monotone non-increasing function, we can also define its generalized inverse

$$G^{-1}(z) = \inf\{t : G(t) \leq z\},$$

Lemma B.15.1 gives a natural way to set C and γ in Equation (B.16) to get an UMP test.

Lemma B.15.1 (An UMP test)**.** *Adopting the convention that* $0/0 = 0$*, taking* $C = G^{-1}(\alpha)$ *and* $\gamma = (\alpha - G^+(C))/(G(C) - G^+(C))$ *in Equation* (B.16) *gives an UMP test.*

Proof. At continuity points t of $G(\cdot)$, we have $G(G^{-1}(t)) = t$ and also $P(T(X) = t) = 0$. Thus, if $C = G^{-1}(\alpha)$ is a continuity point of $G(\cdot)$, then $G^+(C) = P_{\theta_0}(T(X) > C) = P_{\theta_0}(T(X) \geq C) = G(C) = \alpha$ and $\gamma = 0$, so the constraint (B.17) is immediately satisfied.

If $C = G^{-1}(\alpha)$ is not a continuity point of $G(\cdot)$, then $G(C) - G^+(C) > 0$ and the constraint (B.17) is also immediately satisfied. To ensure that we still have a valid test, however, we need $\gamma \in [0, 1]$. This is true so long as $\alpha \in [G^+(C), G(C)]$. We know that $G(t) \leq \alpha$ for any $t > C$, so $G^+(C) \leq \alpha$. If $G(C) < \alpha$, then we

could find some $t^- < C$ such that $G(C) < \alpha$ by left-continuity, but this would contradict that $C = G^{-1}(\alpha)$, finishing the proof. □

Letting $U_{aux} \sim \text{Unif}([0,1])$ be auxiliary randomness that is independent of X, a simple way to instantiate the test from Lemma B.15.1 is to reject whenever $T(X) > C$ or when $T(X) = C$ and $U_{aux} \leq \gamma$. Lemma B.15.2 explains how this is the same as rejecting when the p-value, termed as a fuzzy p-value in Geyer and Meeden [2005],

$$p = G^+(T(X)) + U_{aux}(G(T(X)) - G^+(T(X))) \tag{B.18}$$

is at most α.

Lemma B.15.2 (Fuzzy p-value is UMP)**.** *Rejecting $H_0 : \theta \leq \theta_0$ when the fuzzy p-value (B.18) is at most α instantiates the test from Lemma B.15.1, and is therefore UMP.*

Proof. We rewrite

$$p = (1 - U_{aux})G^+(T(X)) + U_{aux}G(T(X))$$

and consider four cases.

- If $t < G^{-1}(\alpha)$ then we can find some $t^+ > t$ such that $G(t^+) > \alpha$. Thus $G(t) \geq G^+(t) > \alpha$. So $p > \alpha$ whenever $T(X) < G^{-1}(\alpha)$
- If $t = G^{-1}(\alpha)$ and $G^{-1}(\alpha)$ is a continuity point of $G(\cdot)$, then $G^+(t) = G(t) = \alpha$. Thus, in this case $p \leq \alpha$ whenever $T(X) = G^{-1}(\alpha)$ and $U_{aux} \leq \frac{\alpha - G^+(G^{-1}(\alpha))}{G(G^{-1}(\alpha)) - G^+(G^{-1}(\alpha))} = \infty$.
- If $t = G^{-1}(\alpha)$ and $G^{-1}(\alpha)$ is not a continuity point of $G(\cdot)$, then we must have that $G(t) - G^+(t) > 0$. Also by right continuity we have $G(t) \geq \alpha$ and by how $G^{-1}(\cdot)$ is defined we have $G^+(t) \leq \alpha$. In this case $p \leq \alpha$ also whenever $T(X) = G^{-1}(\alpha)$ and $U_{aux} \leq \frac{\alpha - G^+(G^{-1}(\alpha))}{G(G^{-1}(\alpha)) - G^+(G^{-1}(\alpha))}$.
- If $t > G^{-1}(\alpha)$ then $G^+(t) \leq G(t) \leq \alpha$. So $p \leq \alpha$ whenever $T(X) > G^{-1}(\alpha)$.

This implies the following set equalities:

$$\begin{aligned}\{p \leq \alpha\} &= \{T(X) > G^{-1}(\alpha)\} \cup \left\{T(X) = G^{-1}(\alpha), U_{aux} \leq \frac{\alpha - G^+(G^{-1}(\alpha))}{G(G^{-1}(\alpha)) - G^+(G^{-1}(\alpha))}\right\} \\ &= \{T(X) > C\} \cup \{T(X) = C, U_{aux} \leq \gamma\}\end{aligned}$$

□

Lemma B.15.3 shows that $p \sim \text{Unif}([0,1])$ under P_{θ_0}, which will be useful for us later.

Lemma B.15.3 (Fuzzy p-value is uniform at null boundary)**.** *Under P_{θ_0}, the p-value (B.18) has a Unif$([0,1])$ distribution.*

Proof. Using the set equality from Lemma B.15.2 but replacing α with $z \in (0,1)$, we find

$$
\begin{aligned}
&P_{\theta_0}(G^+(T(X)) + U_{aux}(G(T(X)) - G^+(T(X))) \leq z) \\
&= P_{\theta_0}(T(X) > G^{-1}(z)) + P_{\theta_0}\left(T(X) = G^{-1}(z), U \leq \frac{z - G^+(G^{-1}(z))}{G(G^{-1}(z)) - G^+(G^{-1}(z))}\right) \\
&= P_{\theta_0}(T(X) > G^{-1}(z)) + P_{\theta_0}(T(X) = G^{-1}(z)) P_{\theta_0}\left(U \leq \frac{z - G^+(G^{-1}(z))}{G(G^{-1}(z)) - G^+(G^{-1}(z))}\right) \\
&= G^+(G^{-1}(z)) + (G(G^{-1}(z)) - G^+(G^{-1}(z))) \cdot \frac{z - G^+(G^{-1}(z))}{G(G^{-1}(z)) - G^+(G^{-1}(z))} \\
&= z.
\end{aligned}
$$

□

Now we can show that p is a selectively dominant p-value for testing the null $H_0 : \theta \leq \theta_0$. In what follows, we consider some fixed $\theta \leq \theta_0$ and prove some facts that allow us to relate the distribution of $T(X)$ under P_{θ_0} to its distribution under P_θ.

Lemma B.15.4 (Distribution of $T(X)$)**.** *Let $g_\theta(T(x))$ be a non-increasing function that equals the likelihood ratio $p_\theta(x)/p_{\theta_0}(x)$ on the support and*

$$\nu(A) = \int I(T(x) \in A) p_{\theta_0}(x) \mu(dx)$$

be the measure of $T(X)$ under $X \sim P_{\theta_0}$. Then

$$P_\theta(T(X) \in A) = \int_A g_\theta(t) \nu(dt)$$

Proof. We know that

$$P_\theta(T(X) \in A) = \int I(T(x) \in A) p_{\theta_0}(x) g_\theta(T(x)) \mu(dx),$$

so we need to show that

$$\int I(T(x) \in A) p_{\theta_0}(x) g_\theta(T(x)) \mu(dx) = \int_A g_\theta(t) \nu(dt) \tag{B.19}$$

If $g_\theta(T(x)) = I(T(x) \in A')$ happens to be an indicator then (B.19) holds. Therefore, we can apply the standard machine (see the discussion after Equation 42 in Lei and Fithian [2018]) to show that (B.19) holds for all non-negative functions $g_\theta(\cdot)$. □

Lemma B.15.5 (Distribution of $(T(X), U_{aux})$)**.** *If ω denotes the product measure of ν and Lebesgue measure λ on $[0, 1]$, i.e., the distribution of $(T(X), U_{aux})$ under P_{θ_0}, then*

$$P_\theta((T(X), U) \in B) = \int_B g_\theta(t) \omega(dt, du) \tag{B.20}$$

Proof. We will first argue that (B.20) holds for any B which is a product set $A_1 \times A_2$. We can further reduce

to the case that $g_\theta(T(x)) = I(T(x) \in A_1')$ is an indicator. Then we see using our previous lemma that

$$\begin{aligned} P_\theta((T(X), U_{aux}) \in A_1 \times A_2) &= P_\theta(T(X) \in A_1) P(U_{aux} \in A_2) \\ &= \int_{A_1} g_\theta(t)\nu(dt) \cdot \int_{A_2} \lambda(du) \\ &= \int_{A_1 \cap A_1'} \nu(dt) \cdot \int_{A_2} \lambda(du) \\ &= \int_{A_1 \cap A_1' \times A_2} \omega(dt, du) \\ &= \int_{A_1 \times A_2} g_\theta(t)\omega(dt, du) \end{aligned}$$

To handle the case of general $g_\theta(\cdot)$ we can again simply apply the standard machine.

The full result then follows from an application of the $\pi - \lambda$ theorem: the set of B for which (B.20) holds is a λ-system, and (B.20) holds for every set in the π system of all product sets $B = A_1 \times A_2$. □

Note that our p-value p is a deterministic function of $T(X)$ and U_{aux}:

$$p = m(T(X), U_{aux}) \qquad m(t, u) = G^+(t) + u(G(t) - G^+(t)).$$

As such, we sometimes write our selection function as a function of $T(X)$ and U_{aux}:

$$s(t, u) = s(m(t, u)).$$

We use this abuse of notation in our next lemma, which characterizes the conditional distribution of $T(X)$ given selection.

Lemma B.15.6 (Distribution of $(T(X), U_{aux})$ given selection)**.** *For any selection function $s(x)$ under which p has a positive probability of selection under P_θ,*

$$P_\theta((T(X), U_{aux}) \in B | S = 1) = \frac{\int_B g_\theta(t)s(t, u)\omega(dt, du)}{\int g_\theta(t)s(t, u)\omega(dt, du)}$$

Proof. First note that

$$P_\theta((T(X), U_{aux}) \in B | S = 1) = \frac{P_\theta((T(X), U_{aux}) \in B, S = 1)}{P_\theta(S = 1)}.$$

Thus it suffices to show for any set B that

$$P_\theta((T(X), U_{aux}) \in B, S = 1) = \int_B g_\theta(t)s(t, u)\omega(dt, du).$$

By the definition of conditional expectation

$$\begin{aligned} P_\theta((T(X), U_{aux}) \in B, S = 1) &= E_\theta[E_\theta[I(S = 1) \mid T(X), U_{aux}] I((T(X), U_{aux}) \in B)] \\ &= E_\theta[s(T(X), U_{aux}) I((T(X), U_{aux}) \in B)] \end{aligned}$$

If $s(t,u) = I((t,u) \in B)$ is an indicator function, then the result is implied by our previous lemma. We again get the result for general selection functions $s(t,u)$ by applying the standard machine. □

With these lemmas under our belt, we can show Proposition B.15.1, the main result of this sub-section. Since $p \sim_{P_{\theta_0}} \text{Unif}([0,1])$ by Lemma B.15.3, this proposition is sufficient to imply selective dominance.

Proposition B.15.1. *For any selection function $s(x)$ for which p has positive probability of selection under P_θ,*

$$P_\theta(p \leq z | S = 1) \leq P_{\theta_0}(p \leq z | S = 1).$$

Proof. First we show that $P_\theta(S=1) > 0 \implies P_{\theta_0}(S=1) > 0$. We show the contrapositive. If

$$\begin{aligned} 0 = P_{\theta_0}(S=1) &= E_{\theta_0}[E_{\theta_0}[S|p]] \\ &= E_{\theta_0}[s(p)] \\ &= E_{\theta_0}[s(T(X), U_{aux})], \end{aligned}$$

then $s(T(X), U_{aux}) = 0$ a.e. under P_{θ_0}. Therefore,

$$P_\theta(S=1) = E_\theta[s(T(X), U_{aux})] = E_{\theta_0}[g_\theta(T(X)) s(T(X), U_{aux})] = 0.$$

Now, fix $z \in (0,1)$. If z is such that $P_\theta(p \leq z | S = 1) = 0$ then the desired inequality is trivial. To handle the non-trivial case, we recall Lemma B.15.2 and note three facts:

- If $(t,u) \in m^{-1}([0,z])$ then $t \geq G^{-1}(z)$,
- If $(t,u) \in m^{-1}((z,1])$ then $t \leq G^{-1}(z)$,
- The sets $m^{-1}([0,z])$ and $m^{-1}((z,1])$ are disjoint.

Thus,

$$\begin{aligned} \frac{1}{P_\theta(p \leq z | S=1)} &= \frac{\int_{m^{-1}([0,1])} g_\theta(t) s(t,u) \omega(dt, du)}{\int_{m^{-1}([0,z])} g_\theta(t) s(t,u) \omega(dt, du)} \\ &= \frac{\int_{m^{-1}([0,z])} g_\theta(t) s(t,u) \omega(dt, du) + \int_{m^{-1}((z,1])} g_\theta(t) s(t,u) \omega(dt, du)}{\int_{m^{-1}([0,z])} g_\theta(t) s(t,u) \omega(dt, du)} \\ &= 1 + \frac{\int_{m^{-1}((z,1])} g_\theta(t) s(t,u) \omega(dt, du)}{\int_{m^{-1}([0,z])} g_\theta(t) s(t,u) \omega(dt, du)} \\ &\geq 1 + \frac{g_\theta(G^{-1}(z)) \int_{m^{-1}((z,1])} s(t,u) \omega(dt, du)}{g_\theta(G^{-1}(z)) \int_{m^{-1}([0,z])} s(t,u) \omega(dt, du)} \\ &= 1 + \frac{\int_{m^{-1}((z,1])} s(t,u) \omega(dt, du)}{\int_{m^{-1}([0,z])} s(t,u) \omega(dt, du)} \\ &= \frac{1}{P_{\theta_0}(p \leq z | S=1)}, \end{aligned}$$

where to finish we have noted that $g_{\theta_0}(t) = 1$ almost everywhere in the measure ω. □

B.15.2 Exponential families

Suppose we observe data $X \in \mathbb{R}^m$ from an exponential family P_θ parameterized by $\theta \in \mathbb{R}^n$ i.e., under P_θ the data X has density

$$g_\theta(x) = \exp(\theta_1 T_1(x) + \cdots + \theta_n T_n(x) - \psi(\theta))g(x)$$

with respect to some carrier measure μ. We consider the problem of testing $H_0 : \theta_i \leq \theta_{0,i}$.

The UMPU test for $H_0 : \theta_i \leq \theta_{0,i}$ is valid conditional on $T_{-i}(X)$. More specifically, Theorem 4.4.1 of Lehmann et al. [1986] tells us that any test of the form

$$\phi(t_i, t_{-i}) = \begin{cases} 1 & \text{if } t_i > C_0(t_{-i}) \\ \gamma(t_{-i}) & \text{if } t_i = C_0(t_{-i}) \\ 0 & \text{otherwise} \end{cases}$$

where the functions $\gamma(\cdot)$ and $C_0(\cdot)$ satisfy

$$E_{\theta_i=\theta_{0,i}}[\phi(T_i(X), t_{-i})|T_{-i}(X)] = \alpha \text{ a.e. under } P_{\theta_i=\theta_{0,i}}$$

is UMPU for testing $H_0 : \theta_i \leq \theta_{0,i}$. Lemma 2.7.2 of Lehmann et al. [1986] tells us that the conditional distribution of $T_i(X)$ given $T_{-i}(X) = t_{-i}$ admits a density

$$g_{\theta_i, t_{-i}}(t_i) = \exp(\theta_i t_i - \tilde{\psi}(\theta_i))$$

with respect to some base measure $\mu_{t_{-i}}$. This density has an MLR in t_i (to be specific, we are imagining observing $T_i(X)$ from its conditional distribution $T_{-i}(X)$, and the map $T(\cdot)$ from the previous sub-section is actually the identity). Hence, a concrete UMPU test is to just run our UMP test from the previous section using the conditional distribution given $T_{-i}(X) = t_{-i}$. In particular, our work from the previous section implies that it is UMPU to reject using the p-value

$$p = G^+(T_i(X)|T_{-i}(X)) + U_{aux}(G(T_i(X)|T_{-i}(X)) - G_i^+(T_i(X)|T_{-i}(X))), \tag{B.21}$$

where U_{aux} is an uniform random variable independent of the data and

$$G(t_i|t_{-i}) = P_{\theta_0}(T_i(X) \geq t|T_{-i}(X) = t_{-i}) \qquad G^+(t_i|t_{-i}) = \lim_{u \downarrow t_i} G(u|t_{-i}).$$

We argue that this p-value is selectively dominant given Z. Fix a distribution P in the null. If we consider some selection function $s(x, z)$ such that $P(S = 1) = 0$, then (4.3) holds trivially. Instead consider the case that $P(S = 1) > 0$ and let $f_z(x)$ denote the conditional PDF of p given $Z = z$. Proposition B.15.1 along with Lemma B.15.3 tells us exactly that for z such that $\int_0^1 s(x, z) f_z(x) dx > 0$,

$$\frac{\int_0^t s(x, z) f_z(x)}{\int_0^1 s(x, z) f_z(x)} \leq \frac{\int_0^t s(x, z)}{\int_0^1 s(x, z)} \text{ for all } t \in [0, 1].$$

Arguments in Appendix B.19.1 and Appendix B.19.2 then tell us that the following three facts are true a.e. under $P(\cdot|S = 1)$: (1) $\int_0^1 s(x, Z) f_Z(x) dx > 0$, (2) the left-hand side of the above equals the left-hand side

of (4.3), and (3) the right-hand side of the above equals the right-hand side of (4.3). Combined these facts imply that p is selectively dominant given Z.

