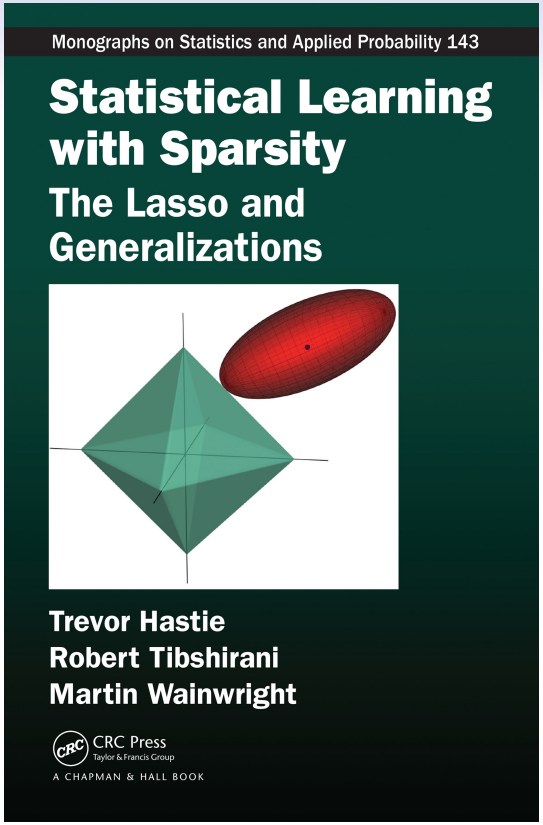
Statistical Learning with Sparsity

2019 Wald Lecture I

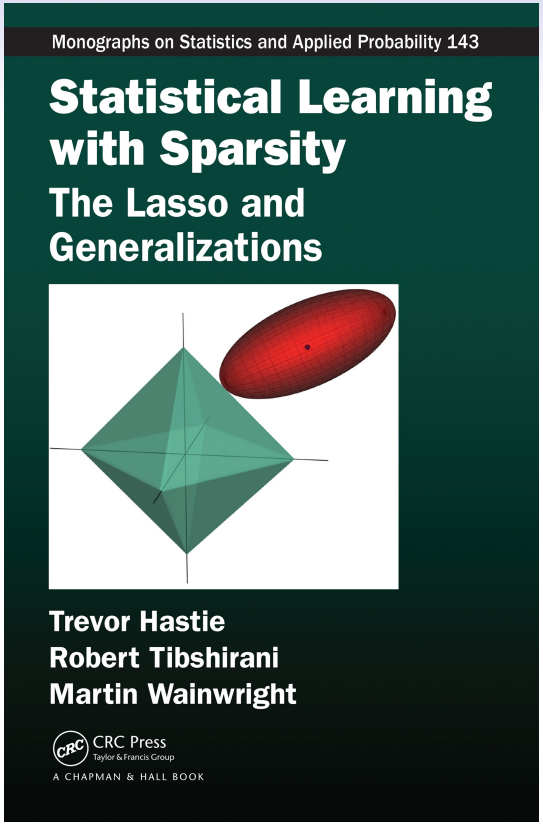
Trevor Hastie
Department of Statistics
Department of Biomedical Data Science
Stanford University




Monographs on Statistics and Applied Probability 143
Statistical Learning with Sparsity
The Lasso and Generalizations
Trevor Hastie
Robert Tibshirani
Martin Wainwright
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

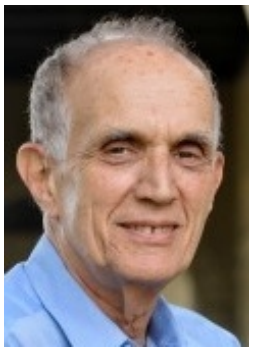




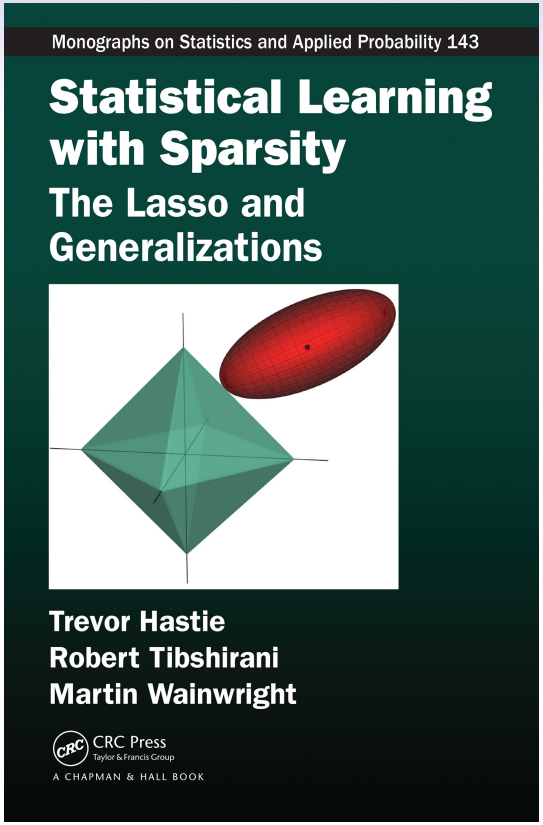

Monographs on Statistics and Applied Probability 143
Statistical Learning with Sparsity
The Lasso and Generalizations
Trevor Hastie
Robert Tibshirani
Martin Wainwright
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK





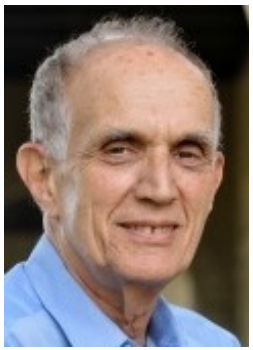





Monographs on Statistics and Applied Probability 143
Statistical Learning with Sparsity
The Lasso and Generalizations
Trevor Hastie
Robert Tibshirani
Martin Wainwright
CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK















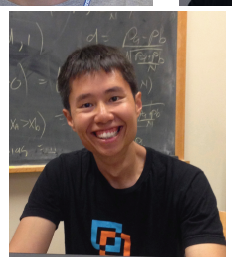


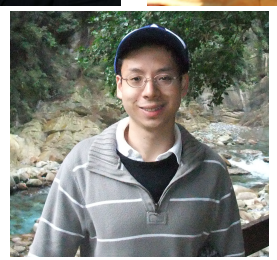



STANFORD

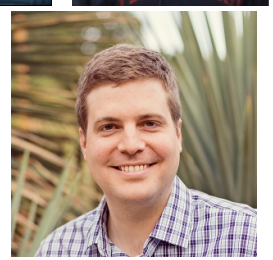














Confessions

Confessions

- I like simple ideas.

Confessions

- I like simple ideas.
  Good ideas are often simple, so sometimes I'm in luck!

Confessions

- I like simple ideas.
  Good ideas are often simple, so sometimes I'm in luck!
  Simple methods often work as well as complex ones — we are all in luck!

Confessions

- I like simple ideas.
  Good ideas are often simple, so sometimes I'm in luck!
  Simple methods often work as well as complex ones — we are all in luck!
- I learned a few good tricks early on, then spent the rest of my career finding places to use them!

Confessions

- I like simple ideas.
  Good ideas are often simple, so sometimes I'm in luck!
  Simple methods often work as well as complex ones — we are all in luck!
- I learned a few good tricks early on, then spent the rest of my career finding places to use them!
- Good collaborators are essential, especially ones armed with whips!

Confessions

- I like simple ideas.
  Good ideas are often simple, so sometimes I'm in luck!
  Simple methods often work as well as complex ones — we are all in luck!
- I learned a few good tricks early on, then spent the rest of my career finding places to use them!
- Good collaborators are essential, especially ones armed with whips!
- I love developing algorithms and software. For a technique to get used and take off, good software is essential.

Linear Models for Wide Data

As datasets grow *wide*—i.e. many more features than samples—the linear model has regained favor as the tool of choice.

Document classification: bag-of-words easily leads to $p = 20K$ features and $N = 5K$ document samples. Much more if bigrams, trigrams etc, or documents from Facebook, Google, Twitter!

Genomics, microarray studies: $p = 40K$ genes are measured for each of $N = 300$ subjects.

Genome-wide association studies: $p = 1$–$2M$ SNPs measured for $N = 2$–$500K$ subjects.

Web activity: Millions of possible webpages visited, search terms used.

Linear Models for Wide Data

As datasets grow *wide*—i.e. many more features than samples—the linear model has regained favor as the tool of choice.

Document classification: bag-of-words easily leads to $p = 20K$ features and $N = 5K$ document samples. Much more if bigrams, trigrams etc, or documents from Facebook, Google, Twitter!

Genomics, microarray studies: $p = 40K$ genes are measured for each of $N = 300$ subjects.

Genome-wide association studies: $p = 1$–$2M$ SNPs measured for $N = 2$–$500K$ subjects.

Web activity: Millions of possible webpages visited, search terms used.

In examples like these we tend to use linear models — e.g. linear regression, logistic regression, Cox model. Since $p \gg N$, we cannot fit these models using standard approaches.

Regularization

$$\eta(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

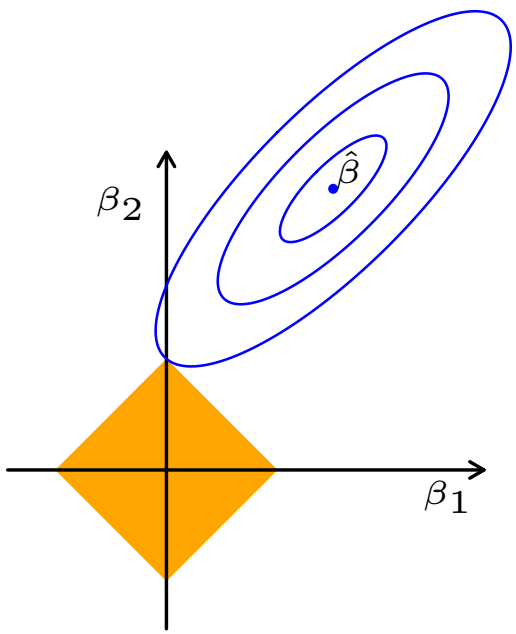
We cannot fit linear models with $p > N$ without some constraints. Common approaches are

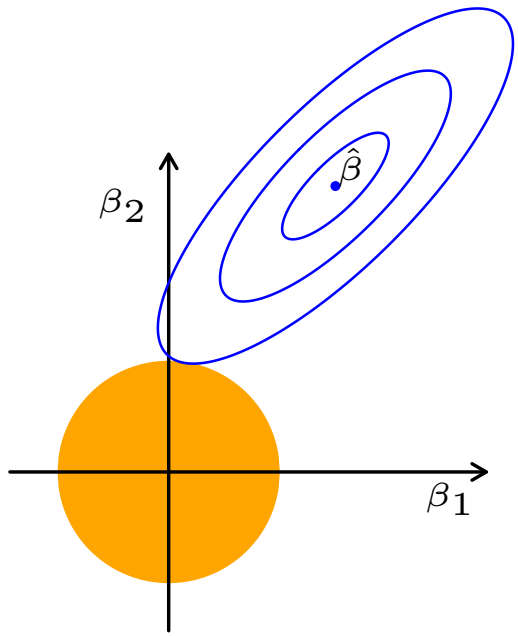
Forward stepwise adds variables one at a time and stops when overfitting is detected. Regained popularity for $p \gg N$, since it is the most feasible method among it's subset cousins: best-subset, backward stepwise.

Ridge regression fits the model subject to the constraint $\sum_{j=1}^{p} \beta_j^2 \leq t$. Shrinks coefficients toward zero, and hence controls variance. Allows linear models of arbitrary size p to be fit.

Lasso regression (Tibshirani, 1996) fits the model subject to the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$.

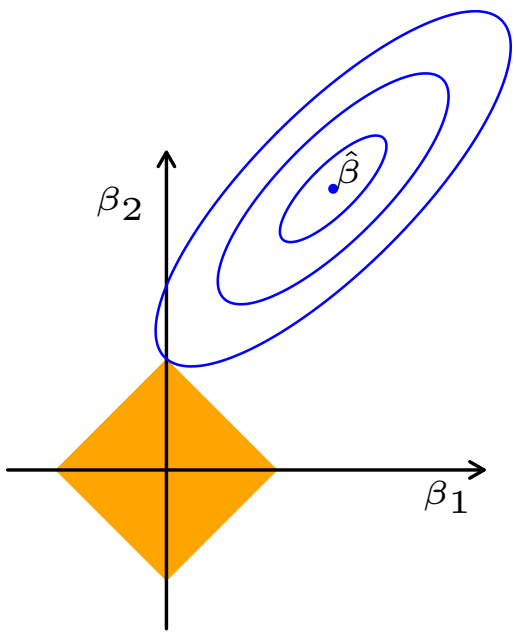
Lasso does variable selection and shrinkage, while ridge only shrinks (a lot to control variance!)

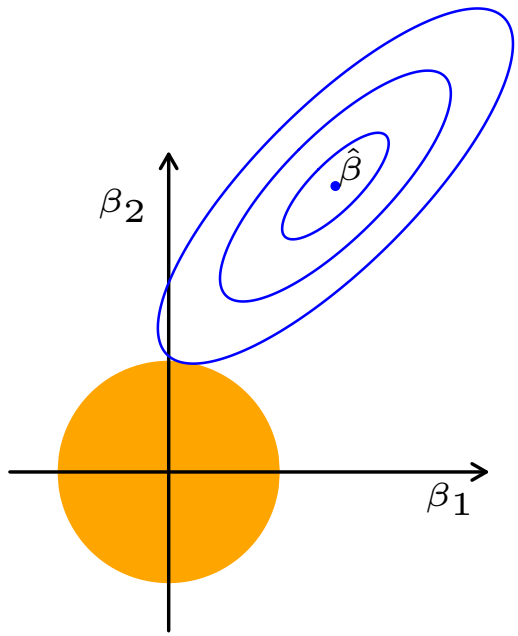




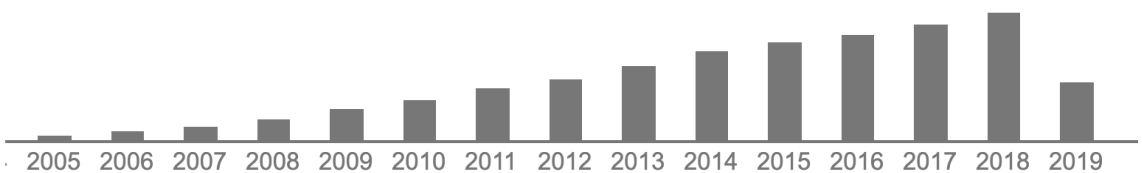
Lasso regression (Tibshirani, 1996) fits the model subject to the constraint $\sum_{j=1}^{p} |\beta_j| \leq t$.

Lasso does variable selection and shrinkage, while ridge only shrinks (a lot to control variance!)

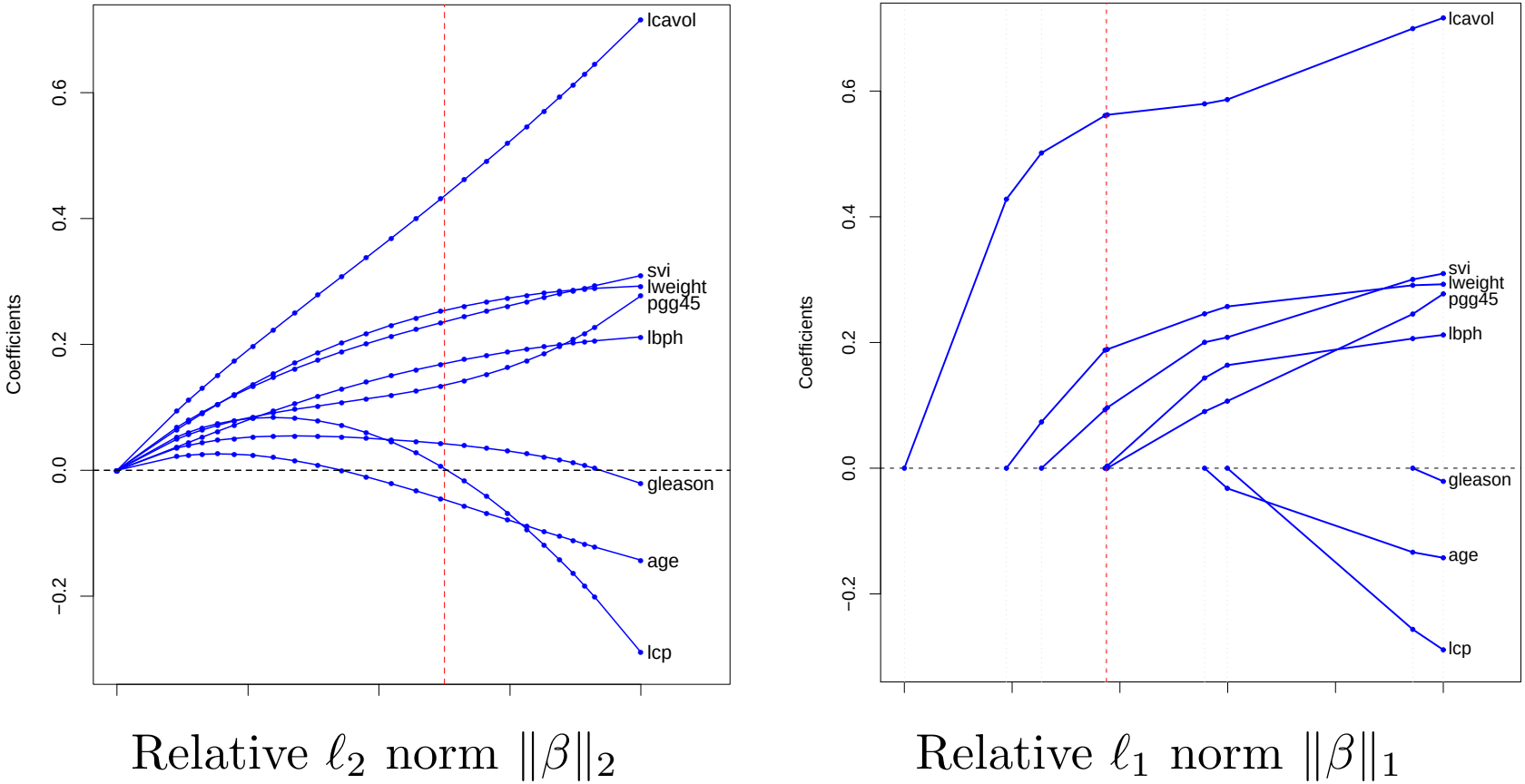




28867 citations (Google) by July 2019!