B.15.3 Monotonicity of selective MLR p-values

In this sub-section, we consider data (X, Z) where the conditional distribution $X|Z = z \sim P_{\theta,z}$ is parameterized by $\theta \in \mathbb{R}$ and has an MLR in $T(x)$. Because we do everything conditional on $Z = z$, without loss of generality we can just work with X and understand that the results will hold a.e. over Z. Letting

$$G^{\theta_0}(t) = P_{\theta_0}(T(X) \geq t) \qquad G^{\theta_0,+}(t) = \lim_{u \downarrow t} G^{\theta_0}(u)$$

we let

$$p^{\theta_0} = G^{\theta_0,+}(T(X)) + U_{aux}(G^{\theta_0}(T(X)) - G^{\theta_0,+}(T(X)))$$

be the UMP p-value for testing $H_0 : \theta \leq \theta_0$. Again, U_{aux} is a uniform random variable independent of the data. Let

$$m^{\theta_0}(t, u) = G^{\theta_0,+}(t) + u(G^{\theta_0}(t) - G^{\theta_0,+}(t))$$

be the map such that $m^{\theta_0}(T(X), U) = p^{\theta_0}$.

Considering a class of selection functions $s^{\theta_0}(x)$ such that $\int_0^1 s^{\theta_0}(x) > 0$ (when Z is present we need $\int_0^1 s^{\theta_0}(x, Z) > 0$ almost surely), we want to show that the selective p-values from Theorem 5,

$$p^{\theta_0}_{sel} = \frac{\int_0^{p^{\theta_0}} s^{\theta_0}(x)dx}{\int_0^1 s^{\theta_0}(x)dx},$$

are monotone non-decreasing in θ_0. Specifically, we will show that this is true when the selection function $s^{\theta_0}(x)$ is independent of θ_0 once written in terms of the data (i.e., selection can be stated in terms of the data without reference to the null parameter being tested). Formally, we establish monotonicity whenever there exists $\tilde{s}(\cdot, \cdot)$, independent of θ_0, such that

$$s^{\theta_0}(x) = \tilde{s}(t, u) \text{ for all } t, u \text{ with } x = m^{\theta_0}(t, u),$$

Like before, we now consider some fixed $\theta \leq \theta_0$ and let $g_\theta(T(x))$ be the non-increasing function that equals the likelihood ratio $p_\theta(x)/p_{\theta_0}(x)$ on the support. The next lemma allows us to rewrite the selective p-value in a useful way.

Lemma B.15.7. *Letting $E_{r,s} = \{(t, u) : t > r \text{ or } t = r \text{ and } u \leq s\}$. For any t and u such that $m^\theta(t, u) = y$,*

$$\int_0^y s^\theta(x)dx = \int_{E_{t,u}} g_\theta(t)\tilde{s}(t, u)\omega(dt, du).$$

Proof. First we handle the case that $s^\theta(x) = I(x \in B_\theta)$ is an indicator of membership to some set. Then the left-hand side of the equation is the Lebesgue measure of the set $\{x : x \leq y\}$ intersected with B_θ. By Lemma B.15.3 this is the same as the probability under P_θ of the p-value p^θ being at most y and in B_θ. By our choice of t and u, the set difference between $E_{t,u}$ and the set $\{(t, u) : m^{\theta_0}(t, u) \leq y\}$ is measure zero under P_θ. Thus by Lemma B.15.5 and how $\tilde{s}$ is defined, the right-hand side of the equation is also the

probability probability under P_θ of the p-value p^θ being at most y and in B_θ. We again get the result for general selection functions $s^\theta(x)$ by applying the standard machine. □

According the the previous lemma,

$$p^\theta_{sel} = \frac{\int_0^{p^\theta} s^\theta(x)dx}{\int_0^1 s^\theta(x)dx} = \frac{\int_{E_{T(X),U_{aux}}} g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}{\int g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}$$

Hence it would suffice to show that for all r and s,

$$\frac{\int_{E_{r,s}} g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}{\int g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}$$

is monotone in non-decreasing in θ. This is the subject of our next lemma.

Lemma B.15.8. *For any r and s, the quantity*

$$\frac{\int_{E_{r,s}} g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}{\int g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}$$

is monotone non-decreasing in θ.

Proof. The proof strategy is the same as our earlier results. We would like to show that

$$\frac{\int_{E_{r,s}} g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}{\int g_\theta(t)\tilde{s}(t,u)\omega(dt,du)} \leq \frac{\int_{E_{r,s}} g_{\theta_0}(t)\tilde{s}(t,u)\omega(dt,du)}{\int g_{\theta_0}(t)\tilde{s}(t,u)\omega(dt,du)}$$

If the numerator of the left-hand side is zero then the inequality must hold. In the other case we see that

$$\begin{aligned}
\frac{\int g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}{\int_{E_{r,s}} g_\theta(t)\tilde{s}(t,u)\omega(dt,du)} &= 1 + \frac{\int_{E^c_{r,s}} g_\theta(t)\tilde{s}(t,u)\omega(dt,du)}{\int_{E_{r,s}} g_\theta(t)\tilde{s}(t,u)\omega(dt,du)} \\
&\geq 1 + \frac{g_\theta(r)\int_{E^c_{r,s}} \tilde{s}(t,u)\omega(dt,du)}{g_\theta(r)\int_{E_{r,s}} \tilde{s}(t,u)\omega(dt,du)} \\
&= 1 + \frac{\int_{E^c_{r,s}} \tilde{s}(t,u)\omega(dt,du)}{\int_{E_{r,s}} \tilde{s}(t,u)\omega(dt,du)} \\
&= \frac{\int g_{\theta_0}(t)\tilde{s}(t,u)\omega(dt,du)}{\int_{E_{r,s}} g_{\theta_0}(t)\tilde{s}(t,u)\omega(dt,du)}
\end{aligned}$$

where we have noted that if $(t,u) \in E_{r,s}$ then $t \geq r$ and if $(t,u) \in E_{r,s}$ then $t \leq r$, and also that $g_{\theta_0}(t) = 1$ almost everywhere in the measure ω. □

B.16 Selecting multiple p-values for inference

In this appendix, we generalize our selective dominance framework to allow us to select multiple p-values for inference.

Suppose we have k p-values for the nulls $H_{0,i}$ that are selectively dominant given some common Z. We require that these p-values are conditionally independent given Z. Define k binary selection random variables $S_j \in \{0,1\}$, where $S_j = 1$ when p_j is selected. The relationship between p_j, Z, and S_j is governed by the selection functions [1]

$$s_j(x,z) = p(S_j = 1 | p_j = x, Z = z)$$

Furthermore, we demand that the different selections happen independently, i.e.,

$$P(S_1 = 1, \ldots, S_k = 1 \mid p_1, \ldots, p_k, Z) = \prod_{j=1}^{k} P(S_j = 1 \mid p_j, Z) \text{ a.e. under } P \tag{B.22}$$

Define the selective p-values from Theorem 5 $p_{sel,j} = \int_0^{p_j} s_j(x,Z)dx / \int_0^1 s_j(x,Z)dx$. Our goal is to show that these selective p-values remain independent and are also valid p-values after selection:

$$P_{\cap_{i=1}^k H_{0,i}}(p_{sel,1} \leq t_1, \ldots, p_{sel,j} \leq t_1 | S_1 = 1, \ldots, S_j = 1) \leq \prod_{j=1}^{k} t_j \text{ for all } t \in [0,1]^k \tag{B.23}$$

Fix a distribution P in the global null. As this statement is vacuously true when $P(S_1 = 1, \ldots, S_j = 1) = 0$, it suffices to suppose that $P(S_1 = 1, \ldots, S_j = 1) > 0$, in which case we can define the conditional probability measures

$$Q(A) = P(A | S_1 = 1, \ldots, S_k = 1) = \frac{P(A, S_1 = 1, \ldots, S_k = 1)}{P(S_1 = 1, \ldots, S_k = 1)}$$

and

$$Q_j(A) = P(A | S_j = 1) = \frac{P(A, S_j = 1)}{P(S_j = 1)}$$

Note that Q is absolutely continuous with respect to all the Q_j and also P, so any statement which is true a.e. under any Q_j or a.e under P is also true a.e. under Q.

By taking expectations of (B.22) conditional on Z with respect to both sides, we find that the S_j are independent given Z a.e. under P (and thus also a.e. under Q). The following equalities hold a.e. under P.

$$\begin{aligned}
P(S_1 = 1, \ldots, S_k = 1 \mid Z) &= E[P(S_1 = 1, \ldots, S_k = 1 | p_1, \ldots, p_k, Z) \mid Z] \\
&= E\left[\prod_{j=1}^{k} P(S_j = 1 | p_j, Z) \,\middle|\, Z \right] && \text{(Equation (B.22))} \\
&= \prod_{j=1}^{k} E[P(S_j = 1 | p_j, Z)] && (p_j \text{ conditionally independent given } Z) \\
&= \prod_{j=1}^{k} P(S_j = 1 | Z)
\end{aligned}$$

where we have used that the p_j are conditionally independent given Z to move the expectation inside the product.

Now, we recall from Appendix B.19.1 that the selective p-value $p_{sel,j}$ is a continuous function of p_j for

[1] Again, formally $s(p_j, Z) = E[S | p_j, Z]$.

a.e. Z under Q_j, and therefore this is true for a.e. Z under Q. Namely, we have for all $t \in [0,1]^k$ that

$$Q(p_{sel,1} \leq t_1, \ldots, p_{sel,k} \leq t_k | Z) = Q(p_1 \in A_{1,Z,t}, \ldots, p_k \in A_{k,Z,t} | Z) \text{ a.e. under } Q$$

To be specific, for z such that $\int_0^1 s_j(x,z)dx > 0$, the $A_{j,z,t}$ are Borel sets such that $\int_0^y s_j(x,z) / \int_0^1 s_j(x,z) \leq t_j \iff y \in A_z$. Now, the following equalities hold for all $t \in [0,1]^k$ a.e. under Q:

$$\begin{aligned}
&E_Q[I(p_1 \in A_{1,Z,t}) \ldots I(p_k \in A_{k,Z,t}) \mid Z] \\
&= \frac{E_P[I(p_1 \in A_{1,Z,t}) S_1 \ldots I(p_k \in A_{k,Z,t}) S_k \mid Z]}{E_P[S_1 \ldots S_k \mid Z]} && \text{(Lemma B.19.1)} \\
&= \frac{E_P[I(p_1 \in A_{1,Z,t}) \ldots I(p_k \in A_{k,Z,t}) E_P[S_1 \ldots S_k \mid p_1, \ldots, p_k, Z] \mid \mid Z]}{E_P[S_1 \ldots S_k \mid Z]} && \text{(LoTE)} \\
&= \frac{E_P[I(p_1 \in A_{1,Z,t}) E_P[S_1 = 1 | p_k, Z] \ldots I(p_k \in A_{k,Z,t}) E_P[S_k = 1 | p_k, Z] \mid \mid Z]}{E_P[S_1 \ldots S_k \mid Z]} && \text{(Equation (B.22))} \\
&= \prod_{i=1}^k \frac{E_P[I(p_i \in A_{i,Z,t}) E_P[S_i = 1 | p_i, Z] | Z]}{E_P[S_{i,Z} | Z]} && \text{(earlier result)}
\end{aligned}$$

$$\begin{aligned}
&= \prod_{i=1}^k \frac{E_P[I(p_i \in A_{i,Z,t}) S_i | Z]}{E_P[S_i | Z]} && \text{(LoTE)} \\
&= \prod_{i=1}^k E_{Q_i}[I(p_i \in A_{i,Z,t}) | Z] && \text{(Lemma B.19.1)} \\
&= \prod_{i=1}^k E_{Q_i}[I(p_{sel,i} \leq t_i) | Z] && \text{(earlier reasoning)} \\
&\leq \prod_{i=1}^k t_i && \text{(Theorem 5)}
\end{aligned}$$

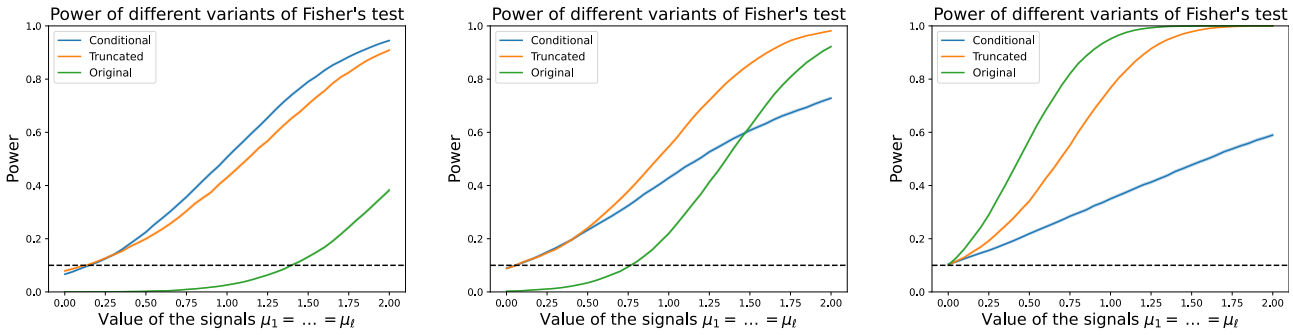
Of course, the fact that

$$P(p_{sel,1} \leq t_1, \ldots, p_{sel,k} \leq t_k | Z, S_1 = 1, \ldots, S_j = 1) \leq \prod_{j=1}^k t_j \text{ for all } t \in [0,1]^k \text{ a.e. under } Q$$

is sufficient to imply (B.23).

B.16.1 Adaptive versions of Fisher's combination test

We can use our generalized framework to come up with variants of Fisher's combination test that are more powerful when some null p-values are conservative (i.e., they have super-uniform distributions). Corollary 10, which is of similar flavor to the conditional inference on winners procedure, gives a conditional version of Fisher's combination test that only uses the bottom k p-values for inference. In Corollary 10 s procedure, k must be pre-specified.



(a) $\ell = 3$ (b) $\ell = 5$ (c) $\ell = 10$

Figure B.3: For $\ell = 3$ (left), $\ell = 5$ (middle), and $\ell = 10$ (right), power of the $k = 3$ conditional, $\tau = 0.5$ truncated, and original Fisher's combination test for data drawn from $N(\mu, I_{10})$ with $\mu_1 = \cdots = \mu_\ell$ varying according to the x-axis and $\mu_{\ell+1} = \cdots = \mu_n = -2$. Power results from an average over $N = 10^4$ trials, error bands (barely visible) denote one standard error, and the level $\alpha = 0.1$ is denoted by the dashed line.

Corollary 10 (Fisher's top-k combination test)**.** *Suppose that p_i are n independent and selectively dominant p-values for the nulls $H_{0,i}$, and let $H_{0,(j)}$ denote the null corresponding to the jth smallest p-value (with ties broken randomly). For some fixed $k \leq n$, rejecting the data-dependent global null $\cap_{j=1}^{k} H_{0,(j)}$ (and therefore also the global null $\cap_{i=1}^{n} H_{0,i}$) when*

$$-2\sum_{j=1}^{k} \log(p_{(j)}/p_{(k+1)}) \geq \mathit{Quantile}(1-\alpha, \chi^2_{2k}) \tag{B.24}$$

controls Type I error at level α conditional on the indices of the smallest k p-values (and therefore also marginally).

Examining (B.24), we see that when the $(k+1)$st p-value provides essentially no evidence against the null (so $p_{(k+1)} \approx 1$), running Corollary 10's test is like running Fisher's test using just the bottom k p-values and ignoring that any selection took place. In this case, our test statistic will be essentially identical to Fisher's original test statistic, but the critical value required for rejection will be much smaller. On the flip-side, if many of the $p_{(j)}$ for $j \leq k$ are not sufficiently smaller than $p_{(k+1)}$, then Corollary 10's test statistic will be small and the test will fail to reject.

Another approach to improving Fisher's combination test is truncation: only use a p-value for inference if it is below some fixed threshold $\tau \in \mathbb{R}$ [Zaykin et al., 2002]. Corollary 11, which is identical to the test from Proposition 2 of Zhao et al. [2019], gives a version of Fisher's truncated combination test that is still valid whenever the p-values are independent and selectively dominant [2].

Corollary 11. *Suppose that p_i are n independent and selectively dominant p-values for the nulls $H_{0,i}$, and fix n thresholds $\tau_i \in [0, 1]$. Letting $j \in J$ denote the random set of indices for which $p_j \leq \tau_j$, rejecting the data-dependent global null $\cap_{j \in J} H_{0,j}$ (and therefore also the global null $\cap_{i=1}^{n} H_{0,i}$) when*

$$-2\sum_{j=1}^{k} \log(p_j/\tau_j) \geq \mathit{Quantile}(1-\alpha, \chi^2_{2|J|})$$

[2] We give a variant that is valid conditional on which p-values are selected. If we just care about rejecting the global null $\cap_{i=1}^{n} H_{0,i}$, we can give a more powerful test via marginalization, as Zaykin et al. [2002] do.

controls Type I error at level α conditional on J (and therefore also marginally).

If some p_j are substantially lower than their truncation point τ_j but most are above it, then Corollary 11 will be powerful. In this case, Corollary 11's test will have a slightly smaller statistic compared to Fisher's original combination test, but a much smaller critical value. Hence, the truncated Fisher test is most powerful when some p-values come from strong alternatives but many come from conservative nulls. As such, Corollary 11 generalizes the truncated Fisher test to the settings where it is most applicable. On the flip-side, if most of the p_j are below τ_j, the truncated test statistic will pay a penalty due to selection, while the critical value required for rejection will remain essentially unchanged compared to Fisher's original test.

To illustrate the benefits and drawbacks of these methods, we display their power alongside that of Fisher's original test for a simple $n = 10$ dimensional Gaussian problem. We sample $X \sim N(\mu, I_n)$ and use the p-values $p_i = 1 - \Phi(X_i)$ try and detect the existence of a positive mean. For $\ell \in \{3, 5, 10\}$, we vary the strength $\mu_1 = \cdots = \mu_\ell > 0$ of our signals and set $\mu_{\ell+1} = \cdots = \mu_n = -2$ to be conservative nulls. We do inference using the bottom $k = 3$ p-values for the conditional version of Fisher's method and set the truncation $\tau = 0.5$ for the truncated version (i.e., we include p_i for which $X_i > 0$). The results are displayed in Figure B.3.

As expected, the new methods outperform Fisher's original method when conservative nulls are present. When $\ell = 3$ and the bottom three p-values are much smaller than the rest, the conditional method does incredibly well. As expected, its performance quickly degrades when $\ell = 5, 10$ and the fourth smallest p-value becomes close to the bottom three. The truncated method is more robust, and still considerably improves power when $\ell = 5$. Unsurprisingly, both selective methods perform worse than Fisher's original method when $\ell = 10$ and every μ_i is a signal (i.e., all p-values are sub-uniform).

B.16.2 Proof of Corollary 10

Suppose we have n independent and selectively dominant p-values p_i for the nulls $H_{0,i}$. We consider a subset $p_{i_1}, \ldots, p_{i_k}$ of them to use for inference. Denoting the remaining as $Z = (p_{i_{k+1}}, \ldots, p_{i_n})$, we have that $p_{i_1}, \ldots, p_{i_k}$ are conditionally independent given Z. We select the p_{i_j} for $j \leq k$ to use for inference when they are smaller than all the p-values in Z, i.e., we use the selection functions $s_j(x, z) = I(x \leq \min_k z_k)$. For these selection functions, the condition Equation (B.22) is obviously satisfied, and Equation (B.23) tells us that, under the global null $\cap_{j=1}^k H_{0,i_j}$, the selective p-values $p_{i_j} / \min_{\ell \notin \{i_1,\ldots,i_k\}} p_\ell$ are independent and valid p-values after selection. In particular, if we denote the $1 - \alpha$ quantile of the χ^2_d distribution as Q_d, a consequence of Equation (B.23) is that

$$P_{\cap_{j=1}^k H_{0,i_j}}(-2 \sum_{j=1}^{k} \log(p_{i_j} / \min_{\ell \notin \{i_1,\ldots,i_k\}} p_\ell) \geq Q_{2k} | S_{i_1} = 1, \ldots, S_{i_k} = 1) \tag{B.25}$$

To show conditional error control, we first note that if $\cap_{j=1}^k H_{0,i_j}$ is false, then trivially

$$P(\text{falsely reject } \cap_{i=1}^k H_{0,(i)} | S_{i_1} = 1, \ldots, S_{i_k} = 1) = 0 \leq \alpha.$$

For the case that $\cap_{j=1}^k H_{0,i_j}$ is true, then the event $S_{i_1} = 1, \ldots, S_{i_k} = 1$ is the same as selecting the $p_{i_1}, \ldots, p_{i_k}$

for inference in (B.25), so

$$\begin{aligned}
&P(\text{falsely reject } \cap_{i=1}^{k} H_{0,(i)} | S_{i_1} = 1, \ldots, S_{i_k} = 1) \\
&= P(-2 \sum_{j=1}^{k} \log(p_{(j)}/p_{(k+1)}) \geq Q_{2k} | S_{i_1} = 1, \ldots, S_{i_k} = 1) \\
&= P(-2 \sum_{j=1}^{k} \log(p_{i_j} / \min_{\ell \notin \{i_1, \ldots, i_k\}} p_\ell) \geq Q_{2k} | S_{i_1} = 1, \ldots, S_{i_k} = 1) \\
&\leq \alpha.
\end{aligned}$$

Now, letting J denote the set of indices of the bottom k p-values, It follows from the definition of conditional expectation that $P(A|J) = P(A|S_j = 1 \text{ for all } j \in J)$ a.e. under P, and the above therefore implies that $P(\text{falsely reject } \cap_{i=1}^{k} H_{0,(i)}|J) \leq \alpha$ a.e. under P. Marginal error control then follows from the law of total expectation.

To see that the same applies for the global null repeat the same argument and replace any mentions of $\cap_{j=1}^{k} H_{0,i_j}$ and $\cap_{j=1}^{k} H_{0,(j)}$ with $\cap_{i=1}^{n} H_{0,i}$.

B.16.3 Proof of Corollary 11

Suppose we have n independent and selectively dominant p-values p_i for the nulls $H_{0,i}$. We select p_j to use for inference when it is smaller than some pre-specified threshold τ_j, i.e., we use the selection functions $s_j(x) = I(x \leq \tau_j)$. Let S_j be the indicator that the jth p-value is selected and J be the random set of indices of the selected p-values. For these selection functions, the condition (B.22) is obviously satisfied, and (B.23) tells us that, under the global null $\cap_{j=1}^{k} H_{0,i_j}$, the selective p-values p_{i_j}/τ_j are independent and valid p-values after selection. In particular, if we denote the $1-\alpha$ quantile of the χ^2_d distribution as Q_d, a consequence of Equation (B.23) is that

$$P_{\cap_{j=1}^{k} H_{0,i_j}}(-2 \sum_{j=1}^{k} \log(p_{i_j}/\tau_j) \geq Q_{2k} | S_{i_1} = 1, \ldots, S_{i_k} = 1)$$

To show conditional error control, we first note that if $\cap_{j=1}^{k} H_{0,i_j}$ is false, then trivially

$$P(\text{falsely reject } \cap_{j \in J}^{k} H_{0,j} | S_{i_1} = 1, \ldots, S_{i_k} = 1) = 0 \leq \alpha.$$

For the case that $\cap_{j=1}^{k} H_{0,i_j}$ is true, then the event $S_{i_1} = 1, \ldots, S_{i_k} = 1$ is the same as selecting the $p_{i_1}, \ldots, p_{i_k}$ for inference in (B.16.3), so

$$\begin{aligned}
&P(\text{falsely reject } \cap_{j \in J}^{k} H_{0,j} | S_{i_1} = 1, \ldots, S_{i_k} = 1) \\
&= P(-2 \sum_{j \in J} \log(p_j/\tau_j) \geq Q_{2k} | S_{i_1} = 1, \ldots, S_{i_k} = 1) \\
&= P(-2 \sum_{j=1}^{k} \log(p_{i_j}/\tau_j) \geq Q_{2k} | S_{i_1} = 1, \ldots, S_{i_k} = 1) \\
&\leq \alpha.
\end{aligned}$$

It follows from the definition of conditional expectation that $P(A|J) = P(A|S_j = 1 \text{ for all } j \in J)$ a.e. under P, and the above therefore implies that $P(\text{falsely reject } \cap_{j\in J} H_{0,j}|J) \leq \alpha$ a.e. under P. Marginal error control then follows from the law of total expectation.

To see that the same applies for the global null repeat the same argument and replace any mentions of $\cap_{j=1}^{k} H_{0,i_j}$ and $\cap_{j\in J} H_{0,j}$ with $\cap_{i=1}^{n} H_{0,i}$.

B.17 Simultaneous approach

For sake of comparison, we consider simultaneous inference approaches that leverage the known covariance structure Σ.

B.17.1 Tukey's test

The typical simultaneous inference approach that is cognizant of the covariance structure Σ is Tukey's honestly significant difference (HSD) test [Tukey, 1951]. Tukey's test, once adapted to our problem, enables us to reject $\cup_{i\in S, j\notin S} H_{ij}^{\delta}$ whenever

$$\min_{i\in S, j\notin S} \frac{X_i - X_j - \delta}{v_{ij}} \geq h_{1-\alpha}, \qquad h_{1-\alpha} = \text{Quantile}\left(1-\alpha, \max_{i\neq j} \frac{|Z_i - Z_j|}{v_{ij}}\right) \text{ with } Z = X - \mu. \tag{B.26}$$

We justify that rejecting $\cup_{i\in S, j\notin S} H_{ij}^{\delta}$ whenever (B.26) is satisfied is an error controlling procedure later. In contrast, our selective test is guaranteed to reject $\cup_{i\in S, j\notin S} H_{ij}^{\delta}$ at least whenever

$$\min_{i\in S, j\notin S} \frac{X_i - X_j - \delta}{v_{ij}} \geq z_{1-\alpha/2}, \qquad z_{1-\alpha/2} = \text{Quantile}\left(1-\alpha, Z\right) \text{ with } Z \sim N(0,1) \tag{B.27}$$

When $n = 2$, these two rejection regions coincide. But as soon as $n > 2$, the quantile $h_{1-\alpha}$ becomes strictly larger than $z_{1-\alpha/2}$, and our test's rejection region becomes a strict superset of Tukey's HSD rejection region. In the case that the X_i are independent, the growth of $h_{1-\alpha}$ is at least on the order of $\sqrt{\log n}$ (we will show this later). The quantile $z_{1-\alpha/2}$ that our procedure uses, however, stays fixed.

B.17.2 A refinement of Tukey's HSD

Based off of the approach in Bofinger [1983, 1985b] and Hsu [1981b], we derive an approach that dominates Tukey's HSD test. These earlier works assume an isotropic covariance. To handle the case with general covariances, we will have to come up with a slightly different procedure than what currently exists in the literature. It is not tractable, but it still helps inform us of the limits of simultaneous inference.

Let $Z = X - \mu$ be a centered version of X, so that $Z \sim N(0, \Sigma)$, and let Π denote the set of permutations over n elements. Defining

$$q_{1-\alpha} = \max_{\pi\in\Pi} \text{Quantile}\left(1-\alpha, \max\left\{\frac{Z_{\pi^{-1}(K)} - Z_{\pi^{-1}(K+1)}}{v_{\pi^{-1}(K),\pi^{-1}(K+1)}}, \max_{\substack{i\geq K+1,\\ j\leq K}} \frac{Z_{\pi^{-1}(i)} - Z_{\pi^{-1}(j)}}{v_{\pi^{-1}(i),\pi^{-1}(j)}}\right\}\right), \tag{B.28}$$

we argue that we can safely draw the inference $\min_{i \in S} \mu_i > \max_{j \notin S} \mu_j + \delta$ whenever

$$\min_{i \in S, j \notin S} \frac{X_i - X_j - \delta}{v_{ij}} \geq q_{1-\alpha}. \tag{B.29}$$

Assume that the entries of μ are sorted (with ties dealt with arbitrarily). Define I' to be the largest index in S (corresponding to the smallest mean) and J' to be the smallest index not in S (corresponding to the largest mean). Define $I \in S$ and $J \in S$ to be the indices that minimize $\frac{X_i - X_j - \delta}{v_{ij}}$ for $i \in S$ and $j \in S$. A false rejection happens exactly when $\frac{X_I - X_J - \delta}{v_{IJ}} > q_{1-\alpha}$ and also $\mu_{I'} - \mu_{J'} - \delta \leq 0$. Note that $I' \geq K$, and $I' = K \implies J' = K+1$ and $I' > K \implies J' \leq K$. With this in mind, we can bound

$$\begin{aligned}
P(\text{false rejection}) &= P\left(\mu_{I'} - \mu_{J'} - \delta \leq 0, \frac{X_I - X_J - \delta}{v_{IJ}} - q_{1-\alpha} \geq 0\right) \\
&\leq P\left(\frac{\mu_{I'} - \mu_{J'} - \delta}{v_{I'J'}} \leq 0, \frac{X_{I'} - X_{J'} - \delta}{v_{I'J'}} - q_{1-\alpha} \geq 0\right) \\
&\leq P\left(\frac{\mu_{I'} - \mu_{J'} - \delta}{v_{I'J'}} \leq \frac{X_{I'} - X_{J'} - \delta}{v_{I'J'}} - q_{1-\alpha}\right) \\
&\leq P\left(q_{1-\alpha} \leq \frac{X_{I'} - \mu_{I'}}{v_{I'J'}} - \frac{X_{J'} - \mu_{J'}}{v_{I'J'}}\right) \\
&\leq P\left(q_{1-\alpha} \leq \frac{Z_{I'} - Z_{J'}}{v_{I'J'}}\right) \\
&\leq P\left(q_{1-\alpha} \leq \max\left\{\frac{Z_K - Z_{K+1}}{v_{K,K+1}}, \max_{\substack{i \geq K+1, \\ j \leq K}} \frac{Z_i - Z_j}{v_{ij}}\right\}\right) \\
&\leq \alpha
\end{aligned}$$

where the last inequality follows from the definition of $q_{1-\alpha}$.

Having proven the validity of this simultaneous method, we make two points. First, we could not use the $1-\alpha$ quantile of

$$\max\left\{\frac{Z_K - Z_{K+1}}{v_{K,K+1}}, \max_{\substack{i \geq K+1, \\ j \leq K}} \frac{Z_i - Z_j}{v_{ij}}\right\}$$

in our procedure because, in the non-isotropic covariance case, computing this quantile requires us to know how to order the samples by their means. Second, $q_{1-\alpha}$ is certainly at most

$$h_{1-\alpha} = \text{Quantile}\left(1-\alpha, \max_{i \neq j} \frac{|Z_i - Z_j|}{v_{ij}}\right), \tag{B.30}$$

which justifies the validity of the Tukey's HSD variant that we proposed earlier.

Now, let us compare our selective test to this simultaneous approach. It is easy to see that $q_{1-\alpha} \geq z_{1-\alpha/2}$, where the inequality is strict exactly when $n > 2$. Therefore, the rejection region (B.27) matches the simultaneous rejection region (B.29) when $n = 2$, and is a strict superset of it when $n > 2$.

In the case that the X_i are independent, we can more explicitly quantify the difference in the two rejection regions. Suppose that n is arbitrarily large and, without loss of generality, that $K < n/2$ (the case that $K \geq n/2$ can be handled with an identical argument). Let π be the permutation such that $\pi^{-1}(K)$ satisfies

$\Sigma_{\pi^{-1}(K),\pi^{-1}(K)} \leq \Sigma_{m,m}$ for all m. Then,

$$
\begin{aligned}
&\max\left\{\frac{Z_{\pi^{-1}(K)} - Z_{\pi^{-1}(K+1)}}{v_{\pi^{-1}(K),\pi^{-1}(K+1)}}, \max_{\substack{i\geq K+1,\\ j\leq K}} \frac{Z_{\pi^{-1}(i)} - Z_{\pi^{-1}(j)}}{v_{\pi^{-1}(i),\pi^{-1}(j)}}\right\} \\
&\geq \max_{i\geq K+1} \frac{Z_{\pi^{-1}(i)} - Z_{\pi^{-1}(K)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} \\
&\geq \max_{i\geq K+1} \frac{Z_{\pi^{-1}(i)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} - \max_{i\geq K+1} \frac{Z_{\pi^{-1}(K)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} \\
&\geq \max_{i\geq K+1} \frac{Z_{\pi^{-1}(i)}}{v_{\pi^{-1}(i),\pi^{-1}(K)}} - \max_{i\geq K+1} \frac{Z_{\pi^{-1}(K)} I(Z_{\pi^{-1}(K)} > 0)}{v_{\pi^{-1}(i),\pi^{-1}(K)}} \\
&= \max_{i\geq K+1} \frac{Z_{\pi^{-1}(i)} I(Z_{\pi^{-1}(i)} > 0)}{v_{\pi^{-1}(i),\pi^{-1}(K)}} - \frac{Z_{\pi^{-1}(K)} I(Z_{\pi^{-1}(K)} > 0)}{\sqrt{2}\sqrt{\Sigma_{\pi^{-1}(K),\pi^{-1}(K)}}} + o_p(1) \\
&\geq \max_{i\geq K+1} \frac{Z_{\pi^{-1}(i)} I(Z_{\pi^{-1}(i)} > 0)}{\sqrt{2}\sqrt{\Sigma_{\pi^{-1}(i),\pi^{-1}(i)}}} + O_p(1) \\
&\geq \frac{1}{\sqrt{2}} \max_{i\geq K+1} \frac{Z_{\pi^{-1}(i)}}{\sqrt{\Sigma_{\pi^{-1}(i),\pi^{-1}(i)}}} + O_p(1) \\
&= O(\sqrt{\log n}) + O_p(1)
\end{aligned}
$$

where the last equality follows from applying standard extreme value theory results regarding the concentration of the maximum of independent standard Gaussians [Haan and Ferreira, 2006] and the fact that $n - K \geq n/2$ per our assumption. As a consequence, $q_{1-\alpha}$ must grow at least on the order of $\sqrt{\log n}$ as well. This implies that, in the independent case, the HSD quantile (B.30) grows at least on the order of $\sqrt{\log n}$ also.

B.18 Confidence lower bound for rank verification

By inverting the test (6.3) for different values of δ (i.e., considering the set of δ for which we fail to reject), we get a confidence region for the gap $\min_{i\in S} \mu_i - \max_{j\neq S} \mu_j$ that is valid conditional on S. It is not immediately clear, however, that this region will result in a confidence lower bound (i.e., there is some smallest δ for which we fail to reject). We provide an argument that it does.

Appendix B.15.3 tells us that, because our original marginal p-values p^{δ}_{ij} in (B.36) come from the UMP test in a MLR family, and because our selection event does not depend on the parameter δ we are testing, the selective p-values $p^{\delta}_{sel,ij}$ from (B.37) are non-decreasing in δ. If, for $i \in S$ and $j \notin S$ we define

$$
\hat{\mu}_{ij} = \begin{cases} \infty, & p^{\delta}_{sel,ij} < \alpha \text{ for all } \delta, \\ -\infty, & p^{\delta}_{sel,ij} > \alpha \text{ for all } \delta, \\ \sup\{\delta : p^{\delta}_{sel,ij} = \alpha\} & \text{otherwise,} \end{cases}
$$

then it is straightforward to argue that Theorem 9's procedure will fail to reject if and only if $\delta > \min_{i\in S, j\notin S} \hat{\mu}_{ij}$. Therefore the inverted confidence region does indeed correspond to a confidence lower bound.

B.19 Proofs

B.19.1 Proof of Theorem 5

First we prove a useful lemma.