Ridge versus Lasso



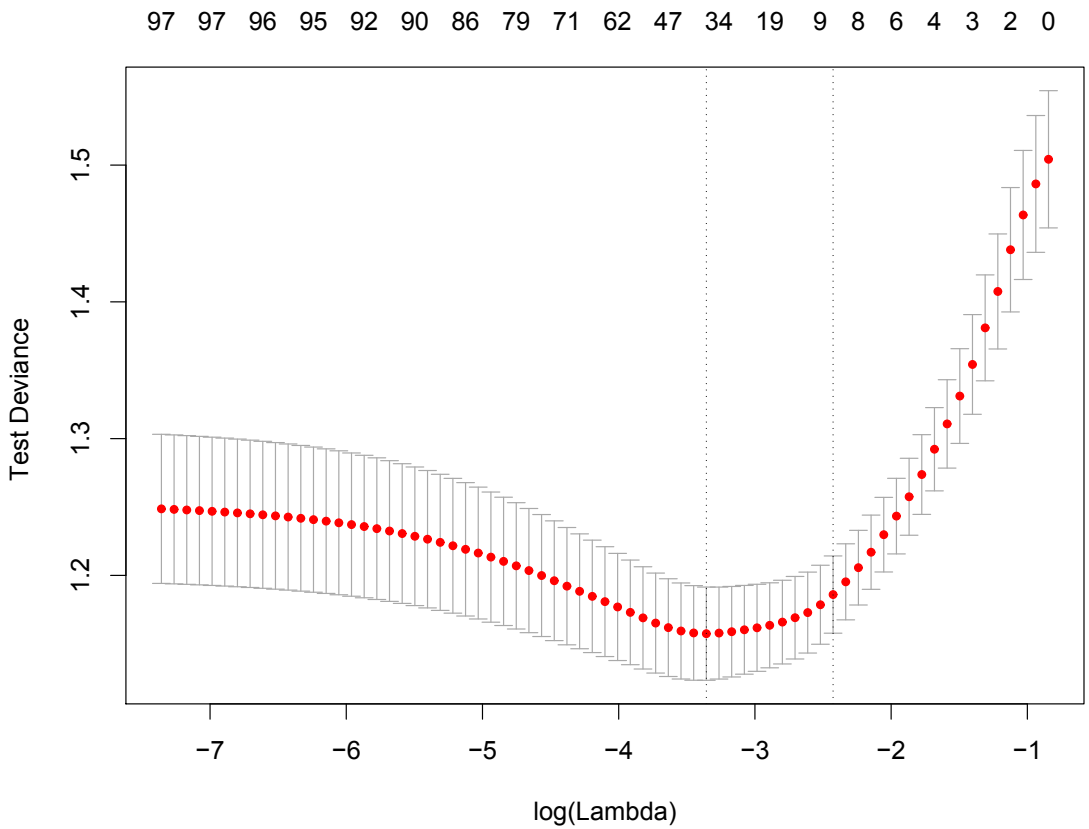
Relative ℓ_2 norm $\|\beta\|_2$ Relative ℓ_1 norm $\|\beta\|_1$

- With ridge, *all* coefficients non-zero, and heavily shrunk for large p
- Lasso does variable selection and shrinkage

Brief History of ℓ_1 Regularization

- Wavelet *Soft Thresholding* (Donoho and Johnstone 1994) in orthonormal setting.
- Tibshirani introduces *Lasso* for regression in 1995.
- Same idea used in *Basis Pursuit* (Chen, Donoho and Saunders 1996).
- Extended to many linear-model settings e.g. Survival models (Tibshirani, 1997), logistic regression, and so on.
- Gives rise to a new field *Compressed Sensing* (Donoho 2004, Candes and Tao 2005)—near exact recovery of sparse signals in very high dimensions. In many cases ℓ_1 a good surrogate for ℓ_0.

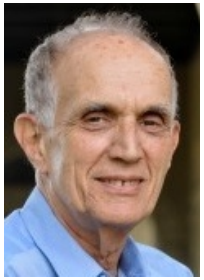
Cross Validation to select λ



Divide the data into K folds. Fit the model path K times on K-1 folds and predict on the fold omitted. Average results.

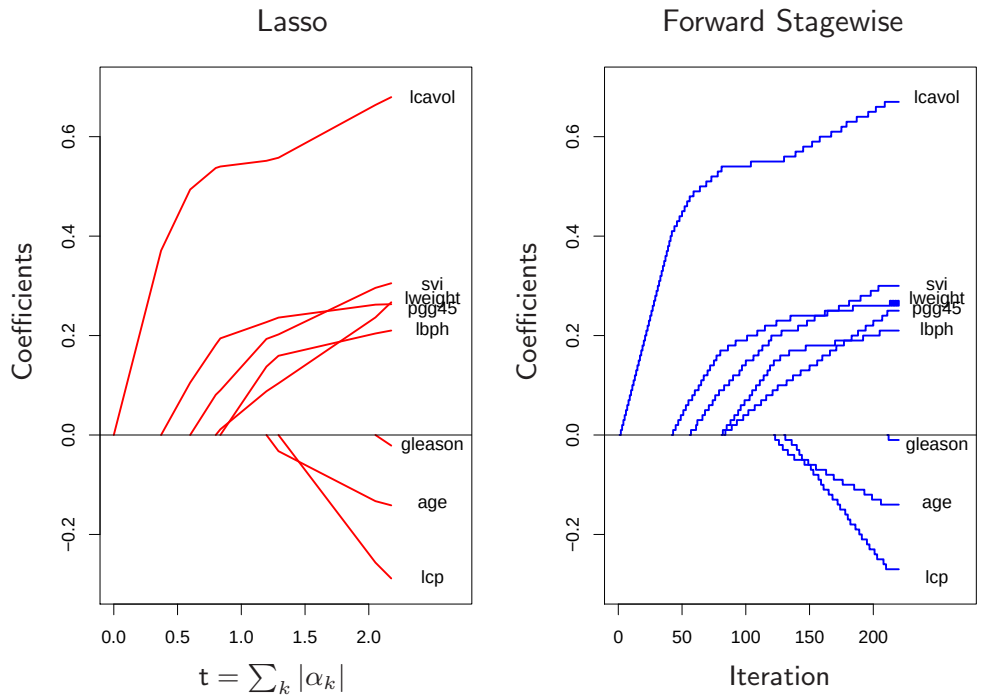
Need fast algorithms for fitting paths!

Least Angle Regression

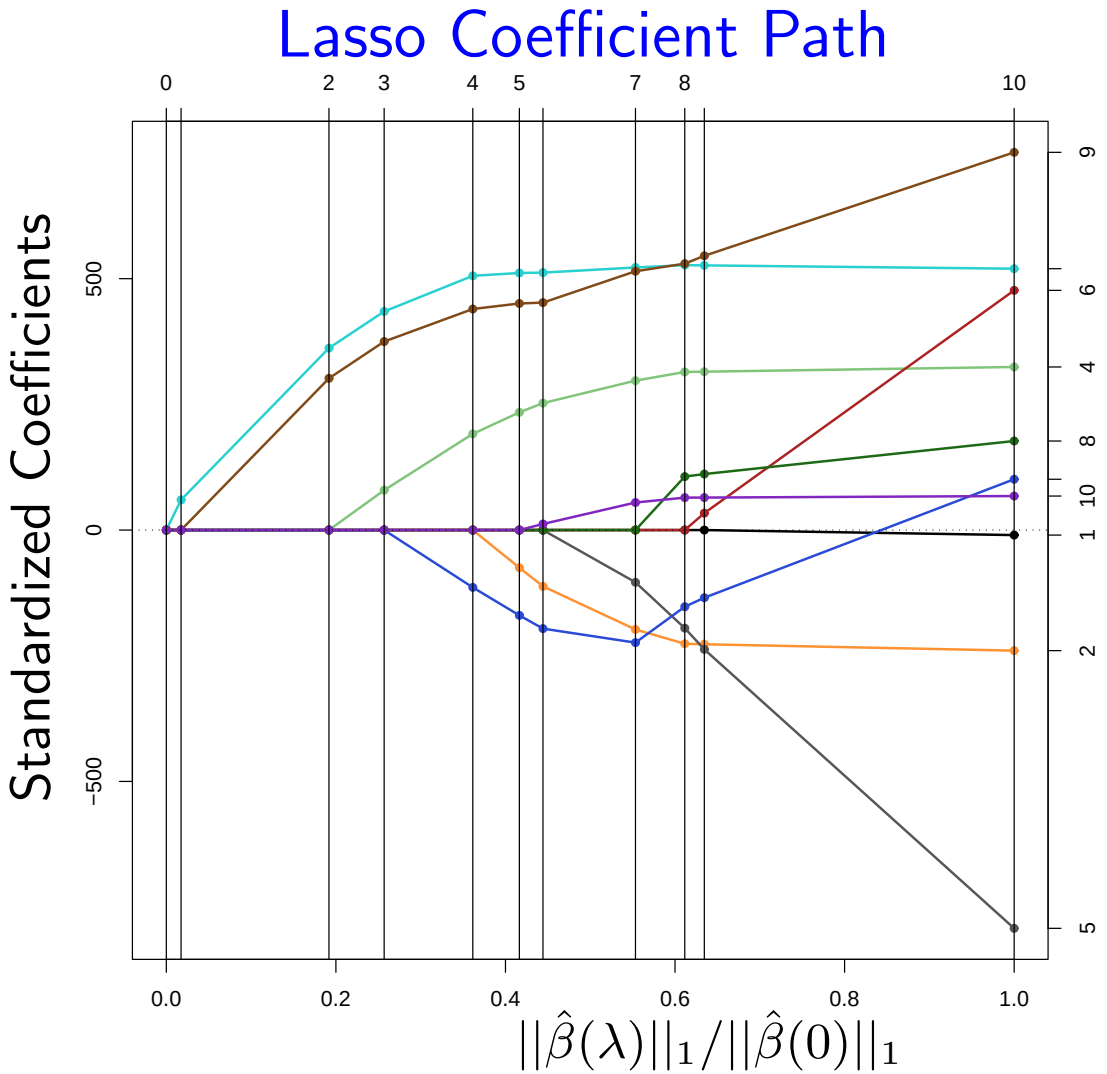


Efron, H, Johnstone, Tibshirani (2002)

The figure below is on page 609 in ESL, and shows the similarity between lasso and a version of least-squares boosting. Brad let the step-size go to zero, with some wonderful consequences.



Lasso: $\hat{\beta}(\lambda) = \text{argmin}_\beta \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 + \lambda \|\beta\|_1$



fit using LARS package in R.

Least Angle Regression — LAR

A *democratic* version of forward stepwise regression.

- Start with $r = y$, $\hat{\beta}_1, \hat{\beta}_2, \ldots \hat{\beta}_p = 0$. Assume x_j standardized.
- Find predictor x_j most correlated with r.
- Increase β_j in the direction of $\text{sign}\langle r, x_j \rangle$ until some other competitor x_k has as much correlation with current residual as does x_j.
- Move $(\hat{\beta}_j, \hat{\beta}_k)$ in the joint least squares direction for (x_j, x_k) until some other competitor x_ℓ has as much correlation with the current residual
- Continue in this way until all predictors have been entered. Stop when $\langle r, x_j \rangle = 0 \; \forall \; j$, i.e. OLS solution.

This *almost* gives the lasso path; a simple tweak and it does.

History of Path Algorithms

Efficient path algorithms for $\hat{\beta}(\lambda)$ allow for easy and exact cross-validation and model selection.

- In 2001 the LARS algorithm (Efron et al) provides a way to compute the entire lasso coefficient path efficiently at the cost of a full least-squares fit.

History of Path Algorithms

Efficient path algorithms for $\hat{\beta}(\lambda)$ allow for easy and exact cross-validation and model selection.

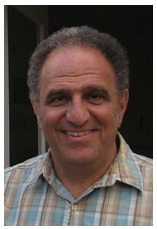
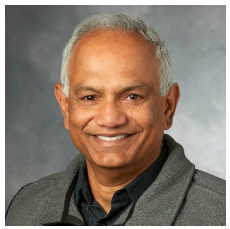
- In 2001 the LARS algorithm (Efron et al) provides a way to compute the entire lasso coefficient path efficiently at the cost of a full least-squares fit.
- 2001 – 2008: path algorithms pop up for a wide variety of related problems: Group lasso (Yuan & Lin 2006), support-vector machine (H, Rosset, Tibshirani & Zhu 2004), elastic net (Zou & H 2004), quantile regression (Li & Zhu, 2007), logistic regression and glms (Park & H, 2007), Dantzig selector (James & Radchenko 2008), ...

History of Path Algorithms

Efficient path algorithms for $\hat{\beta}(\lambda)$ allow for easy and exact cross-validation and model selection.

- In 2001 the LARS algorithm (Efron et al) provides a way to compute the entire lasso coefficient path efficiently at the cost of a full least-squares fit.
- 2001 – 2008: path algorithms pop up for a wide variety of related problems: Group lasso (Yuan & Lin 2006), support-vector machine (H, Rosset, Tibshirani & Zhu 2004), elastic net (Zou & H 2004), quantile regression (Li & Zhu, 2007), logistic regression and glms (Park & H, 2007), Dantzig selector (James & Radchenko 2008), ...
- Many of these do not enjoy the piecewise-linearity of LARS, and seize up on very large problems.

GLMNET and coordinate descent




Team: Friedman, H, Tibshirani, Narasimhan, Simon, Qian, Balasubramaniam

- Solve the lasso problem by coordinate descent: optimize each parameter separately, holding all the others fixed. Updates are trivial. Cycle around till coefficients stabilize.
- Do this on a grid of λ values, from λ_{max} down to λ_{min} (uniform on log scale), using warms starts.
- Can use a variety of loss functions and additive penalties.

Coordinate descent achieves dramatic speedups over all competitors, by factors of 10, 100 and more.

Example: Newsgroup data: 11K obs, 778K features (sparse), 100 values λ across entire range, lasso logistic regression; time 29s on a 2010 Macbook Pro.

Brief History of Coordinate Descent for the Lasso

1997 Tibshirani's student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn't fully appreciate it.

Brief History of Coordinate Descent for the Lasso

1997 Tibshirani's student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn't fully appreciate it.

2002 Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for ℓ_1 image deblurring. Hastie implements it, makes an error, and Hastie + Tibshirani conclude that the method doesn't work.

Brief History of Coordinate Descent for the Lasso

1997 Tibshirani's student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn't fully appreciate it.

2002 Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for ℓ_1 image deblurring. Hastie implements it, makes an error, and Hastie + Tibshirani conclude that the method doesn't work.

2006 Friedman is external examiner at PhD oral of Anita van der Kooij (Leiden) who uses coordinate descent for elastic net. Friedman, Hastie + Tibshirani revisit this problem.

Brief History of Coordinate Descent for the Lasso

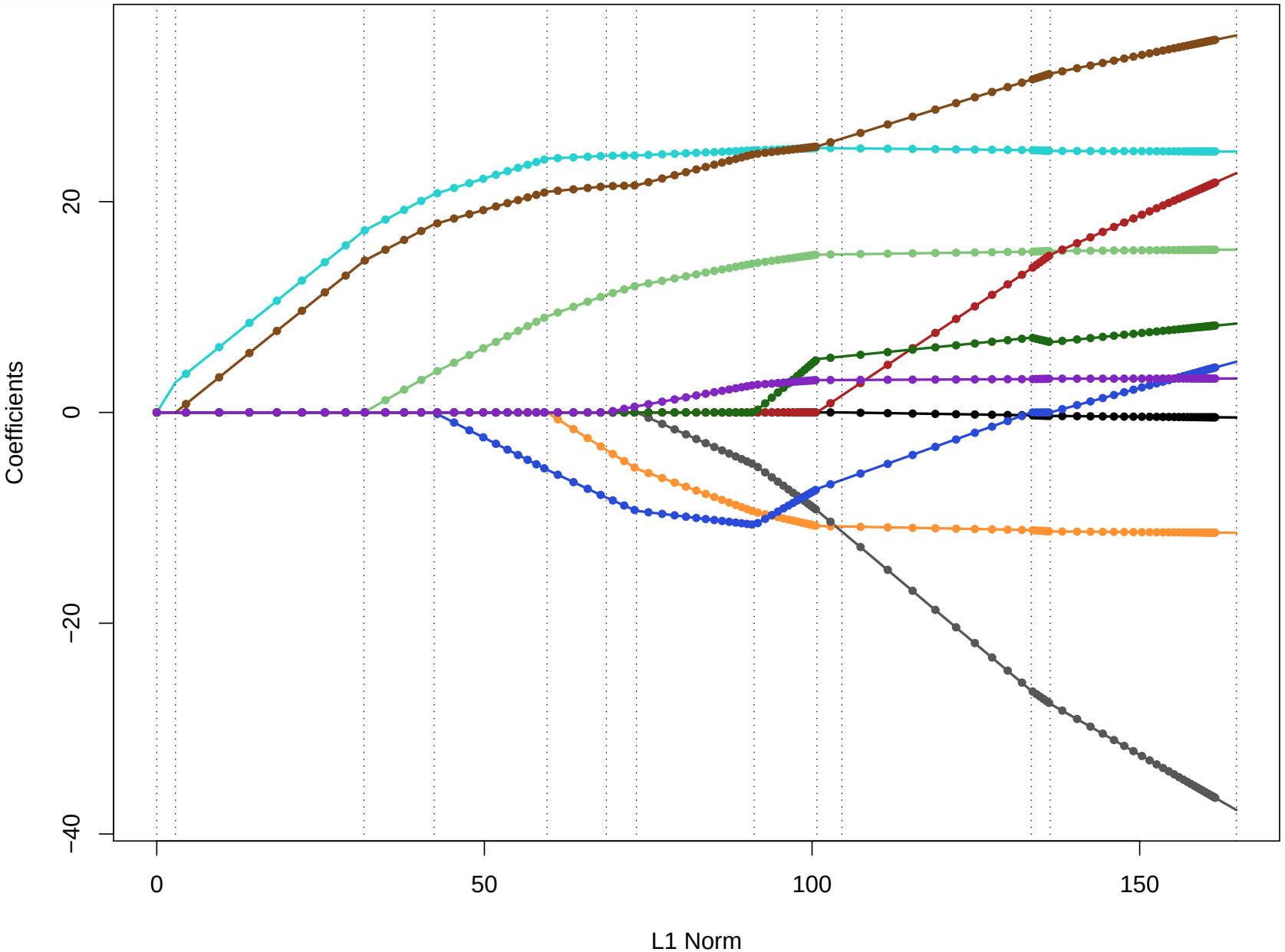
1997 Tibshirani's student Wenjiang Fu at U. Toronto develops the *shooting algorithm* for the lasso. Tibshirani doesn't fully appreciate it.