Lemma B.19.1. *For a measure P and an event B with $P(B) > 0$ define the conditional measure $Q(E) = P(E|B) = P(E \cap B)/P(B)$. Then, for a random variable X,*

$$Q(A|X) = \frac{P(A, B|X)}{P(B|X)} \text{ a.e. under } Q$$

Proof. First note that $E_P[I(P(B|X) = 0)I(B)] = E_P[I(P(B|X) = 0)E_P[I(B)|X]] = 0$, so $P(B|X) > 0$ a.e. on the set $I(B)$, Thus it must be the case that $P(B|X) > 0$ a.e. under Q. Now we can compute for any Borel C that

$$\begin{aligned}
&E_Q\left[\frac{P(A,B|X)}{P(B|X)} I(P(B|X) > 0) I(X \in C)\right] \\
&= E_P\left[\frac{P(A,B|X)}{P(B|X)} I(P(B|X) > 0) I(X \in C) I(B)\right] / P(B) && \text{(standard machine)} \\
&= E_P\left[\frac{P(A,B|X)}{P(B|X)} I(P(B|X) > 0) I(X \in C) P(B \mid X)\right] / P(B) && \text{(tower property)} \\
&= E_P[P(A, B \mid X) I(P(B|X) > 0) I(X \in C)]/P(B) \\
&= P(A, B, P(B|X) > 0, X \in C)/P(B) && \text{(definition of conditional expectation)} \\
&= Q(A, P(B|X) > 0, X \in C) && \text{(definiton of } Q\text{)} \\
&= E_Q[I(A) I(X \in C)] && (P(B|X) \text{ positive } Q \text{ a.e.}) \\
&= E_Q[Q(A \mid X) I(X \in C)]. && \text{(definition of conditional expectation)}
\end{aligned}$$

Thus $\frac{P(A,B|X)}{P(B|X)} I(P(B|X) > 0)$ satisfies the definition of the conditional expectation of $I(A)$ given X under Q. Since $I(P(B|X) > 0) = 1$ a.e. under Q, we are done.

□

Equipped with this lemma, we can provide a proof.

Fix a probability distribution P in H_0 and let $f_z(x)$ be the conditional PDF of p given $Z = z$. If $P(S = 1) = 0$, then our convention is that $P(p_{sel} \leq \alpha|Z, S = 1)$ and $P(p_{sel} \leq \alpha|S = 1)$ are identically zero, so the result is immediate.

Now consider the non-trivial case, where $P(S = 1) > 0$ and the conditional probability measure $Q(A) = P(A|S = 1) = P(A, S = 1)/P(S = 1)$ is well defined. We first show that (4.5) holds.

We start by showing that $P(S' = 1) > 0$ also. First, we compute

$$E_P[S|Z] = E_P[E_P[S|p, Z]|Z] = E_P[s(p, Z)|Z] = \int_0^1 s(x, Z) f_Z(x) dx$$

and likewise

$$E_P[S'|Z] = E_P[E_P[S|U, Z]|Z] = E_P[s(U, Z)|Z] = \int_0^1 s(x, Z) dx.$$

Whenever $\int_0^1 f_z(x)s(x,z)dx > 0$, it must be the case this z that the function $s(x,z)$ is not zero Lebesgue almost everywhere in x. Therefore $\int_0^1 s(x,z)dx > 0 \implies \int_0^1 f_z(x)s(x,z)dx > 0$. Thus we have both that

$$E_P[S|Z] > 0 \implies E_P[S'|Z] > 0 \text{ everywhere}$$

and the contrapositive

$$E_P[S'|Z] = 0 \implies E_P[S|Z] = 0 \text{ everywhere } P$$

Now, since $P(S=1) = E_P[E_P[S|Z]]$, it cannot be the case that $E_P[S|Z] = 0$ a.e. under P. By the first implication above we must then also have $P(S'=1) = E_P[E_P[S'|Z]] > 0$ as well. Thus we can also define the conditional probability measure $Q'(A) = P(A|S'=1) = P(A, S'=1)/P(S'=1)$. Now, with both these probability measures at our disposal, we prove two lemmas.

Lemma B.19.2. *$E_P[S|Z] > 0$ a.e. under Q and $E_P[S'|Z] > 0$ a.e. under Q'.*

Proof. The fact that $E_P[S|Z] > 0$ a.e. under Q follows from the fact that

$$\begin{aligned} P(E_P[S|Z] = 0, S = 1) &= E_P[I(E_P[S|Z] = 0)S] \\ &= E_P[I(E_P[S|Z] = 0)E_P[S|Z]] \\ &= 0 \end{aligned}$$

The fact that $E_P[S'|Z] > 0$ a.e. under Q' can be shown by an identical proof.

□

Lemma B.19.3. *$Q'(Z \in A) = 0 \implies Q(Z \in A) = 0$, and as a consequence $Q'(Z \in A) = 1 \implies Q(Z \in A) = 1$*

Proof. Since $Q'(Z \in A) = P(Z \in A, S'=1)/P(S'=1)$, we have

$$\begin{aligned} &Q'(Z \in A) = 0 \\ &\iff P(Z \in A, S'=1) = E_P[I(Z \in A)E_P[S'|Z]] = 0 \\ &\iff I(Z \in A)E_P[S'|Z] = 0 \text{ a.e. under } P \\ &\iff P(Z \in A, E_P[S'|Z] > 0) = 0 \end{aligned}$$

By our earlier results, however, $P(Z \in A, E_P[S|Z] > 0) \leq P(Z \in A, E_P[S'|Z] > 0) = 0$. Thus $P(Z \in A, E_P[S|Z] > 0) = 0$, and the same set of equivalences tell us that $Q(Z \in A) = 0$.

The consequence follows from the fact that $Q'(Z \in A) = 1 \implies Q'(Z \in A^c) = 0 \implies Q(Z \in A^c) = 0 \implies Q(Z \in A) = 1$. □

Now, define the function

$$F_{U|z,S'=1}(t) = \frac{\int_0^t s(x,z)dx}{\int_0^1 s(x,z)dx}$$

so that $p_{sel} = F_{U|Z,S'}(p)$. For z such that $\int_0^1 s(x,z)dx > 0$, the above is a monotone non-decreasing and

continuous function of t. For such z, we can define its generalized inverse

$$F^{-1}_{U|z,S'=1}(x) = \inf\{t : F_{U|z,S'=1}(t) > x\}$$

which satisfies $F_{U|z,S'=1}(F^{-1}_{U|z,S'=1}(x)) = x$ and $F_{U|z,S'=1}(t) \leq x \iff t \leq F^{-1}_{U|z,S'=1}(x)$. By Lemma B.19.2 and the fact that $E_P[S'|Z] > E_P[S|Z]$ everywhere, $F_{U|ZS'=1}(t)$ will be a continuous function of x with probability one both under Q and Q'. The following string of equalities and inequalities then hold almost surely under Q:

$$\begin{aligned}
&Q(F_{U|Z,S'=1}(p) \leq t|Z) \\
&= Q(p \leq F^{-1}_{U|Z,S'=1}(t)|Z) && (F_{U|ZS'=1}(t) \text{ continuous a.e. under } Q) \\
&\leq Q'(U \leq F^{-1}_{U|Z,S'=1}(t)|Z) && \text{(selective dominance)} \\
&= t && \text{(Lemma B.19.1)}
\end{aligned}$$

We justify the last equality. First we use Lemma B.19.1 to note that a.e. under Q' we have

$$\begin{aligned}
Q'(U \leq t|Z) &= \frac{P(U \leq t, S' = 1|Z)}{P(S' = 1|Z)} \\
&= \frac{E_P[I(U \leq t)P(S' = 1|U, Z)|Z]}{E_P[P(S' = 1|U, Z)|Z]} \\
&= \frac{\int_0^t s(x, Z)dx}{\int_0^1 s(x, Z)dx} \\
&= F_{U|Z,S'=1}(t)
\end{aligned}$$

Thus a.e. under Q' it holds that

$$Q'(U \leq F^{-1}_{U|Z,S'=1}(t)|Z) = F_{U|Z,S'=1}(F^{-1}_{U|Z,S'=1}(t)) = t$$

Writing the event $Q'(U \leq F^{-1}_{U|Z,S'=1}(t)|Z) = t$ as $Z \in A$ for some Borel set A, Lemma B.19.3 tells us that $Q'(Z \in A) = 1 \implies Q(Z \in A) = 1$, which establishes the claim. Finally, we (4.6) from the law of total expectation.

In the case that $Q(p \leq t|Z) = t$ a.e. under Q, then by the same computations as earlier we can directly compute that $Q(p \leq t|Z) = F_{U|Z,S'=1}(t)$ a.e. under Q, and we can directly compute that $Q(p \leq F^{-1}_{U|Z,S'=1}(t)|Z) = F_{U|Z,S'=1}(F^{-1}_{U|Z,S'=1}(t)) = t$ a.e. under Q, i.e., we get exact equality a.e. under Q as claimed in (4.7).

B.19.2 Proof of Theorem 6

Fix a probability distribution P in H_0 and let $f_z(x)$ be the conditional PDF of p given $Z = z$. If $P(S = 1) = 0$, then our convention is that $P(p \leq \alpha|Z, S = 1) = 0$, so (4.3) holds automatically. Thus we can consider only selection functions $s(x, z)$ for which $P(S = 1) > 0$. We argued in Appendix B.19.1 that when $P(S = 1) > 0$ we also have $P(S' = 1) > 0$. Correspondingly, we define the conditional measures

$Q(A) = P(A|S=1) = P(A, S=1)/P(S=1)$ and $Q'(A) = P(A|S=1) = P(A, S=1)/P(S=1)$.

For now, consider fixed z for which $\int_0^1 s(x,z)f_z(x)dx > 0$. For such z, we argued in Appendix B.19.1 that $\int_0^1 s(x,z)f_z(x)dx > 0$ also. For such z, we argue that the inequality

$$\frac{\int_0^t s(x,z)f_z(x)dx}{\int_0^1 s(x,z)f_z(x)dx} \leq \frac{\int_0^t s(x,z)dx}{\int_0^1 s(x,z)dx}$$

holds for all $t \in [0,1]$. If $\int_0^t s(x,z)f_z(x)dx = 0$ then the inequality trivially holds. Otherwise, we see that

$$\begin{aligned}
\frac{\int_0^1 s(x,z)f_z(x)dx}{\int_0^t s(x,z)f_z(x)dx} &= 1 + \frac{\int_t^1 s(x,z)f_z(x)dx}{\int_0^t s(x,z)f_z(x)dx} \\
&\geq 1 + \frac{f_z(t)\int_t^1 s(x,z)dx}{f_z(t)\int_0^t s(x,z)dx} \\
&= 1 + \frac{\int_t^1 s(x,z)dx}{\int_0^t s(x,z)dx} \\
&= \frac{\int_0^1 s(x,z)dx}{\int_0^t s(x,z)dx},
\end{aligned}$$

which is sufficient to imply the claim.

To finish the argument, we note by Lemma B.19.2 that $E[S|Z] = \int_0^1 s(x,Z)f_Z(x)dx > 0$ a.e. under Q. Applying Lemma B.19.1, we find that a.e. under Q

$$\begin{aligned}
Q(p \leq t|Z) &= \frac{P(p \leq t, S=1|Z)}{P(S=1|Z)} \\
&= \frac{E_P[I(p \leq t)P(S=1|p,Z)|Z]}{E_P[I(p \leq t)P(S=1|p,Z)|Z]} \\
&= \frac{\int_0^t s(x,Z)f_Z(x)dx}{\int_0^1 s(x,Z)f_Z(x)dx}
\end{aligned}$$

and a.e. under Q'

$$\begin{aligned}
Q'(U \leq t|Z) &= \frac{P(U \leq t, S'=1|Z)}{P(S'=1|Z)} \\
&= \frac{E_P[I(U \leq t)P(S'=1|U,Z)|Z]}{E_P[P(S'=1|U,Z)|Z]} \\
&= \frac{\int_0^t s(x,Z)dx}{\int_0^1 s(x,Z)dx}
\end{aligned}$$

But, we can write the event that $Q'(U \leq t|Z) = \frac{\int_0^t s(x,Z)dx}{\int_0^1 s(x,Z)dx}$ as $Z \in A$ for some Borel set A, and Lemma B.19.3 tells us that since $Z \in A$ is measure one under Q' it is also measure one under Q. That is, we have a.e.

under Q that

$$Q'(U \leq t|Z) = \frac{\int_0^t s(x,Z)dx}{\int_0^1 s(x,Z)dx}$$

Combined, we have a.e. under Q that $Q(p \leq t|Z) \leq Q'(U \leq t|Z)$, which is the desired claim.

Now we show the converse part of the statement. Again fix a probability distribution P in H_0, and suppose there is a set B such that $P(Z \in B) > 0$ and for all $z \in B$ the conditional density f_z is everywhere continuous and not non-decreasing. More specifically, all we require is that for all $z \in E$, there are two points $y_1(z) < y_2(z)$ such that $f_z(x)$ is strictly larger in a neighborhood around $y_1(z)$ than in a neighborhood around $y_2(z)$, and these neighborhoods are disjoint. In particular for $\epsilon > 0$ let $N_\epsilon(y) = (y - \epsilon, y + \epsilon)$ be a ball around y. Then we need there to be $y_1(z)$, $y_2(z)$, $\epsilon(z) > 0$, and some $\eta(z) > 0$ such that, for all $w_1 \in N_{\epsilon(z)}(y_1(z))$ and $w_2 \in N_{\epsilon(z)}(y_2(z))$, $w_1 < w_2$ but $f_z(w_2) + \eta(z) < f_z(w_1)$. If f_z is everywhere continuous and not non-decreasing, this will be true. Define $H(z) = \inf\{f_z(w_1) : w_1 \in N_{\epsilon(z)}(y_1(z))\}$ and $L(z) = \sup\{f_z(w_2) : w_2 \in N_{\epsilon(z)}(y_2(z))\}$ so $H(z) > L(z)$. Then consider the selection function

$$s(x,z) = \begin{cases} 1 & \text{if } x \in N_{\epsilon(z)}(y_1(z)) \cup N_{\epsilon(z)}(y_2(z)) \text{ and } z \in B \\ 0 & \text{otherwise .} \end{cases}$$

We show that, for this selection function, selection happens with positive probability:

$$\begin{aligned} P(S=1) &= E_P[E_P[P(S=1|p,Z)|Z]] \\ &= E_P[\int_0^1 s(x,Z) f_Z(x)dx] \\ &\geq E_P[\int_0^1 s(x,Z) f_Z(x)dx I(Z \in B)] \\ &\geq E_P[2\epsilon(Z)\eta(Z) I(Z \in B)] > 0 \end{aligned}$$

Thus the conditional probability measure Q and Q' are well defined. Next we show that $Q(Z \in E) = 1$, so we know that $Z \in E$ a.e. under Q:

$$\begin{aligned} Q(Z \in E) &= \frac{P(Z \in E, S=1)}{P(S=1)} \\ &= \frac{E_P[I(Z \in E) E_P[P(S=1|p,Z)|Z]]}{E_P[E_P[P(S=1|p,Z)|Z]]} \\ &= \frac{E_P[I(Z \in E) \int_0^1 s(x,Z) f_Z(x)dx]}{E_P[\int_0^1 s(x,Z) f_Z(x)dx]} \\ &= \frac{E_P[I(Z \in E) \int_0^1 s(x,Z) f_Z(x)dx]}{E_P[I(Z \in E) \int_0^1 s(x,Z) f_Z(x)dx]} \\ &= 1, \end{aligned}$$

where we have noted that $s(x,z)$ is zero everywhere if $z \notin E$.

Finally, for $z \in E$, define $t(z)$ to be a value such that $t(z) > w_1$ for all $w_1 \in N_{\epsilon(z)}(y_1(z))$ and $t(z) < w_2$

for all $w_2 \in N_{\epsilon(z)}(y_2(z))$. We know from earlier that a.e. under Q

$$Q(p \leq t(Z)|Z) = \frac{\int_0^{t(Z)} s(x,Z) f_Z(x) dx}{\int_0^1 s(x,Z) f_Z(x) dx},$$

where the numerator is at least $2\epsilon(Z)\eta(Z) > 0$, so also a.e. under Q so,

$$\begin{aligned}
\frac{1}{Q(p \leq t(Z)|Z)} &= 1 + \frac{\int_{t(Z)}^1 s(x,Z) f_Z(x) dx}{\int_0^{t(Z)} s(x,Z) f_Z(x) dx} \\
&\leq 1 + \frac{2\epsilon(Z) L(Z)}{2\epsilon(Z) H(Z)} \\
&< 2,
\end{aligned}$$

and

$$Q(p \leq t(Z)|Z) > 1/2.$$

On the other hand, we also have from above that a.e. under Q

$$\begin{aligned}
Q'(U \leq t(Z)|Z) &= \frac{\int_0^{t(Z)} s(x,Z) dx}{\int_0^1 s(x,Z) dx} \\
&= \frac{2\epsilon(Z)}{4\epsilon(Z)} \\
&= 1/2.
\end{aligned}$$

Thus we have shown that, a.e. under Q, there exists some $t \in [0,1]$ such that $Q(p \leq t|Z) > Q'(U \leq t)$, which means that p is not selectively dominant given Z.

B.19.3 Proof of Corollary 3

Suppose we have n independent and selectively dominant p-values p_i for the null hypotheses $H_{0,i}$. Suppose we use p_j to test $H_{0,j}$ only when we observe that p_j is strictly larger than β_n but still the smallest of all the p-values. We can apply Chapter 4's framework with $p = p_j$, $Z = p_{-j}$ and the selection function $s(x,z) = I(\beta_n < x < \min_k z_k)$. It is straightforward to see that Theorem 5's selective p-value is p_{sel} is $(p_j - \beta_n)/(\min_{i \neq j} p_i - \beta_n)$, and Theorem 5 therefore tells us that

$$P_{H_{0,j}}\left(\frac{p_j - \beta_n}{\min_{i \neq j} p_i - \beta_n} \leq \frac{\alpha - \beta}{1 - \beta} \,\middle|\, S = 1 \right) \leq \frac{\alpha - \beta}{1 - \beta}.$$

Re-arranging things we get

$$P_{H_{0,j}}\left(p_j \leq \frac{\alpha - \beta}{1 - \beta} \min_{i \neq j} p_i + \left(1 - \frac{\alpha - \beta}{1 - \beta}\right) \beta_n \,\middle|\, S = 1 \right) \leq \frac{\alpha - \beta}{1 - \beta}. \tag{B.31}$$

Letting W be the index of the smallest p-value (with ties broken randomly), we can now prove the claim

that rejecting $H_{0,W}$ when

$$p_{(1)} \leq \frac{\alpha-\beta}{1-\beta} p_{(2)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_n$$

controls Type I error at level α. Considering any distribution P, let B be the event that a p-value corresponding to a true null is at most β_n. Then $P(B) \leq \beta$ and $P(B^c) \geq 1-\beta$. If $H_{0,j}$ is false, we trivially have have $P(\text{falsely reject } H_{0,W} | W = j, B^c) = 0 \leq (\alpha-\beta)/(1-\beta)$. If $H_{0,j}$ is true, the event $W = j \text{and} B^c$ is the same event as selecting p_j for inference in (B.31), so

$$\begin{aligned}
P(\text{falsely reject } H_{0,W} | W = j, B^c) &= P\left(p_{(1)} \leq \frac{\alpha-\beta}{1-\beta} p_{(2)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_n \;\middle|\; W = j, B^c\right) \\
&= P\left(p_j \leq \frac{\alpha-\beta}{1-\beta} \min_{i \neq j} p_i + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_n \;\middle|\; W = j, B^c\right) \\
&\leq \frac{\alpha-\beta}{1-\beta}.
\end{aligned}$$

Our result then follows from law of total probability:

$$\begin{aligned}
&P(\text{falsely reject } H_{0,W}) \\
&= P(\text{falsely reject } H_{0,W}, B) + P(\text{falsely reject } H_{0,W}, B^c) \\
&\leq P(B) + P(B^c) P(\text{falsely reject } H_{0,W} | B^c) \\
&\leq P(B) + P(B^c) \sum_{j=1}^{n} P(\text{falsely reject } H_{0,W} | W = j, B^c) P(W = j | B^c) \\
&= P(B) + (1 - P(B)) \frac{\alpha-\beta}{1-\beta} \sum_{j=1}^{n} P(W = j | B^c) \\
&\frac{\alpha-\beta}{1-\beta} + \frac{1-\alpha}{1-\beta} P(B) \\
&\leq \frac{\alpha-\beta}{1-\beta} + \frac{1-\alpha}{1-\beta} \beta \\
&= \alpha.
\end{aligned}$$

B.19.4 Proof of Corollary 2 and Corollary 4

It suffices to argue that closing our hybrid global null testing procedure rejects $H_{0,(k)}$ if and only if

$$p_{(j)} \leq \frac{\alpha-\beta}{1-\beta} p_{(j+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-j+1}$$

for all $j \leq k$. We will define $p_{(n+1)} = 1$ so that the right-hand side of the above equals α when $j = n$. Correspondingly, for subsets $I \subseteq [p]$ of size one, we define the hybrid procedure to reject the global null $H_{I,0}$ when the lone p-value is at most α. For subsets I of size strictly more than one, supposing that the smallest p-value in I is the ℓth smallest p-value and the second smallest p-value in I is the mth smallest p-value, the hybrid procedure rejects the global null $H_{I,0}$ when

$$p_{(\ell)} \leq \frac{\alpha-\beta}{1-\beta} p_{(m)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{|I|}.$$

Necessity: For $1 \leq j \leq k$, let I_{n-j+1} be the size $n-j+1$ subset that excludes the $j-1$ smallest p-values (when $j=1$ then $I=[p]$). This subset includes the index of the kth smallest p-value, so we must reject $H_{0,I}$ to reject $H_{0,(k)}$. It rejects exactly when

$$p_{(j)} \leq \frac{\alpha-\beta}{1-\beta} p_{(j+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-j+1}$$

so our conditions are necessary.

Sufficiency: Consider a subset I that contains the index of the kth smallest p-value. If it is size-one, then we reject because

$$p_{(k)} \leq \frac{\alpha-\beta}{1-\beta} p_{(k+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-k+1} \leq \frac{\alpha-\beta}{1-\beta} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta = \alpha.$$

Now suppose that I is size $n-j+1$ for some $j < n$. Its smallest p-value is the ℓth smallest p-value for some $\ell \leq k$ and $\ell \leq j$, and its second smallest p-value is the mth smallest p-value for some $m > \ell$. We reject because

$$\begin{aligned} p_{(\ell)} &\leq \frac{\alpha-\beta}{1-\beta} p_{(\ell+1)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-\ell+1} \\ &\leq \frac{\alpha-\beta}{1-\beta} p_{(m)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{n-j+1} \\ &= \frac{\alpha-\beta}{1-\beta} p_{(m)} + \left(1 - \frac{\alpha-\beta}{1-\beta}\right) \beta_{|I|}. \end{aligned}$$

B.19.5 Proof of Theorem 7

We complete the proof in three steps.

Applying Neymann-Pearson

First we argue that it is sufficient to show that

$$\liminf_{n\to\infty} P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_R^{(n)})} \leq \alpha\right) \geq \liminf_{n\to\infty} P(1-\Phi(X_1^{(n)}) \leq \alpha),$$

Our goal is to show that

$$\liminf_{n\to\infty} \inf_{\phi\in\mathcal{H}_n} P(\phi_n^* \text{ rejects }) - P(\phi_n \text{ rejects }) \geq 0,$$

or equivalently,

$$\liminf_{n\to\infty} P(\phi_n^* \text{ rejects }) - \sup_{\phi\in\mathcal{H}_n} P(\phi_n \text{ rejects }) \geq 0.$$

Defining the point nulls $\widetilde{H}_{0,i}^{(n)} : \mu_i^{(n)} = 0$, let $\widetilde{H}_0 = \cap_{i=1}^n \widetilde{H}_0$ be the global point null that all the means are identically zero, and also let $\widetilde{\mathcal{H}}_n$ be the set of tests that maintain level α Type I error when testing $\widetilde{H}_0$. Then

$$\liminf_{n\to\infty} P(\phi_n^* \text{ rejects }) - \sup_{\phi\in\mathcal{H}_n} P(\phi_n \text{ rejects }) \geq \liminf_{n\to\infty} P(\phi_n^* \text{ rejects }) - \sup_{\phi\in\widetilde{\mathcal{H}}_n} P(\phi_n \text{ rejects }).$$

The term $\sup_{\phi \in \widetilde{\mathcal{H}}_n} P(\phi_n \text{ rejects })$ is the highest power that can be achieved by a level α test trying to distinguish between the point null that all the means are 0 and the alternative that the means are given by $\mu^{(n)}$. The most powerful test is given by Neymann-Pearson [Lehmann et al., 1986, Theorem 3.2.1], and because $\mu_j^{(n)} = 0$ for $j > 1$, it is not hard to verify that it rejects when X_1 is large (it is an identical computation to [Lehmann et al., 1986, Example 3.2.1]). More specifically, this test will reject when $1 - \Phi(X_1^{(n)}) \leq \alpha$. Therefore

$$\sup_{\phi \in \widetilde{\mathcal{H}}_n} P(\phi_n \text{ rejects }) = P(1 - \Phi(X_1^{(n)}) \leq \alpha),$$

and it suffices for us to show

$$\liminf_{n \to \infty} P\left(\frac{1 - \Phi(X_W^{(n)})}{1 - \Phi(X_R^{(n)})} \leq \alpha \right) - P(1 - \Phi(X_1^{(n)}) \leq \alpha) \geq 0,$$

where we have recalled that ϕ_n^* rejects when $(1 - \Phi(X_W^{(n)}))/(1 - \Phi(X_R^{(n)})) \leq \alpha$.

Studying the conditional test

In this part of the proof, which is thanks to Rahul Kanekar, we argue that for any $\epsilon > 0$ there exists a G such that whenever $g \geq G$

$$x_w \geq z_{1-\alpha} + \epsilon \implies \frac{1 - \Phi(x_w)}{1 - \Phi(x_w - g)} \leq \alpha$$

To show this, we define $a = -x_w$ and study

$$\frac{\Phi(a)}{\Phi(a+g)} = \frac{\Phi(x_w)}{1 - \Phi(x_w - g)}$$

By L'Hopital's rule, whenever $g > 0$ we have

$$\lim_{a \to -\infty} \frac{\Phi(a)}{\Phi(a+g)} = \lim_{a \to -\infty} \frac{\phi(a)}{\phi(a+g)} = \lim_{a \to -\infty} \frac{e^{-a^2/2}}{e^{-(a+g)^2/2}} = \lim_{a \to -\infty} e^{ga + \frac{g^2}{2}} = 0 \tag{B.32}$$

Given this, fix $g_1 > 0$ and consider $R \leq z_\alpha - \epsilon$ such that $\Phi(a)/(\Phi(a+g)) \leq \frac{\alpha}{2}$ for all $a \leq R$. Because $\Phi(a)/(\Phi(a+g))$ is decreasing in g for any fixed a, we have for all $g \geq g_1$ that

$$\sup_{a \leq R} \frac{\Phi(a)}{\Phi(a+g)} \leq \sup_{a \leq R} \frac{\Phi(a)}{\Phi(a+g_1)} \leq \frac{\alpha}{2}. \tag{B.33}$$

Furthermore, for all $a \in [R, z_\alpha - \epsilon]$ and any g we have

$$\frac{\Phi(a)}{\Phi(a+g)} \leq \frac{\Phi(z_\alpha - \epsilon)}{\Phi(R+g)}$$

We can take g_2 large enough such that

$$\frac{\Phi(z_\alpha - \epsilon)}{\Phi(R+g_2)} \leq \alpha(1-\epsilon),$$

so that for any $g \geq g_2$

$$\sup_{a \in [R, z_\alpha - \epsilon]} \frac{\Phi(a)}{\Phi(a+g)} \leq \frac{\Phi(z_\alpha - \epsilon)}{\Phi(R+g)} \leq \frac{\Phi(z_\alpha - \epsilon)}{\Phi(R+g_2)} \leq \epsilon(1-\alpha) \tag{B.34}$$

Taking $G = \max(g_1, g_2)$ we see that whenever $g \geq G$,

$$\sup_{a \leq z_\alpha - \epsilon} \frac{\Phi(a)}{\Phi(a+g)} \leq \max(\alpha(1-\epsilon), \frac{\alpha}{2}) < \alpha.$$

Correspondingly, we have that whenever $g \geq G$,

$$\sup_{x_w \geq z_{1-\alpha} + \epsilon} \frac{1 - \Phi(x_w)}{1 - \Phi(x_w - g)} < \alpha,$$

which is sufficient to prove the claim.

Final probabilistic arguments

Now we can put everything together to prove the claim. Fix some $\epsilon > 0$ and take G to be such that whenever $g \geq G$ then

$$x_w \geq z_{1-\alpha} + \epsilon \implies \frac{1 - \Phi(x_w)}{1 - \Phi(x_w - g)} \leq \alpha.$$

$$\begin{aligned}
P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_R^{(n)})} > \alpha\right) &= P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_W^{(n)} - (X_W^{(n)} - X_R^{(n)}))} > \alpha, X_W^{(n)} - X_R^{(n)} \geq G\right) + P(X_W^{(n)} - X_R^{(n)} < G) \\
&= P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_W^{(n)} - (X_W^{(n)} - X_R^{(n)}))} > \alpha, X_W^{(n)} - X_R^{(n)} < G\right) + P(X_W^{(n)} - X_R^{(n)} \geq G)) \\
&\leq P(X_W^{(n)} < z_{1-\alpha} + \epsilon, X_W^{(n)} - X_R^{(n)} < G) + P(X_W^{(n)} - X_R^{(n)} < G) \\
&\leq P(X_W^{(n)} < z_{1-\alpha} + \epsilon) + P(X_W^{(n)} - X_R^{(n)} < G) \\
&\leq P(X_1^{(n)} < z_{1-\alpha} + \epsilon) + P(X_W^{(n)} - X_R^{(n)} < G)
\end{aligned}$$

Because $X_W^{(n)} - X_R^{(n)} \xrightarrow{p} \infty$, we know that $\liminf_{n\to\infty} P(X_W^{(n)} - X_R^{(n)} < G) = 0$, and we can take the limit on both sides to see that

$$\liminf_{n\to\infty} P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_R^{(n)})} > \alpha\right) \leq \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha} + \epsilon)$$

Then, taking the infimum over ϵ gives us that

$$\liminf_{n\to\infty} P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_R^{(n)})} > \alpha\right) \leq \inf_{\epsilon > 0} \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha} + \epsilon)$$

If we knew that

$$\inf_{\epsilon > 0} \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha} + \epsilon) = \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha}),$$

then we would be done:

$$\begin{aligned}
&\liminf_{n\to\infty} P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_R^{(n)})} > \alpha\right) \leq \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha}) \\
&\iff \liminf_{n\to\infty} P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_R^{(n)})} \leq \alpha\right) \geq \liminf_{n\to\infty} P(X_1^{(n)} \geq z_{1-\alpha}) \\
&\iff \liminf_{n\to\infty} P\left(\frac{1-\Phi(X_W^{(n)})}{1-\Phi(X_R^{(n)})} \leq \alpha\right) \geq \liminf_{n\to\infty} P(1-\Phi(X_1^{(n)}) \leq \alpha)
\end{aligned}$$

The remainder of our argument is devoted to showing that this is the case. First we derive that

$$\begin{aligned}
\inf_{\epsilon>0} \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha}+\epsilon) - P(X_1^{(n)} < z_{1-\alpha}) &= \inf_{\epsilon>0} \liminf_{n\to\infty} P(z_{1-\alpha} < X_1^{(n)} < z_{1-\alpha}+\epsilon) \\
&= \inf_{\epsilon>0} \liminf_{n\to\infty} \Phi(z_{1-\alpha}+\epsilon-\mu_1^n) - \Phi(z_{1-\alpha}-\mu_1^{(n)}) \\
&= \inf_{\epsilon>0} \liminf_{n\to\infty} \epsilon\phi(\zeta_n)
\end{aligned}$$

where we have used the mean value theorem and ζ_n is some value between $z_{1-\alpha}+\epsilon-\mu_1^{(n)}$ and $z_{1-\alpha}-\mu_1^{(n)}+\epsilon$. Because $\phi(\cdot)$ is bounded by 1, we must have

$$\inf_{\epsilon>0} \liminf_{n\to\infty} P(X_1^{(n)} = \inf_{\epsilon>0} \epsilon \liminf_{n\to\infty} \phi(\zeta_n) \leq \inf_{\epsilon>0} \epsilon = 0$$

Therefore

$$\begin{aligned}
&\inf_{\epsilon>0} \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha}+\epsilon) - P(X_1^{(n)} < z_{1-\alpha}) = 0 \\
&\implies \inf_{\epsilon>0} \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha}+\epsilon) = \inf_{\epsilon>0} \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha}) = \liminf_{n\to\infty} P(X_1^{(n)} < z_{1-\alpha}).
\end{aligned}$$

and we are done.

B.19.6 Proof of Theorem 8

We will couple our random variables, so $X_i^{(n)} = Z_i + \mu_i^{(n)}$ and consider the sequence of problems given by

$$\mu_i^{(n)} = \begin{cases} \log\log n, & i = 1, \\ \log\log n - 2\sqrt{\log n}, & i = 2, \ldots, \lfloor\sqrt{n}\rfloor \\ -n, & \text{otherwise} \end{cases}$$

We show the result for $\delta = 1$. To show it for $\delta \in (0, 1)$, apply the same argument to the sequence of problems

$$\mu_i^{(n)} = \begin{cases} \Phi^{-1}(z_{1-\alpha}) - \Phi^{-1}(z_{1-\delta}), & i = 1, \\ \Phi^{-1}(z_{1-\alpha}) - \Phi^{-1}(z_{1-\delta}) - 2\sqrt{\log n}, & i = 2, \ldots, \lfloor \sqrt{n} \rfloor \\ -n, & \text{otherwise} \end{cases}$$

Prior to proving the result, we first need to recall facts from extreme value theory facts. Define $M_n = \max_{1 \le i \le n} X_i$ and $m_n = \min_{1 \le i \le n} X_i$. It is known (see [Bovier, 2005, Lemma 1.2.1] that

$$a_n(M_n - b_n) \xrightarrow{d} L, \quad a_n = \sqrt{2 \log n}, \quad b_n = \sqrt{2 \log n} - \frac{\log \log n + \log(4\pi)}{2\sqrt{\log n}}$$

where L is a Gumbel distribution. Essentially, we can write

$$M_n = a_n^{-1} Y_n + b_n$$

where $Y_n = O_p(1)$ and converges in distribution to a Gumbel. Because $m_n \stackrel{d}{=} -M_n$ we can also write

$$m_n = a_n^{-1} W_n - b_n$$

where $W_n = O_p(1)$ and converges in distribution to -1 times a Gumbel. Letting $q_{1-\alpha}$ be the $1 - \alpha$ quantile of a Gumbel distribution, we find that

$$\begin{aligned} z_{1-\alpha_n} &= \text{Quantile}(1 - \alpha, M_n) \\ &= a_n^{-1} \text{Quantile}(1 - \alpha, Y_n) + b_n \\ &= a_n^{-1}(q_{1-\alpha} + o(1)) + b_n \\ &= \sqrt{2 \log n} + O\left(\frac{\log \log n}{\sqrt{\log n}}\right) \end{aligned}$$

Now, we proceed by showing that three sequences of events of interest happen with probability one as $n \to \infty$. Define

$$A_n = X_1^{(n)}$$

$$B_n = \max_{2 \le i \le \lfloor \sqrt{n} \rfloor} X_i^{(n)}$$

$$C_n = \min_{2 \le i \le \lfloor \sqrt{n} \rfloor} X_i^{(n)}$$

$$D_n = \max_{\lfloor \sqrt{n} \rfloor + 1 \le i \le n} X_i^{(n)}$$

- We see that

$$\begin{aligned} P(A_n > B_n) &\ge P(Z_1 + > M_{\lfloor \sqrt{n} \rfloor} - 2\sqrt{\log n}) \\ &= P\left(Z_1 - a_{\lfloor \sqrt{n} \rfloor}^{-1} Y_{\lfloor \sqrt{n} \rfloor} > b_{\lfloor \sqrt{n} \rfloor} - 2\sqrt{\log n}\right) \end{aligned}$$