2002 Ingrid Daubechies gives a talk at Stanford, describes a one-at-a-time algorithm for ℓ_1 image deblurring. Hastie implements it, makes an error, and Hastie + Tibshirani conclude that the method doesn't work.

2006 Friedman is external examiner at PhD oral of Anita van der Kooij (Leiden) who uses coordinate descent for elastic net. Friedman, Hastie + Tibshirani revisit this problem.

Others have too — Shevade and Keerthi (2003), Krishnapuram and Hartemink (2005), Genkin, Lewis and Madigan (2007), Wu and Lange (2008), Meier, van de Geer and Bühlmann (2008).

LARS and GLMNET



GLMNET package in R

Fits coefficient paths for a variety of different GLMs and the *elastic net* family of penalties.
Some features of `glmnet`:

- Models: linear, logistic, multinomial (grouped or not), Poisson, Cox model, and multiple-response grouped linear.
- Elastic net penalty includes *ridge* and *lasso*, and hybrids in between.
- Can handle large number of variables p. Along with screening rules we can fit GLMs on GWAS scale.
- Cross-validation functions for all models.
- Can allow for sparse matrix formats for $\mathbf{X}$, and hence massive problems (eg $N = 11K$, $p = 750K$ logistic regression).

Python and Matlab versions as well on my webpage

GLMNET package in R (continued)

- Can provide lower and upper bounds for each coefficient; eg: positive lasso
- Useful bells and whistles:
  - offsets — as in GLM, can have part of the linear predictor that is given and not fit. Often used in Poisson models (sampling frame)
  - penalty strengths — can alter relative strength of penalty on different variables. Zero penalty means a variable is *always in* the model. Useful for adjusting for demographic variables
  - observation weights
  - can fit no-intercept models
  - session-wise parameters can be set with new `glmnet.options` command

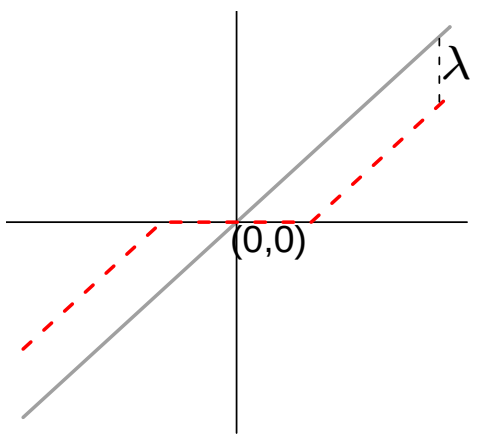
Coordinate descent for the lasso

$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Suppose the p predictors and response are standardized to have mean zero and variance 1. Initialize all the $\beta_j = 0$.
Cycle over $j = 1, 2, \ldots, p, 1, 2, \ldots$ till convergence:

- Compute the partial residuals $r_{ij} = y_i - \sum_{k \neq j} x_{ik}\beta_k$.
- Compute the simple least squares coefficient of these residuals on jth predictor: $\beta_j^* = \frac{1}{N} \sum_{i=1}^{N} x_{ij} r_{ij}$
- Update β_j by *soft-thresholding*:

$$\begin{aligned} \beta_j &\leftarrow S(\beta_j^*, \lambda) \\ &= \text{sign}(\beta_j^*)(|\beta_j^*| - \lambda)_+ \end{aligned}$$



Elastic-net penalty family

Family of convex penalties proposed in Zou and H (2005) for $p \gg N$ situations, where predictors are correlated in groups.

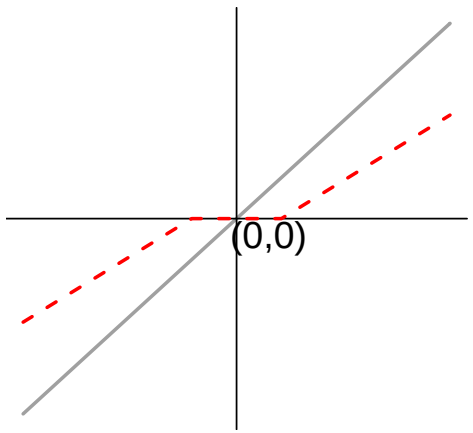
$$\min_{\beta} \frac{1}{2N} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \left[\tfrac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|\right]$$

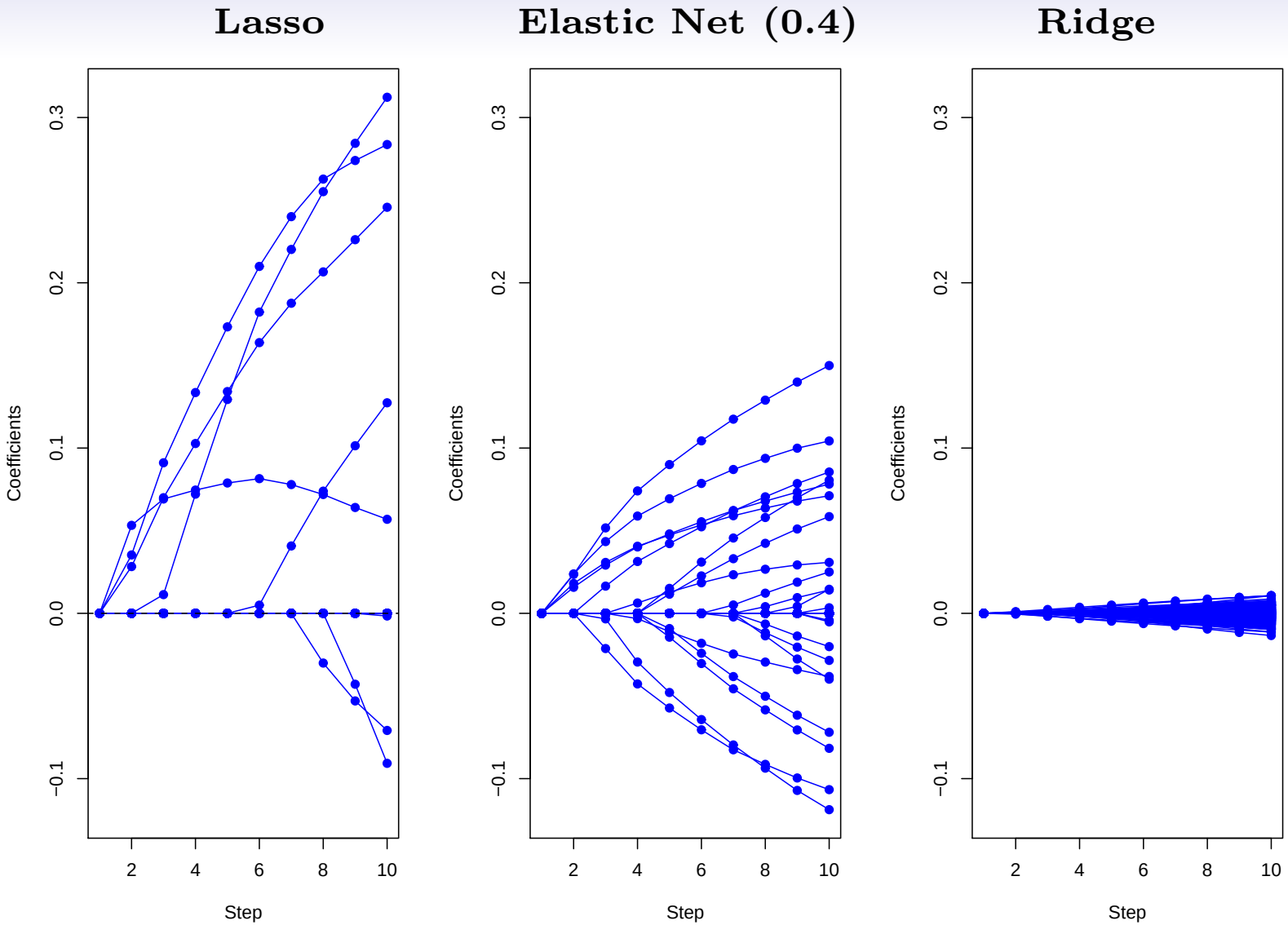
α creates a compromise between the *lasso* and *ridge*.

Coordinate update is now

$$\beta_j \leftarrow \frac{S(\beta_j^*,\ \lambda\alpha)}{1+\lambda(1-\alpha)}$$

where $\beta_j^* = \frac{1}{N}\sum_{i=1}^{N} x_{ij} r_{ij}$ as before.





Leukemia Data, Logistic, N=72, p=3571, first 10 steps shown

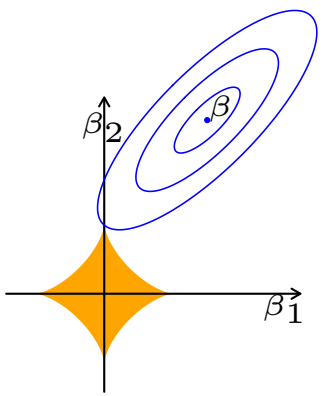
Sparser than lasso —ℓ_0 and best-subset

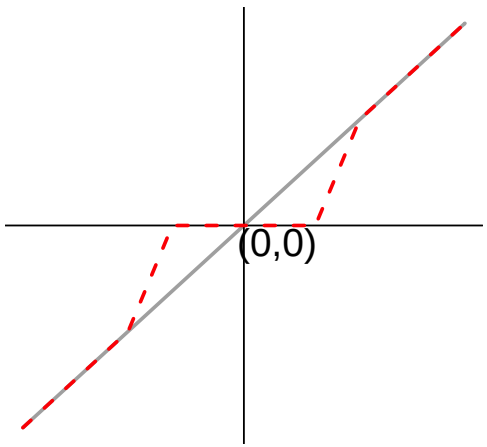
Mazumder, Friedman and H. (JASA, 2011).

SPARSENET: extends elastic-net family into concave domain. Bridges ℓ_1 and ℓ_0 via MC+ penalty (Zhang 2010) and coordinate descent.

See also PICASSO package.







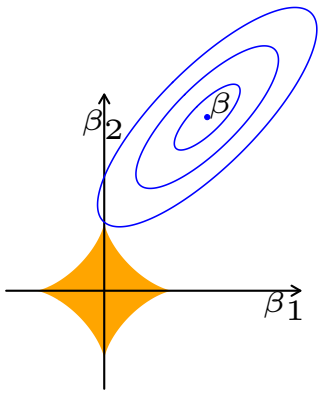
Sparser than lasso —ℓ_0 and best-subset

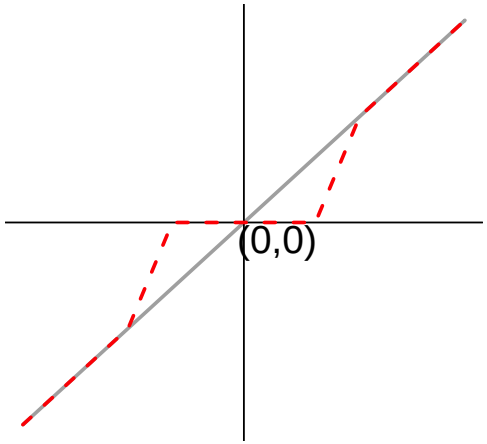
Mazumder, Friedman and H. (JASA, 2011).

SPARSENET: extends elastic-net family into concave domain. Bridges ℓ_1 and ℓ_0 via MC+ penalty (Zhang 2010) and coordinate descent.

See also PICASSO package.



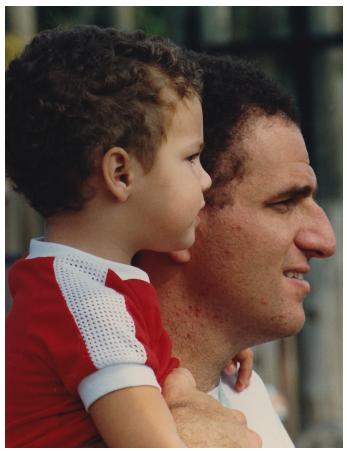




Mazumder (with Bertsimas and King, AoS 2016) crack forty year old best-subset selection bottleneck! They use *mixed-integer programming (MIO)* along with the GUROBI solver.

Best-subset, forward-stepwise or lasso?

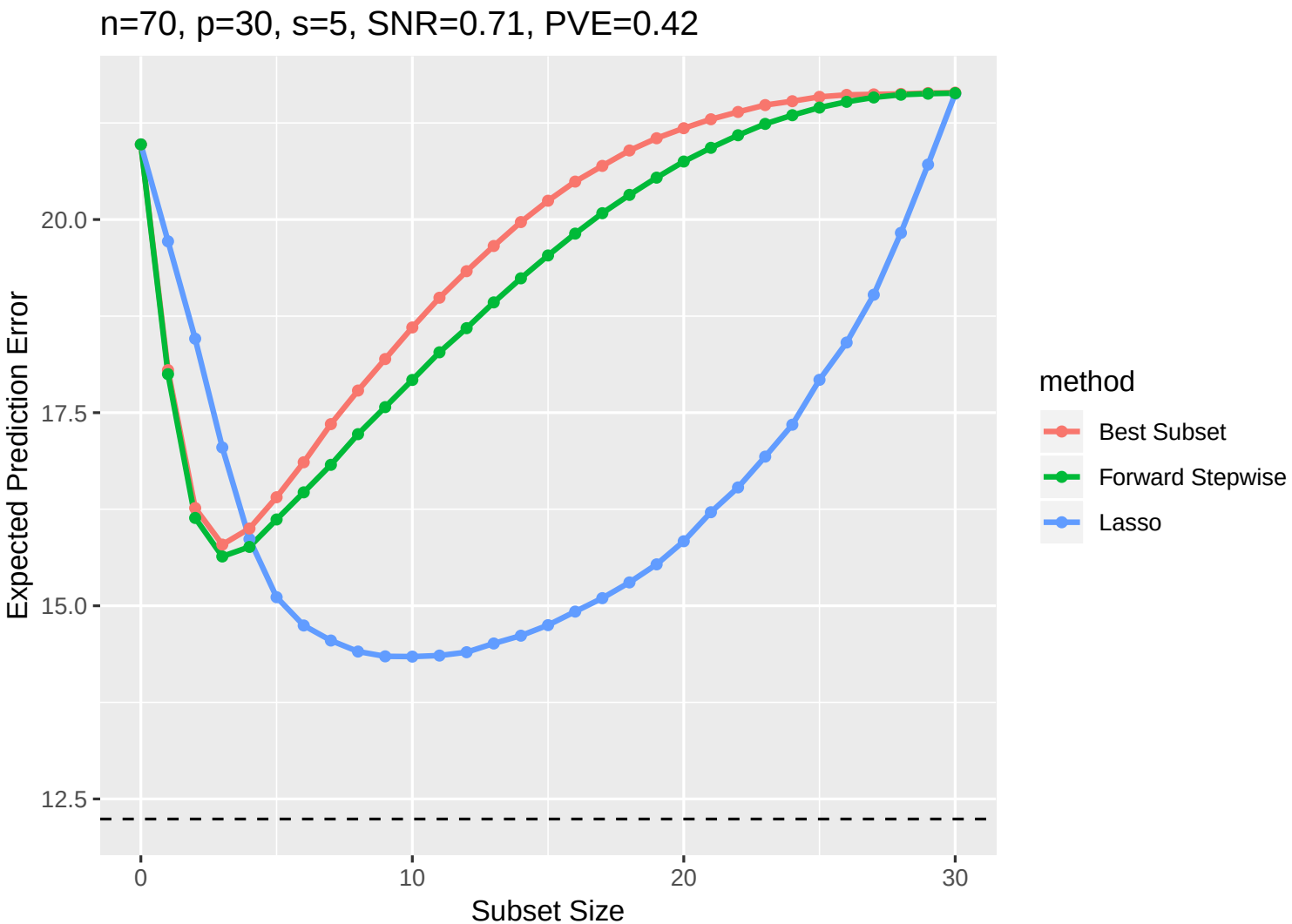
Discussion and comparisons via simulations in H, Tibshirani and Tibshirani *SS* 2019 🙏
Variety of settings varying SNR, # variables, correlations, and model structure.



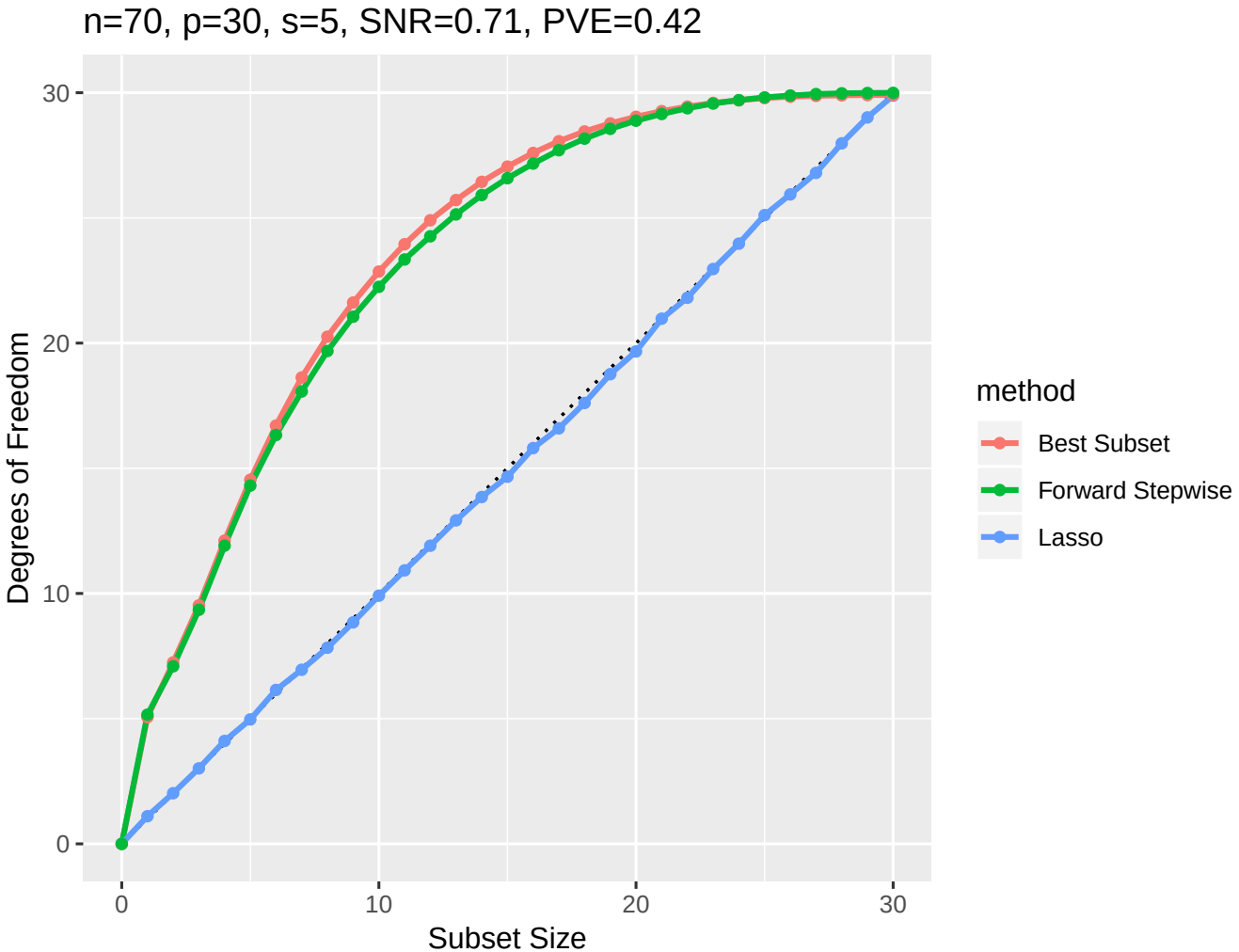
Conclusions:

- forward stepwise very close to best subset, but much faster.
- In wide-data settings, and low SNR, lasso can beat best subset and forward stepwise.
- relaxed lasso overall winner (except in very high SNR regime, where best-subset wins), and fastest by far.

Lasso versus Subset Regression: ℓ_1 vs ℓ_0



Effective degrees of freedom *df*



$$\mathrm{df} = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(y_i, \hat{y}_i)$$

Relaxed Lasso

Originally proposed by Meinshausen (2006). We present a simplified version.

- Suppose $\hat{\beta}_\lambda$ is the lasso solution at λ, and let A_λ be the *active set* of indices with nonzero coefficients in $\hat{\beta}_\lambda$.
- Let $\hat{\beta}^{LS}_{A_\lambda}$ be the coefficients in the least squares fit, using only the variables in A_λ. Let $\hat{\beta}^{LS}_\lambda$ be the full-sized version of this coefficient vector, padded with zeros.

  $\hat{\beta}^{LS}_\lambda$ debiases the lasso, while maintaining its sparsity.
- Define the *Relaxed Lasso*

$$\hat{\beta}^{RELAX}_\lambda(\gamma) = \gamma \cdot \hat{\beta}_\lambda + (1-\gamma) \cdot \hat{\beta}^{LS}_\lambda$$

Relaxed Lasso

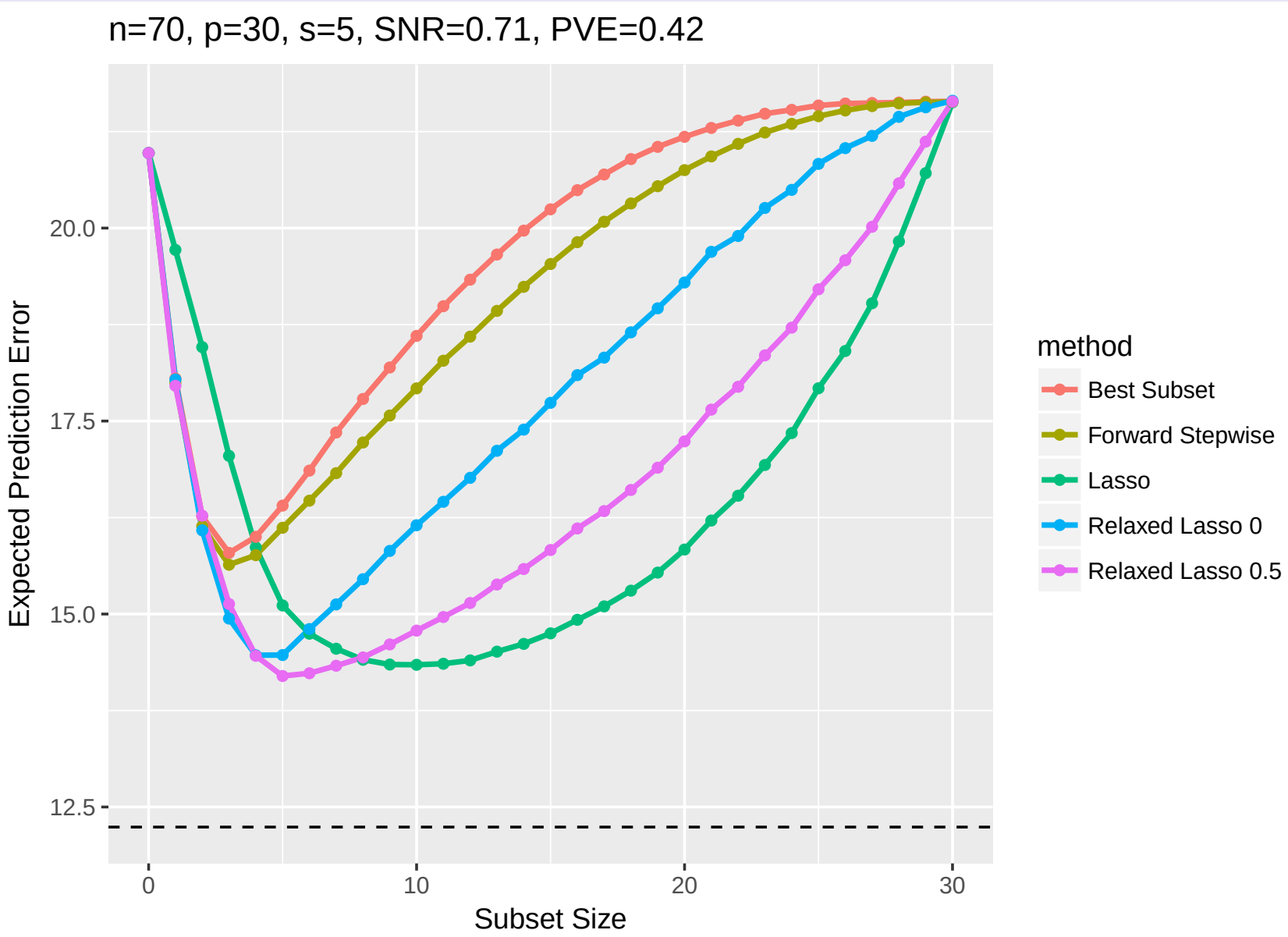
Originally proposed by Meinshausen (2006). We present a simplified version.

- Suppose $\hat{\beta}_\lambda$ is the lasso solution at λ, and let A_λ be the *active set* of indices with nonzero coefficients in $\hat{\beta}_\lambda$.
- Let $\hat{\beta}^{LS}_{A_\lambda}$ be the coefficients in the least squares fit, using only the variables in A_λ. Let $\hat{\beta}^{LS}_\lambda$ be the full-sized version of this coefficient vector, padded with zeros.