Because $Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} Y_{\lfloor\sqrt{n}\rfloor} = O_p(1)$ and $b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n} \to -\infty$, we have that

$$\begin{aligned}
&\liminf_{n\to\infty} P(A_n > B_n) \geq \liminf_{n\to\infty} P\left(Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} Y_{\lfloor\sqrt{n}\rfloor} > b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n}\right) = 1 \\
&\implies \lim_{n\to\infty} P(A_n > B_n) = 1
\end{aligned}$$

- We see that

$$\begin{aligned}
P(C_n > D_n) &\geq P(m_{\lfloor\sqrt{n}\rfloor} + \log\log n - 2\sqrt{\log n} > M_n - n) \\
&= P(a^{-1}_{\lfloor\sqrt{n}\rfloor} W_{\lfloor\sqrt{n}\rfloor} - a_n^{-1} Y_n > b_{\lfloor\sqrt{n}\rfloor} + b_n + \log\log n - 2\sqrt{\log n} - n)
\end{aligned}$$

Because $a^{-1}_{\lfloor\sqrt{n}\rfloor} W_{\lfloor\sqrt{n}\rfloor} - a_n^{-1} Y_n = O_p(1)$ and $b_{\lfloor\sqrt{n}\rfloor} + b_n + \log\log n - 2\sqrt{\log n} - n \to -\infty$, we have that

$$\begin{aligned}
&\liminf_{n\to\infty} P(C_n > D_n) \geq \liminf_{n\to\infty} P(m_{\lfloor\sqrt{n}\rfloor} + \log\log n - 2\sqrt{\log n} > M_n - n) \\
&\implies \lim_{n\to\infty} P(C_n > D_n) = 1
\end{aligned}$$

- We see that

$$\begin{aligned}
P(A_n - C_n < 4z_{1-\beta_n^{(n)}/2}) &\geq P(A_n - C_n < 4z_{1-\alpha_n}) \\
&\geq P(Z_1 - m_{\lfloor\sqrt{n}\rfloor} + 2\sqrt{\log n} < 4z_{1-\alpha_n}) \\
&= P(Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} W_{\lfloor\sqrt{n}\rfloor} < 4z_{1-\alpha_n} - b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n})
\end{aligned}$$

Because $Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} W_{\lfloor\sqrt{n}\rfloor} = O_p(1)$ and $4z_{1-\alpha_n} - b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n} \to \infty$, we have that

$$\begin{aligned}
&\liminf_{n\to\infty} P(A_n - C_n < 4z_{1-\beta_n^{(n)}/2}) \geq \liminf_{n\to\infty} P(Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} W_{\lfloor\sqrt{n}\rfloor} < 4z_{1-\alpha_n} - b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n}) = 1 \\
&\implies \lim_{n\to\infty} P(A_n - C_n < 4z_{1-\beta_n^{(n)}}/2) = 1
\end{aligned}$$

Based on the above we know that the sequence of events

$$G_n = \{A_n > B_n > C_n > D_n \text{ and } A_n - C_n < 4z_{1-\beta_n^{(n)}}\}$$

is such that

$$\lim_{n\to\infty} P(G_n) = 1.$$

Formally, this can be shown by writing G^c as the union of finitely many events whose probabilities all tend to zero, and applying a union bound.

We finish the proof by showing three things, that in junction imply the result.

Assumption that $X^{(n)}_{W(n)} - X^{(n)}_{R(n)} \xrightarrow{p} \infty$ holds

Fixing any constant K we see that

$$\begin{aligned}\limsup_{n\to\infty} P(X^{(n)}_{W(n)} - X^{(n)}_{R(n)} \leq K) &\leq \limsup_{n\to\infty} P(A_n - B_n \leq B, G_n) + P(G_n^c) \\ &\leq \limsup_{n\to\infty} P(A_n - B_n \leq K)\end{aligned}$$

So it suffices to show that $\limsup_{n\to\infty} P(A_n - B_n \leq K) = 0$, or equivalently that $\liminf_{n\to\infty} P(A_n - B_n > K) = 1$. We see that

$$\begin{aligned}P(A_n - B_n > K) \geq P(&Z_1 - M_{\lfloor\sqrt{n}\rfloor} + 2\sqrt{\log n} > K) \\ &P(Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} Y_{\lfloor\sqrt{n}\rfloor} > K + b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n})\end{aligned}$$

Because $Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} Y_{\lfloor\sqrt{n}\rfloor} = O_p(1)$, and $b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n}$.

$$\liminf_{n\to\infty} P(A_n - B_n > K) \geq \liminf_{n\to\infty} P(Z_1 - a^{-1}_{\lfloor\sqrt{n}\rfloor} Y_{\lfloor\sqrt{n}\rfloor} > K + b_{\lfloor\sqrt{n}\rfloor} - 2\sqrt{\log n}) = 1$$

LSI is powerless

Note that on the event G_n, the number of observations that are within $4z_{1-\beta^{(n)}_n/2}$ (i.e., the value N from (5.5)) of the winner is at least $\lfloor\sqrt{n}\rfloor - 1$. Leveraging this fact, we can use a union bound to show that LSI is powerless in the limit. Recalling $\gamma^{(n)} = \alpha - \beta^{(n)}$ and $\gamma^{(n)}_k = 1 - (1-\alpha)^{1/k}$, we have

$$\begin{aligned}P(\phi_{LSI} \text{ rejects}) &\leq P(\phi_{LSI} \text{ rejects}, G_n) + P(G_n^c) \\ &\leq P(Z_1 + \log\log n \geq z_{1-\alpha_n}, G_n) + P(Z_1 + \log\log n \geq z_{1-\gamma^{(n)}_{\lfloor\sqrt{n}\rfloor-1}}, G_n) + P(G_n^c) \\ &\leq 2P(Z_1 \geq z_{1-\alpha_{\lfloor\sqrt{n}\rfloor-1}} - \log\log n) + P(G_n^c)\end{aligned}$$

Because $Z_1 = O_p(1)$, $z_{1-\alpha_{\lfloor\sqrt{n}\rfloor-1}} - \log\log n \to \infty$, and also $P(G_n^c) \to 0$, we have

$$\limsup_{n\to\infty} P(\phi_{LSI} \text{ rejects}) \leq \limsup_{n\to\infty} 2P(Z_1 \geq z_{1-\alpha_{\lfloor\sqrt{n}\rfloor-1}} - \log\log n) + P(G_n^c) = 0.$$

Other tests are powerful

The test that just uses the first coordinate attains power one:

$$\begin{aligned}\liminf_{n\to\infty} \inf_{\phi_n \in \mathcal{H}_n} -P(\phi_n \text{rejects}) &= \liminf_{n\to\infty} \sup_{\phi_n \in \mathcal{H}_n} P(\phi_n \text{rejects}) \\ &\geq \liminf_{n\to\infty} P(1 - \Phi(X_1^{(n)}) \leq \alpha) \\ &= \liminf_{n\to\infty} P(Z_1 \geq z_{1-\alpha} - \log\log n) \\ &= 1.\end{aligned}$$

B.19.7 Proof of Proposition 3

First, we show that $P(\mu \in CR_{1-\alpha}^{marg}) = 1 - \alpha$. This is straightforward:

$$\begin{aligned}P(\mu \in CR_{1-\alpha}^{marg}) &= P(X \in R_{1-\alpha}^{marg}(\mu)) \\ &= P(X_W - \mu_W \leq q_{1-\alpha}^{marg}(\mu)) \\ &= 1 - \alpha\end{aligned}$$

Now for the second part of the statement, let $\mathbf{1} \in \mathbb{R}^n$ be the n-dimensional ones vector. Then

$$\begin{aligned}q_{1-\alpha}^{marg}(\eta\mathbf{1}) &= \text{Quantile}(1-\alpha, X_W - \mu_W) \\ &= \text{Quantile}(1-\alpha, \max_{1\leq i\leq n} X_i - \eta) \\ &= \text{Quantile}(1-\alpha, \max_{1\leq i\leq n} (X_i - \eta)) \\ &= \text{Quantile}(1-\alpha, \max_{1\leq i\leq n} Z_i) \\ &= z_{1-\alpha_n}\end{aligned}$$

where the Z_i are i.i.d standard normal random variables. Therefore, if $\eta > X_W - z_{1-\alpha_n}$, we must have $\eta\mathbf{1} \in CR_{1-\alpha}^{marg}$,

$$\eta \geq X_W - z_{1-\alpha_n} \implies (\eta\mathbf{1})_W \geq X_W - q_{1-\alpha}^{marg}(\eta\mathbf{1})$$

Since $1 - \Phi(X_W) > \alpha_n \implies X_W - z_{1-\alpha_n} < 0$, this means we can find $\eta < 0$ such that $\eta\mathbf{1} \in CR_{1-\alpha}^{marg}(X)$, which finishes the proof.

B.19.8 Proof of Theorem 9

We show the equivalent statement that rejecting the the data dependent union null $\cup_{i\in S, j\notin S} H_{ij}^{\delta}$ when

$$\max_{i\in S, j\notin S} \frac{\left[1 - \Phi(D_{ij}^{\delta})\right] - \left[1 - \Phi\left(D_{ij}^{\delta} - \max\limits_{\substack{k\in S,\, \ell\notin S: \\ \rho_{ij,k\ell}<0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0\right)\right]}{\left[1 - \Phi\left(D_{ij}^{\delta} - \min\limits_{\substack{k\in S,\, \ell\notin S: \\ \rho_{ij,k\ell}>0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0\right)\right] - \left[1 - \Phi\left(D_{ij}^{\delta} - \max\limits_{\substack{k\in S,\, \ell\notin S \\ \rho_{ij,k\ell}<0}} \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^0\right)\right]} \leq \alpha. \tag{B.35}$$

ensures that, conditional on S, the probability of a false rejection is at most α.

Without loss of generality, we perform our analysis conditional on the specific event $S = \{1, \dots, K\}$ that the $X_1, \dots, X_K$ are larger than the $X_{K+1}, \dots X_n$. Our strategy will mimic that presented in Hung and Fithian [2019a] and Fithian [2015]. First, for pairs $i \leq K$ and $j > K$, we come up with a test for rejecting H_{ij}^δ that maintains error control conditional on $S = \{1, \dots, K\}$. We reject $\cup_{i \leq K, j > K} H_{ij}^\delta$ when these tests reject for all $i \leq K$, $j > K$. Lemma 4 of Hung and Fithian [2019a], which is adopted from Berger [1982], ensures that in doing so we maintain error control conditional on $S = \{1, \dots, K\}$.

Fix some $i \leq K$ and $j > K$. To design a test for rejecting H_{ij}^δ that maintains error control conditional on $S = \{1, \dots, K\}$, we will use the selective dominance machinery Chapter 4. Normally, to maintain marginal Type I error control, we reject the null H_{ij}^δ using the p-value

$$p_{ij}^\delta = 1 - \Phi(D_{ij}^\delta). \tag{B.36}$$

Defining

$$\epsilon_{ij,k\ell}^\delta = D_{k\ell}^0 - \rho_{ij,k\ell} D_{ij}^\delta$$

it is straightforward to verify that the random vector ϵ_{ij}^δ, which consists of the entries $\epsilon_{ij,k\ell}^\delta$ for pairs $k \leq K$ and $\ell > K$, is independent of D_{ij}^δ and therefore also p_{ij}^δ. Since the p-value (B.36) corresponds to running a one-sided uniformly most powerful (UMP) test in a monotone likelihood ratio family (MLR), Example 3 tells us that it is selectively dominant given $Z = \epsilon_{ij}^\delta$ (where Z is the random vector such that $Z_{k\ell} = \epsilon_{ij,k\ell}^\delta$).

All that remains to do is characterize when we are selecting the p-value p_{ij}^δ to use for inference (i.e., when $S = \{1, \dots, K\}$, what values is this p-value taking?). Theorem 5 then tells how to adjust the p-value to get a selective p-value. Rejecting H_{ij}^δ when this selective p-value is below α maintains error control conditional on $S = \{1, \dots, K\}$.

To do so, we consider $k \leq K$ and $\ell > K$ and rewrite the event

$$\begin{aligned} X_k > X_\ell &\iff D_{k\ell}^0 > 0 \\ &\iff \epsilon_{ij,k\ell}^\delta + \rho_{ij,k\ell} D_{ij}^\delta > 0 \end{aligned}$$

This leads to three cases:

1. If $\rho_{ij,k\ell} > 0$ then $X_k > X_\ell \iff D_{ij}^\delta > -\frac{1}{\rho_{ij,k\ell}} \epsilon_{ij,k\ell}^\delta$,
2. If $\rho_{ij,k\ell} = 0$ then $X_k > X_\ell \iff \epsilon_{ij,k\ell}^\delta > 0$,
3. If $\rho_{ij,k\ell} < 0$ then $X_k > X_\ell \iff D_{ij}^\delta < -\frac{1}{\rho_{ij,k\ell}} \epsilon_{ij,k\ell}^\delta$

Ultimately, we see that

$$\begin{aligned} S = \{1, \dots, K\} &\iff X_k > X_\ell \text{ for all } k \leq K \text{ and } \ell > K \\ &\iff D_{ij}^\delta \in \left(\max_{\substack{k \leq K, \ell > K: \\ \rho_{ij,k\ell} > 0}} -\frac{1}{\rho_{ij,k\ell}} \epsilon_{ij,k\ell}^\delta, \min_{\substack{k \leq K, \ell > K: \\ \rho_{ij,k\ell} < 0}} -\frac{1}{\rho_{ij,k\ell}} \epsilon_{ij,j\ell}^\delta \right) \text{ and } \min_{\substack{k \leq K, \ell > K: \\ \rho_{ij,k\ell} = 0}} \epsilon_{ij,k\ell}^\delta > 0. \end{aligned}$$

Essentially, the selection event $S = \{1, \dots, K\}$ corresponds to selecting p_{ij}^δ (B.36) to use for inference

according to the selection function

$$s(x,z) = \begin{cases} 1 & \text{if } x \in \left(1-\Phi\left(\min_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}} -\frac{1}{\rho_{ij,k\ell}} z_{k\ell}\right), 1-\Phi\left(\max_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}} -\frac{1}{\rho_{ij,k\ell}} z_{k\ell}\right)\right) \text{ and } \min_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}=0}} z_{k\ell} > 0. \\ 0 & \text{otherwise} \end{cases}$$

Theorem 5 then tells us that the selective p-value is

$$p^{\delta}_{sel,ij} = \frac{\left[1-\Phi(D^{\delta}_{ij})\right] - \left[1-\Phi\left(\min_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^{\delta}_{ij,jk}\right)\right]}{\left[1-\Phi\left(\max_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^{\delta}_{ij,jk}\right)\right] - \left[1-\Phi\left(\min_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}} -\frac{1}{\rho_{ij,k\ell}}\epsilon^{\delta}_{ij,jk}\right)\right]}.$$

Recalling the definition of $\epsilon_{ij,k\ell}$, we can rewrite this selective p-value as

$$p^{\delta}_{sel,ij} = \frac{\left[1-\Phi(D^{\delta}_{ij})\right] - \left[1-\Phi\left(D^{\delta}_{ij} - \max_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}} \frac{1}{\rho_{ij,k\ell}} D^{0}_{k\ell}\right)\right]}{\left[1-\Phi\left(D^{\delta}_{ij} - \min_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}} \frac{1}{\rho_{ij,k\ell}} D^{0}_{k\ell}\right)\right] - \left[1-\Phi\left(D^{\delta}_{ij} - \max_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}} \frac{1}{\rho_{ij,k\ell}} D^{0}_{k\ell}\right)\right]}. \tag{B.37}$$

Our proposed procedure of rejecting $\cup^{n}_{i\le K, j>K} H^{\delta}_{ij}$ whenever (B.35) holds corresponds exactly to rejecting when all these selective p-values are at most α, establishing the validity of the procedure.

B.19.9 Proof of Theorem 10

We again without loss of generality perform our analysis conditional on the specific event $S = \{1, \ldots, K\}$. Again fix $i \le K$ and $j > K$. Recall that if $a > b > c$, then $(b-c)/(a-c) \le b/a$. Using this fact, we can bound every selective p-value (B.37):

$$
\begin{aligned}
&\frac{\left[1-\Phi(D_{ij}^{\delta})\right]-\left[1-\Phi\left(D_{ij}^{\delta}-\max\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}}\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)\right]}{\left[1-\Phi\left(D_{ij}^{\delta}-\min\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}}\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)\right]-\left[1-\Phi\left(D_{ij}^{\delta}-\max\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}<0}}\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)\right]}\\
&\le\frac{1-\Phi(D_{ij}^{\delta})}{1-\Phi\left(D_{ij}^{\delta}-\min\limits_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}}\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)}\\
&=\max_{\substack{k\le K,\,\ell>K:\\ \rho_{ij,k\ell}>0}}\frac{1-\Phi(D_{ij}^{\delta})}{1-\Phi\left(D_{ij}^{\delta}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)}.
\end{aligned}
$$

We restrict our attention to pairs $k \le K$ and $\ell > K$ such that $\rho_{ij,k\ell} > 0$ and further bound the term inside the above maximium in two separate cases.

Case one: $\boldsymbol{D_{ij}^{\delta} \le D_{k\ell}^{0}}$. Since $\frac{1}{\rho_{ij,k\ell}} \ge 1$, in this case we have $D_{ij}^{\delta} - \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^{0} \le 0$, so

$$
\frac{1-\Phi(D_{ij}^{\delta})}{1-\Phi(D_{ij}^{\delta}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0})} \le \frac{1-\Phi(D_{ij}^{\delta})}{1-\Phi(0)} \le 2 \max_{i\le K, j>K} p_{ij}^{\delta}
$$

Case two: $\boldsymbol{D_{ij}^{\delta} > D_{k\ell}^{0}}$. To handle this case we first show that

$$
f(x) = \frac{1-\Phi(x)}{1-\Phi(x-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0})}
$$

is a non-increasing function of x. The derivative of the function is

$$
f'(x) = \frac{\phi(x)\phi\left(x-\frac{1}{\rho_{ij,j\ell}}D_{k\ell}^{0}\right)}{\left(1-\Phi\left(x-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)\right)^{2}}\left(\frac{1-\Phi(x)}{\phi(x)}-\frac{1-\Phi\left(x-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)}{\phi\left(x-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)}\right) \le 0
$$

where the inequality follows from the fact that the Mills ratio $(1-\Phi(x))/\phi(x)$ is strictly decreasing [Baricz, 2008, Mills, 1926], and we always have

$$
x > x - \frac{1}{\rho_{ij,k\ell}} D_{k\ell}^{0}
$$

because $\rho_{ij,k\ell} \ge 0$ and $D_{k\ell}^{0} \ge 0$. The non-positiveness of the derivative and the fact that $D_{ij}^{\delta} > D_{k\ell}^{0}$ implies that

$$\begin{aligned}\frac{1-\Phi(D_{ij}^{\delta})}{1-\Phi(D_{ij}^{\delta}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0})} &\leq \frac{1-\Phi(D_{k\ell}^{0})}{1-\Phi(D_{k\ell}^{0}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0})} \\ &\leq \frac{1-\Phi(D_{k\ell}^{0})}{1-\Phi(0)} \\ &\leq 2\max_{i\leq K,j>K} p_{ij}^{0}.\end{aligned}$$

where we have that $D_{k\ell}^{0}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\leq 0$ because $\frac{1}{\rho_{ij,k\ell}}\geq 1$ and $D_{k\ell}^{0}\geq 0$.

To finish the proof of the theorem, we note that if $\delta\geq 0$, then $p_{ij}^{\delta}\geq p_{ij}^{0}$ for all $i\leq K, j\geq K$, so

$$2\max_{i\leq K,j>K} p_{ij}^{0}\leq 2\max_{i\leq K,j>K} p_{ij}^{\delta}$$

To finish the proof of the remark following the theorem, note that if $\delta<0$, then $p_{ij}^{0}\geq p_{ij}^{\delta}$ for all $i\leq K, j\geq K$, so

$$2\max_{i\leq K,j>K} p_{ij}^{\delta}\leq 2\max_{i\leq K,j>K} p_{ij}^{0}$$

B.19.10 Proof of Corollary 5

First we show that because Σ is diagonal, for $i,k\neq j,\ell$ that $\rho_{ij,k\ell}\geq 0$. Defining $c_{ij,k\ell}=\mathrm{Cov}(X_i-X_j, X_k-X_\ell)$, it suffices to show that $c_{ij,k\ell}\geq 0$. We see that

$$c_{ij,k\ell}=\Sigma_{ii}I_{i=k}+\Sigma_{jj}I_{j=\ell}\geq 0$$

This simplifies Theorem 9's procedure considerably. Again, without loss of generality, we perform our analysis conditional on $S=\{1,\ldots,K\}$. Because, $\delta=0$ and $\rho_{ij,k\ell}\geq 0$ for $i,k\leq K$ and $j,\ell>K$, the condition (B.35) is satisfied exactly when

$$\max_{i\leq K,j>K}\frac{1-\Phi(D_{ij}^{0})}{1-\Phi\left(D_{ij}^{0}-\min\limits_{\substack{k\leq K,\,\ell>K:\\ \rho_{ij,k\ell}>0}}\frac{1}{\rho_{ij,k\ell}}D_{k\ell}^{0}\right)}\leq\alpha.$$

Letting I and J be as in the statement of the corollary, we have the the bound

$$\max_{i\leq K,j>K}\frac{1-\Phi(D_{ij})}{1-\Phi\left(\max\limits_{\substack{k\leq K,\,\ell>K:\\ \rho_{ij,k\ell}>0}}D_{ij}-\frac{1}{\rho_{ij,k\ell}}D_{k\ell}\right)}\leq 2(1-\Phi(D_{IJ}^{0})),$$

from proof of Theorem 10. It is also the case that

$$\begin{aligned}\max_{i\leq K, j>K} \frac{1-\Phi(D^0_{ij})}{1-\Phi\left(D^0_{ij} - \min_{\substack{k\leq K,\ell>K:\\ \rho_{ij,k\ell}>0}} \frac{1}{\rho_{ij,k\ell}} D^0_{k\ell}\right)} &\geq \frac{1-\Phi(D^0_{IJ})}{1-\Phi\left(D^0_{IJ} - \min_{\substack{k\leq K,\ell>K:\\ \rho_{IJ,k\ell}>0}} \frac{1}{\rho_{IJ,k\ell}} D^0_{k\ell}\right)} \\ &\geq \frac{1-\Phi(D^0_{IJ})}{1-\Phi(D^0_{IJ} - \frac{1}{\rho_{IJ,IJ}} D^0_{IJ})} \\ &= 2(1-\Phi(D^0_{IJ})),\end{aligned}$$

so the two expressions are in fact equal. This is sufficient to imply the claim.

B.19.11 Proof of Corollary 6

Like in the proof of Corollary 5, we first argue for $i, k \neq j, \ell$ that $c_{ij,k\ell} = \text{Cov}(X_i - X_j, X_k - X_\ell)$ is always non-negative. This follows because

$$c_{ij,kl} = (1-\rho)\sigma^2 I_{i=k} + (1-\rho)\sigma^2 I_{j=\ell} \geq 0$$

Then the proof essentially is then identical to that of Corollary 5. We only need to show that the indices I and J of the observations inside and outside the top K that minimize

$$D^0_{ij} = \frac{X_i - X_j}{v_{ij}} = \frac{X_i - X_j}{\sqrt{2\sigma^2 - 2\rho\sigma^2}}$$

over $i \in S$ and $j \notin S$ are the indices of the Kth largest entry and $(K+1)$st largest entry respectively. But the above is clearly minimized if we take i to be the index of the smallest entry in S (i.e., the Kth largest entry) and j to be the index of the largest entry outside of S (i.e., the $(K+1)$st largest entry), so we are done.

B.19.12 Proof of Corollary 7

Similar to the proof of Corollary 5, we first argue for $i \neq j, \ell$ that $c_{ij,i\ell} = \text{Cov}(X_i - X_j, X_i - X_\ell)$ is always non-negative. We can compute

$$c_{ij,i\ell} = \sigma^2 - \sigma^2\rho^{|i-j|} - \sigma^2\rho^{|i-\ell|} + \sigma^2\rho^{|j-\ell|}$$

When $\rho \geq 0$ we have that $c_{ij,i\ell} \geq \sigma^2(1-2\rho)$. When $\rho \leq 0$, we have that $c_{ij,i\ell} \geq \sigma^2(1-|\rho|-\rho^2)$. Thus, as long as $\rho \in [\frac{1-\sqrt{5}}{2}, \frac{1}{2}]$, we have $c_{ij,i\ell} \geq 0$.

The identical argument to Corollary 5 then implies the result for when $K = 1$. The argument for $K = n-1$ is identical.

B.19.13 Proof of Corollary 8

We consider applying our method to the observation $\hat{\pi}$ while using the covariance $\Sigma = \hat{\pi}/t - \hat{\pi}\hat{\pi}^\top/t$. Without loss of generality, suppose that $\hat{\pi}$ is in sorted order, so $\hat{\pi}_1 \geq \cdots \geq \hat{\pi}_n$. Similar to the proof of

Corollary 5, we first show for $1 \neq j, \ell$ that $c_{1j,1\ell} = \text{Cov}(X_i - X_j, X_k - X_\ell)$ is always non-negative. This follows as

$$\begin{aligned} c_{1j,1\ell} &= \frac{1}{t}(\hat{\pi}_1(1-\hat{\pi}_1) + \hat{\pi}_1\hat{\pi}_j + \hat{\pi}_1\hat{\pi}_\ell - \hat{\pi}_j\hat{\pi}_\ell) \\ &= \frac{1}{t}(\hat{\pi}_1(1-\hat{\pi}_1) + \hat{\pi}_1\hat{\pi}_j + (\hat{\pi}_1 - \hat{\pi}_j)\hat{\pi}_\ell) \\ &\geq 0 \end{aligned}$$

Second, we note that since the $\hat{\pi}$ are in sorted order, the index $J \neq 1$ that minimizes

$$D^0_{1j} = \frac{\sqrt{t}(\hat{\pi}_1 - \hat{\pi}_j)}{\sqrt{\hat{\pi}_1(1-\hat{\pi}_1) + \hat{\pi}_j(1-\hat{\pi}_j) + 2\hat{\pi}_1\hat{\pi}_j}} = \frac{\sqrt{t}(\hat{\pi}_1 - \hat{\pi}_j)}{\sqrt{\hat{\pi}_1 + \hat{\pi}_j - (\hat{\pi}_1 - \hat{\pi}_j)^2}}$$

is clearly $J = 2$. With these facts in places, an identical argument to that in Corollary 5 implies the result.

Bibliography

Mary J Allen and Wendy M Yen. *Introduction to measurement theory*. Waveland Press, 2001.

Theodore Wilbur Anderson, Theodore Wilbur Anderson, Theodore Wilbur Anderson, and Theodore Wilbur Anderson. *An introduction to multivariate statistical analysis*, volume 2. Wiley New York, 1958.

TW Anderson and Herman Rubin. Statistical inference in factor analysis. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, December, 1954, July and August, 1955*, volume 1, page 111. Univ of California Press, 1956.

Isaiah Andrews and Maximilian Kasy. Identification of and correction for publication bias. *American Economic Review*, 109(8):2766–94, August 2019. URL https://www.aeaweb.org/articles?id=10.1257/aer.20180310.

Isaiah Andrews, Toru Kitagawa, and Adam McCloskey. Inference on Winners*. *The Quarterly Journal of Economics*, 139(1):305–358, 09 2023. ISSN 0033-5533. URL https://doi.org/10.1093/qje/qjad043.

E.S. Baker and R.D. DeGroat. A correlation-based subspace tracking algorithm. *IEEE Transactions on Signal Processing*, 46(11):3112–3116, 1998. doi: 10.1109/78.726827.

Laura Balzano, Robert Nowak, and Waheed Bajwa. Column subset selection with missing data. In *NIPS workshop on low-rank methods for large-scale machine learning*, volume 1, 2010.

Árpád Baricz. Mills' ratio: Monotonicity patterns and functional inequalities. *Journal of Mathematical Analysis and Applications*, 340(2):1362–1370, 2008. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2007.09.063. URL https://www.sciencedirect.com/science/article/pii/S0022247X07011730.

Friedrich L Bauer and Charles T Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, 1960.

A Alexander Beaujean. Factor analysis using" r". *Practical Assessment, Research & Evaluation*, 18(4):n4, 2013.

Robert E. Bechhofer. A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with known Variances. *The Annals of Mathematical Statistics*, 25(1):16 – 39, 1954. doi: 10.1214/aoms/1177728845. URL https://doi.org/10.1214/aoms/1177728845.

Yuval Benjamini, Jonathan Taylor, and Rafael A Irizarry. Selection-corrected statistical inference for region detection with high-throughput assays. *Journal of the American Statistical Association*, 114(527):1351–1365, 2019.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24(4):295–300, 1982. ISSN 00401706. URL http://www.jstor.org/stable/1267823.

Michael Beuoy. inpredictable - analytics for sports and gambling. https://www.inpredictable.com/, 2025. Accessed: 2025-05-26.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer, 1997.

Christian H Bischof and Gregorio Quintana-Ortí. Computing rank-revealing qr factorizations of dense matrices. *ACM Transactions on Mathematical Software (TOMS)*, 24(2):226–253, 1998.

Eve Bofinger. Multiple comparisons and selection. *Australian Journal of Statistics*, 25(2):198–207, 1983. doi: https://doi.org/10.1111/j.1467-842X.1983.tb00373.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-842X.1983.tb00373.x.

Eve Bofinger. Multiple comparisons and type iii errors. *Journal of the American Statistical Association*, 80 (390):433–437, 1985a.

Eve Bofinger. Multiple comparisons and type iii errors. *Journal of the American Statistical Association*, 80 (390):433–437, 1985b. ISSN 01621459, 1537274X. URL http://www.jstor.org/stable/2287910.

Giorgos Borboudakis and Ioannis Tsamardinos. Forward-backward selection with early dropping. *J. Mach. Learn. Res.*, 20(1):276–314, jan 2019.

Christos Boutsidis, Jimeng Sun, and Nikos Anerousis. Clustered subset selection and its applications on it service metrics. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, page 599–608. Association for Computing Machinery, 2008.

Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *Proceedings of the Twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, page 968–977. Society for Industrial and Applied Mathematics, 2009.

Christos Boutsidis, Petros Drineas, and Malik Magdon-Ismail. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014. doi: 10.1137/12086755X. URL https://doi.org/10.1137/12086755X.

Anton Bovier. Extreme values of random processes. *Lecture Notes Technische Universität Berlin*, 2005.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Michael J Brusco. A comparison of simulated annealing algorithms for variable selection in principal component analysis and discriminant analysis. *Computational Statistics & Data Analysis*, 77:38–53, 2014.

Jorge Cadima, J Orestes Cerdeira, and Manuel Minhoto. Computational aspects of algorithms for variable selection in the context of principal components. *Computational statistics & data analysis*, 47(2):225–236, 2004.

Tony F. Chan. Rank revealing qr factorizations. *Linear Algebra and its Applications*, 88-89:67–82, 1987. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(87)90103-0. URL https://www.sciencedirect.com/science/article/pii/0024379587901030.

Tony F Chan and Per Christian Hansen. Low-rank revealing qr factorizations. *Numerical Linear Algebra with Applications*, 1(1):33–44, 1994.

Shivkumar Chandrasekaran and Ilse CF Ipsen. On rank-revealing factorisations. *SIAM Journal on Matrix Analysis and Applications*, 15(2):592–622, 1994.

Chein-I Chang, Qian Du, Tzu-Lung Sun, and Mark LG Althouse. A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE transactions on geoscience and remote sensing*, 37(6):2631–2641, 1999.

Yifan Chen, Ethan N. Epperly, Joel A. Tropp, and Robert J. Webber. Randomly pivoted cholesky: Practical approximation of a kernel matrix with few entry evaluations, 2023a. URL https://arxiv.org/abs/2207.06503.

Yiqun Chen, Sean Jewell, and Daniela Witten. More powerful selective inference for the graph fused lasso. *Journal of Computational and Graphical Statistics*, 32(2):577–587, 2023b.

Yiqun T Chen and Daniela M Witten. Selective inference for k-means clustering. *Journal of Machine Learning Research*, 24(152):1–41, 2023.

Shuenn-Ren Cheng and S Panchapakesan. Is the selected population the best?—location and scale parameter cases. *Communications in Statistics—Theory and Methods*, 38(10):1553–1560, 2009.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024a. URL https://arxiv.org/abs/2403.04132.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating llms by human preference, 2024b.

Ali Civril and Malik Magdon-Ismail. Column subset selection via sparse approximation of svd. *Theoretical Computer Science*, 421:1–14, 2012.

Lee Anna Clark and David Watson. Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7:309–319, 1995. URL https://api.semanticscholar.org/CorpusID:15716350.

Open Science Collaboration. Estimating the reproducibility of psychological science. *Science*, 349(6251): aac4716, 2015. URL https://www.science.org/doi/abs/10.1126/science.aac4716.

Anna Costello and Jason W. Osborne. Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. *Practical Assessment, Research and Evaluation*, 10:1–9, 2005. URL https://api.semanticscholar.org/CorpusID:3745568.

Lee Joseph Cronbach. Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334, 1951. URL https://api.semanticscholar.org/CorpusID:13820448.

Chen Dan, Kristoffer Arnsfelt Hansen, He Jiang, Liwei Wang, and Yuchen Zhou. Low rank approximation of binary matrices: Column subset selection and generalizations. *arXiv preprint arXiv:1511.01699*, 2015.

Alexandre d'Aspremont, Laurent Ghaoui, Michael Jordan, and Gert Lanckriet. A direct formulation for sparse pca using semidefinite programming. *Advances in neural information processing systems*, 17, 2004.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

Amit Deshpande and Luis Rademacher. Efficient volume sampling for row/column subset selection. In *2010 ieee 51st annual symposium on foundations of computer science*, pages 329–338. IEEE, 2010.

M Mahamunulu Desu. A selection problem. *The Annals of Mathematical Statistics*, 41(5):1596–1603, 1970.

Ameer Dharamshi, Anna Neufeld, Keshav Motwani, Lucy L Gao, Daniela Witten, and Jacob Bien. Generalized data thinning using sufficient statistics. *Journal of the American Statistical Association*, (just-accepted):1–26, 2024.

John M Digman. Personality structure: Emergence of the five-factor model. *Annual review of psychology*, 41(1):417–440, 1990.

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Relative-error cur matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30(2):844–881, 2008.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Tatjana Eitrich, Achim Kless, Claudia Druska, Wolfgang Meyer, and Johannes Grotendorst. Classification of highly unbalanced cyp450 data of drugs using cost sensitive machine learning techniques. *Journal of chemical information and modeling*, 47(1):92–103, 2007.

Harald Engvik and Sten-Erik Clausen. Norwegian short version of the big five inventory (bfi-20). Technical report, University of Zurich, Department of Informatics, September 2011.

Vaclav Fabian. On multiple decision methods for ranking population means. *The Annals of Mathematical Statistics*, pages 248–254, 1962.

Ahmed K Farahat, Ahmed Elgohary, Ali Ghodsi, and Mohamed S Kamel. Greedy column subset selection for large-scale data sets. *Knowledge and Information Systems*, 45:1–34, 2015.

Elaine Fehrman, Awaz K Muhammad, Evgeny M Mirkes, Vincent Egan, and Alexander N Gorban. The five factor model of personality and evaluation of drug consumption risk. In *Data science: innovative developments in data analysis and clustering*, pages 231–242. Springer, 2017.

Martin Fink, Adam Attarian, and Hien Tran. Subset selection for parameter estimation in an hiv model. *PAMM*, 7, 12 2007. doi: 10.1002/pamm.200700319.

William Fithian. *Topics in Adaptive Inference*. Stanford University, 2015.

William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection, 2017. URL https://arxiv.org/abs/1410.2597.

Leslie V. Foster. Rank and null space calculations using matrix decomposition without column interchanges. *Linear Algebra and its Applications*, 74:47–71, 1986. ISSN 0024-3795. doi: https://doi.org/10.1016/0024-3795(86)90115-1. URL https://www.sciencedirect.com/science/article/pii/0024379586901151.

Laura Freijeiro-González, Manuel Febrero-Bande, and Wenceslao González-Manteiga. A critical review of lasso and its derivatives for variable selection under dependence among covariates. *International Statistical Review*, 90(1):118–145, 2022.

Lucy L Gao, Jacob Bien, and Daniela Witten. Selective inference for hierarchical clustering. *Journal of the American Statistical Association*, 119(545):332–342, 2024.

Jean-Yves Gerlitz and Jürgen Schupp. Zur erhebung der big-five-basierten persoenlichkeitsmerkmale im soep. *DIW Research Notes*, 4(10.1016), 2005.

Charles J Geyer and Glen D Meeden. Fuzzy and randomized confidence intervals and p-values. *Statistical Science*, pages 358–366, 2005.

Jeremy Goldwasser, Will Fithian, and Giles Hooker. Gaussian rank verification, 2025. URL https://arxiv.org/abs/2501.14142.

Gene Golub. Numerical methods for solving linear least squares problems. *Numerische Mathematik*, 7: 206–216, 1965.

Ming Gu and Stanley C Eisenstat. Efficient algorithms for computing a strong rank-revealing qr factorization. *SIAM Journal on Scientific Computing*, 17(4):848–869, 1996.

Ming Gu and L. Miranian. Strong rank revealing cholesky factorization. *ETNA. Electronic Transactions on Numerical Analysis [electronic only]*, 17, 01 2004.

Qian Guo, Wen Wu, DL Massart, C Boucon, and Sjoerd de Jong. Feature selection in principal component analysis of analytical data. *Chemometrics and Intelligent Laboratory Systems*, 61(1-2):123–132, 2002.

Shanti S Gupta. On some multiple decision (selection and ranking) rules. *Technometrics*, 7(2):225–245, 1965.

Shanti Swarup Gupta. *On a Decision Rule for a Problem in Ranking Means*. Ph.d. dissertation, University of North Carolina at Chapel Hill, 1956.

Sam Gutmann. Tests uniformly more powerful than uniformly most powerful monotone tests. *Journal of Statistical Planning and Inference*, 17:279–292, 1987. ISSN 0378-3758. doi: https://doi.

org/10.1016/0378-3758(87)90120-0. URL https://www.sciencedirect.com/science/article/pii/0378375887901200.

Sam Gutmann and Zakhar Maymin. Is the selected population the best? *The Annals of Statistics*, pages 456–461, 1987.

Louis Guttman. The determinacy of factor score matrices with implications for five other basic problems of common-factor theory. *British Journal of Statistical Psychology*, 1955.

Laurens Haan and Ana Ferreira. *Extreme value theory: an introduction*, volume 3. Springer, 2006.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Zihuai He, Linxi Liu, Michael E Belloy, Yann Le Guen, Aaron Sossin, Xiaoxia Liu, Xinran Qi, Shiyang Ma, Prashnna K Gyawali, Tony Wyss-Coray, et al. Ghostknockoff inference empowers identification of putative causal variants in genome-wide association studies. *Nature Communications*, 13(1):7209, 2022.

Jesse Hemerik and Jelle Goeman. Exact testing with random permutations. *TEST*, 27(4):811–825, 2018. URL https://doi.org/10.1007/s11749-017-0571-1.

Yoo Pyo Hong and C-T Pan. Rank-revealing qr factorizations and the singular value decomposition. *Mathematics of Computation*, 58(197):213–232, 1992.

Jason C Hsu. Simultaneous confidence intervals for all distances from the" best". *The Annals of Statistics*, pages 1026–1034, 1981a.

Jason C. Hsu. Simultaneous Confidence Intervals for all Distances from the "Best". *The Annals of Statistics*, 9 (5):1026 – 1034, 1981b. doi: 10.1214/aos/1176345582. URL https://doi.org/10.1214/aos/1176345582.

Jason C. Hsu. Constrained simultaneous confidence intervals for multiple comparisons with the best. *The Annals of Statistics*, 12(3):1136–1144, 1984. ISSN 00905364, 21688966. URL http://www.jstor.org/stable/2240990.

Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47 (2):758 – 782, 2019a. doi: 10.1214/17-AOS1634. URL https://doi.org/10.1214/17-AOS1634.

Kenneth Hung and William Fithian. Rank verification for exponential families. *The Annals of Statistics*, 47 (2):758 – 782, 2019b. URL https://doi.org/10.1214/17-AOS1634.

Kenneth Hung and William Fithian. Statistical methods for replicability assessment. *The Annals of Applied Statistics*, 14(3):1063 – 1087, 2020. URL https://doi.org/10.1214/20-AOAS1336.

Sangwon Hyun, Kevin Z Lin, Max G'Sell, and Ryan J Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77(3):1037–1049, 2021.

Benjamin J Isaac, Axel Coussement, Olivier Gicquel, Philip J Smith, and Alessandro Parente. Reduced-order pca models for chemical reacting flows. *Combustion and flame*, 161(11):2785–2800, 2014.

Oliver P John, Eileen M Donahue, and Robert L Kentle. Big five inventory. *Journal of Personality and Social Psychology*, 1991.

I. T. Jolliffe. *Principal Component Analysis*. Springer New York, NY, 2002.

Ian T Jolliffe, Nickolay T Trendafilov, and Mudassir Uddin. A modified principal component technique based on the lasso. *Journal of computational and Graphical Statistics*, 12(3):531–547, 2003.

Lyle V Jones and John W Tukey. A sensible formulation of the significance test. *Psychological methods*, 5 (4):411, 2000.

joriki. Largest eigenvalue of a positive semi-definite matrix is less than or equal to sum of eigenvalues of its diagonal blocks. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/144963. URL:https://math.stackexchange.com/q/144963 (version: 2012-05-14).

Emtiyaz Khan. Updating inverse of a matrix when a column is added/removed, 2008.

Pavel Krömer, Jan Platoš, Jana Nowaková, and Václav Snášel. Optimal column subset selection for image classification by genetic algorithms. *Annals of Operations Research*, 265(2):205–222, 2018.

Peter M Kruyen, Wilco HM Emons, and Klaas Sijtsma. On the shortcomings of shortened tests: A literature review. *International Journal of Testing*, 13(3):223–248, 2013.

Gary W Ladd and Beckie S Golter. Parents' management of preschooler's peer relations: Is it related to children's social competence? *Developmental Psychology*, 24(1):109, 1988.

Jason D Lee and Jonathan E Taylor. Exact post model selection inference for marginal screening. *Advances in neural information processing systems*, 27, 2014.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907 – 927, 2016a. doi: 10.1214/15-AOS1371. URL https://doi.org/10.1214/15-AOS1371.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907 – 927, 2016b. URL https://doi.org/10.1214/15-AOS1371.

Erich Leo Lehmann, Joseph P Romano, and George Casella. *Testing statistical hypotheses*, volume 3. Springer, 1986.

Lihua Lei and William Fithian. AdaPT: An Interactive Procedure for Multiple Testing with Side Information. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 80(4):649–679, 06 2018. ISSN 1369-7412. URL https://doi.org/10.1111/rssb.12274.

James Leiner, Boyan Duan, Larry Wasserman, and Aaditya Ramdas. Data fission: splitting a single data point. *Journal of the American Statistical Association*, pages 1–12, 2023.

Joshua R Loftus and Jonathan E Taylor. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*, 2015.

Timothy Shin Heng Mak, Robert Milan Porsch, Shing Wan Choi, Xueya Zhou, and Pak Chung Sham. Polygenic scores via penalized regression on summary statistics. *Genetic epidemiology*, 41(6):469–480, 2017.

Ruth Marcus, Peritz Eric, and K Ruben Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

Kantilal Varichand Mardia, John T Kent, and John M Bibby. *Multivariate analysis*. Academic Press, 1979.

Jelena Markovic, Jonathan Taylor, and Jeremy Taylor. Inference after black box selection. *arXiv preprint arXiv:1901.09973*, 2019.

Mahdokht Masaeli, Yan Yan, Ying Cui, Glenn Fung, and Jennifer G Dy. Convex principal feature selection. In *Proceedings of the 2010 SIAM international conference on data mining*, pages 619–628. SIAM, 2010.

Zakhar Maymin and Sam Gutmann. Testing retrospective hypotheses. *Canadian Journal of Statistics*, 20 (3):335–345, 1992.

George P McCabe. Principal variables. *Technometrics*, 26(2):137–144, 1984.

Daniel McNeish. Thanks coefficient alpha, we'll take it from here. *Psychological methods*, 23(3):412, 2018.

John P. Mills. Table of the ratio: Area to bounding ordinate, for any portion of normal curve. *Biometrika*, 18(3/4):395–400, 1926. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2331957.

L. Mirsky. A trace inequality of john von neumann. *Monatshefte für Mathematik*, 79(4):303–306, 1975. URL https://doi.org/10.1007/BF01647331.

Alexander Molotaliev. Engine time to failure dataset, 2024. URL https://www.kaggle.com/datasets/m0ntecarl0/engine-time-to-failure.

Anna Neufeld, Ameer Dharamshi, Lucy L Gao, and Daniela Witten. Data thinning for convolution-closed distributions. *Journal of Machine Learning Research*, 25(57):1–35, 2024.

Jana Nowaková, Pavel Krömer, Jan Platoš, and Václav Snášel. Preprocessing covid-19 radiographic images by evolutionary column subset selection. In Leonard Barolli, Kin Fun Li, and Hiroyoshi Miwa, editors, *Advances in Intelligent Networking and Collaborative Systems*, pages 425–436. Springer International Publishing, 2021. ISBN 978-3-030-57796-4.

Masashi Okamoto and Mitsuyo Kanazawa. Minimization of eigenvalues of a matrix and optimality of principal components. *The Annals of Mathematical Statistics*, 39(3):859–863, 1968.

Bruno Ordozgoiti, Sandra Gómez Canaval, and Alberto Mozo. Iterative column subset selection. *Knowledge and Information Systems*, 54:65–94, 2018.

Bruno Ordozgoiti, Alberto Mozo, and Jesús García López de Lacalle. Regularized greedy column subset selection. *Information Sciences*, 486:393–418, 2019.

Art Owen. *Empirical Likelihood*. Chapman and Hall/CRC, 2001.

C-T Pan. On the existence and computation of rank-revealing lu factorizations. *Linear Algebra and its Applications*, 316(1-3):199–222, 2000.

Ching-Tsuan Pan and Ping Tak Peter Tang. Bounds on singular values revealed by qr factorizations. *BIT Numerical Mathematics*, 39:740–756, 1999.

Snigdha Panigrahi, Kevin Fry, and Jonathan Taylor. Exact selective inference with randomization. *Biometrika*, page asae019, 2024.

Algebraic Pavel. Spectrum of schur complement. Mathematics Stack Exchange. URL https://math.stackexchange.com/q/1994997. URL:https://math.stackexchange.com/q/1994997 (version: 2022-11-15).

Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901.

Ronan Perry, Zichun Xu, Olivia McGough, and Daniela Witten. Infer-and-widen versus split-and-condition: two tales of selective inference. *arXiv preprint arXiv:2408.06323*, 2024.

K. B. Petersen and M. S. Pedersen. The matrix cookbook, 2008.

Beatrice Rammstedt and Oliver P John. Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german. *Journal of research in Personality*, 41(1):203–212, 2007.

Joseph P Romano, Azeem M Shaikh, and Michael Wolf. A practical two-step method for testing moment inequalities. *Econometrica*, 82(5):1979–2002, 2014.

Hans Schneeweiss and Harald Mathes. Factor analysis and principal components. *Journal of multivariate analysis*, 55(1):105–124, 1995.

Peter H Schönemann and James H Steiger. Regression component analysis. *British Journal of Mathematical and Statistical Psychology*, 29(2):175–189, 1976.

Peter H Schönemann and James H Steiger. On the validity of indeterminate factor scores. *Bulletin of the Psychonomic Society*, 12(4):287–290, 1978.

Yaroslav Shitov. Column subset selection is np-complete. *Linear Algebra and its Applications*, 610:52–58, 2021.

Uri Simonsohn, Leif D. Nelson, and Joseph P. Simmons. P-curve: A key to the file drawer. *Cognitive Linguistics: Cognition*, 2013. URL https://api.semanticscholar.org/CorpusID:8505270.

Anav Sood. Selective inference is easier with p-values. *arXiv preprint arXiv:2411.13764*, 2024.

Anav Sood. Powerful rank verification for multivariate gaussian data with any covariance structure. *arXiv preprint arXiv:2503.01065*, 2025.

Anav Sood and Trevor Hastie. A statistical view of column subset selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkaf023, 05 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf023. URL https://doi.org/10.1093/jrsssb/qkaf023.

Gunnar Stefansson, Woo-Chul Kim, and Jason C Hsu. On confidence sets in multiple comparisons. *Statistical decision theory and related topics IV*, pages 89–104, 1988.

James H Steiger. The relationship between external variables and common factors. *Psychometrika*, 44(1): 93–97, 1979.

James H Steiger. Factor analysis in the 1980's and the 1990's: Some old debates and some new developments. *Trends and perspectives in empirical social research*, pages 201–224, 1994.

G. W. Stewart. The qlp approximation to the singular value decomposition. *SIAM Journal on Scientific Computing*, 20(4):1336–1348, 1999a. doi: 10.1137/S1064827597319519. URL https://doi.org/10.1137/S1064827597319519.

Gilbert W Stewart. *Matrix algorithms: volume 1: basic decompositions*. SIAM, 1998.

Gilbert W Stewart. Four algorithms for the the efficient computation of truncated pivoted qr approximations to a sparse matrix. *Numerische Mathematik*, 83:313–323, 1999b.

Martin Tobias Strauch. *Column subset selection with applications to neuroimaging data*. PhD thesis, Universität Konstanz, Konstanz, 2014.

Jonathan Taylor and Robert Tibshirani. Post-selection inference for-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61, 2018.

Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*, 354, 2014.

Jonathan E. Taylor, Joshua R. Loftus, and Ryan J. Tibshirani. Inference in adaptive regression via the Kac–Rice formula. *The Annals of Statistics*, 44(2):743 – 770, 2016. URL https://doi.org/10.1214/15-AOS1386.

Xiaoying Tian and Jonathan Taylor. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

Xiaoying Tian, Joshua R Loftus, and Jonathan E Taylor. Selective inference with unknown variance via the square-root lasso. *Biometrika*, 105(4):755–768, 2018.

Ryan J Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016.

Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

Richa Tripathi and Amit Reza. A subset selection based approach to structural reducibility of complex networks. *Physica A: Statistical Mechanics and its Applications*, 540:123214, 2020.

Joel A Tropp. Column subset selection, matrix factorization, and eigenvalue optimization. In *Proceedings of the twentieth annual ACM-SIAM symposium on Discrete algorithms*, pages 978–986. SIAM, 2009.

Lana Tucaković and Boban Nedeljković. From the bfi-44 to bfi-20: Psychometric properties of the short form of the big five inventory. *Psychological reports*, page 00332941231161754, 2023.

John W Tukey. Quick and dirty methods in statistics. part ii. simple analyses for standard designs. *American Society for Quality Control*, pages 189–197, 1951.

A. W. van der Vaart. *Stochastic Convergence*, page 5–24. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 1998. doi: 10.1017/CBO9780511802256.003.

Wayne F Velicer and Douglas N Jackson. Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate behavioral research*, 25(1):1–28, 1990.

Valdiney Veloso Gouveia, RDC Rodrigues Araújo, Isabel Cristina Vasconcelos de Oliveira, Marina Pereira Gonçalves, Taciano Milfont, Gabriel Lins de Holanda Coelho, Walberto Santos, Emerson Diógenes de Medeiros, Ana Karla Silva Soares, Renan Pereira Monteiro, et al. A short version of the big five inventory (bfi-20): Evidence on construct validity. *Interamerican Journal of Psychology*, 2021.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Chi Wang, Maoguo Gong, Mingyang Zhang, and Yongqiang Chan. Unsupervised hyperspectral image band selection via column subset selection. *IEEE Geoscience and Remote Sensing Letters*, 12(7):1411–1415, 2015.

Yining Wang and Aarti Singh. Provably correct algorithms for matrix column subset selection with selectively sampled data. *The Journal of Machine Learning Research*, 18(1):5699–5740, 2017.

Hua-Liang Wei and Stephen A Billings. Feature subset selection and ranking for data dimensionality reduction. *IEEE transactions on pattern analysis and machine intelligence*, 29(1):162–166, 2006.

Samuel S Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

Edwin B. Wilson. On hierarchical correlation systems. *Proceedings of the National Academy of Sciences of the United States of America*, 14(3):283–291, 1928. ISSN 00278424. URL http://www.jstor.org/stable/85427.

Daniela M Witten, Robert Tibshirani, and Trevor Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, 2009.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67, 2006.

Dmitri V Zaykin, Lev A Zhivotovsky, Peter H Westfall, and Bruce S Weir. Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 22(2):170–185, 2002.

Guangjian Zhang, Ge Jiang, Minami Hattori, and Lauren Trichtinger. *EFAutilities: Utility Functions for Exploratory Factor Analysis*, 2022. URL https://CRAN.R-project.org/package=EFAutilities. R package version 2.1.2.

Qingyuan Zhao, Dylan S Small, and Weijie Su. Multiple testing when many p-values are uniformly conservative, with application to testing qualitative interaction in educational interventions. *Journal of the American Statistical Association*, 114(527):1291–1304, 2019.

Hui Zou, Trevor Hastie, and Robert Tibshirani. Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2):265–286, 2006.

Tijana Zrnic and William Fithian. Locally simultaneous inference. *The Annals of Statistics*, 52(3):1227–1253, 2024.