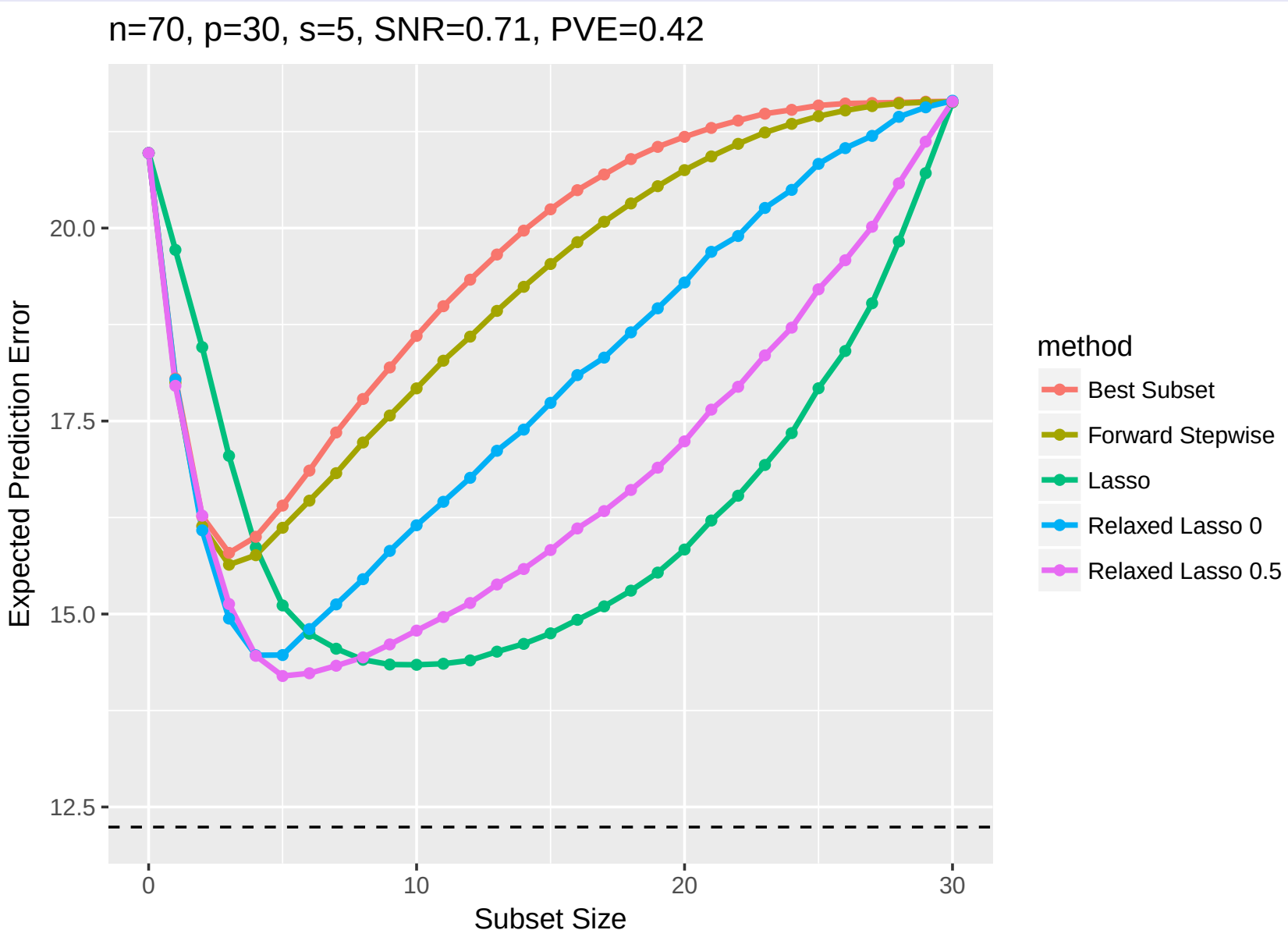
  $\hat{\beta}^{LS}_\lambda$ debiases the lasso, while maintaining its sparsity.
- Define the *Relaxed Lasso*

$$\hat{\beta}^{RELAX}_\lambda(\gamma) = \gamma \cdot \hat{\beta}_\lambda + (1-\gamma) \cdot \hat{\beta}^{LS}_\lambda$$

Once $\hat{\beta}^{LS}_\lambda$ is computed at desired values of λ, the whole family $\hat{\beta}^{RELAX}_\lambda(\gamma)$ comes free of charge!



Relaxed lasso now in GLMNET package



Relaxed lasso now in GLMNET package *(almost!)*

Screening Rules

Logistic regression for small GWAS: $N = 2000$, $p \sim$ million, (Wu et al, 2009)

- Compute $|\langle x_j, y - \bar{y} \rangle|$ for each Snp $j = 1, 2, \ldots, 10^6$, where $\bar{y}$ is the mean of (binary) y.

  Note: the largest of these is λ_{max}: smallest value of λ for which all coefficients are zero.

Screening Rules

Logistic regression for small GWAS: $N = 2000$, $p \sim$ million, (Wu et al, 2009)

- Compute $|\langle x_j, y - \bar{y} \rangle|$ for each Snp $j = 1, 2, \ldots, 10^6$, where $\bar{y}$ is the mean of (binary) y.

  Note: the largest of these is λ_{max}: smallest value of λ for which all coefficients are zero.

- Fit lasso logistic regression path using only largest M of these (e.g. 1000; for small GWAS models with 20–30 terms common).
- Simple KKT confirmations check that omitted Snps would not have entered the model.

Strong Rules

For lasso fit, with *active set* $\mathcal{A}$:

$$\left|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda)\rangle\right| \quad \begin{array}{ll} = \lambda & \forall j \in \mathcal{A} \\ \leq \lambda & \forall j \notin \mathcal{A} \end{array}$$

So variables *nearly in* $\mathcal{A}$ will have inner-products with the residuals *nearly* equal to λ.

Strong Rules

For lasso fit, with *active set* $\mathcal{A}$:

$$\left|\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda)\rangle\right| \quad \begin{array}{ll} = \lambda & \forall j \in \mathcal{A} \\ \leq \lambda & \forall j \notin \mathcal{A} \end{array}$$

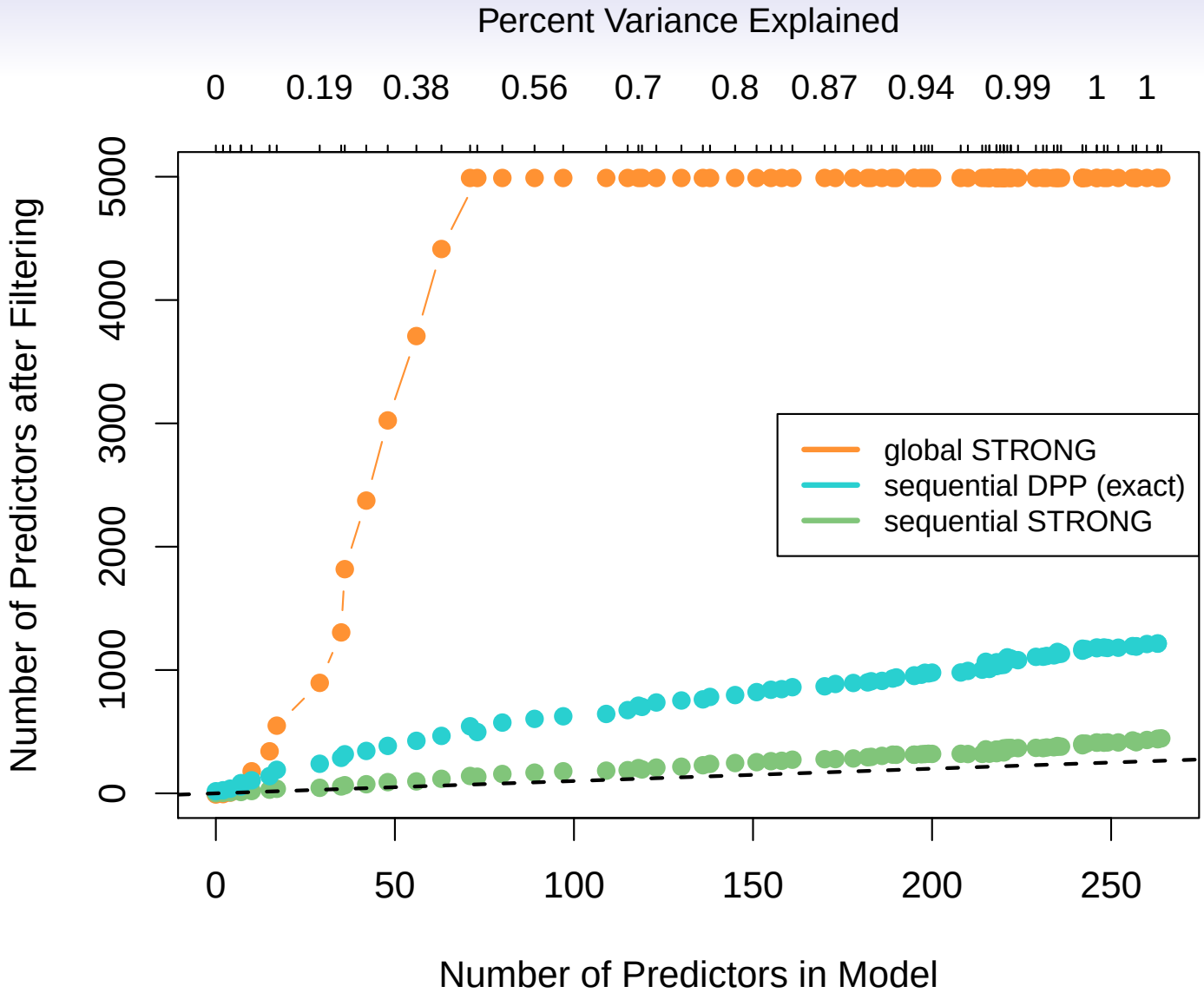
So variables *nearly in* $\mathcal{A}$ will have inner-products with the residuals *nearly* equal to λ.

Suppose fit at λ_ℓ is $\mathbf{X}\hat{\beta}(\lambda_\ell)$, and we want to compute the fit at $\lambda_{\ell+1} < \lambda_\ell$.
Strong rules gamble on set

$$\mathcal{S}_{\ell+1} = \left\{ j : |\langle \mathbf{x}_j, \mathbf{y} - \mathbf{X}\hat{\beta}(\lambda_\ell)\rangle| > \lambda_{\ell+1} - (\lambda_\ell - \lambda_{\ell+1}) \right\}$$

GLMNET screens at every λ step, and after convergence, checks if any violations. Mostly $\mathcal{A}_{\ell+1} \subseteq \mathcal{S}_{\ell+1}$.

* Tibshirani, Bien, Friedman, H, Simon, Taylor, Tibshirani (JRSSB 2012)



Strong rules inspired by El Ghaoui, Viallon and Rabbani (2010)

Sequential DPP due to Wang, Lin, Gong, Wonka and Ye (2013)

Example: UK Biobank

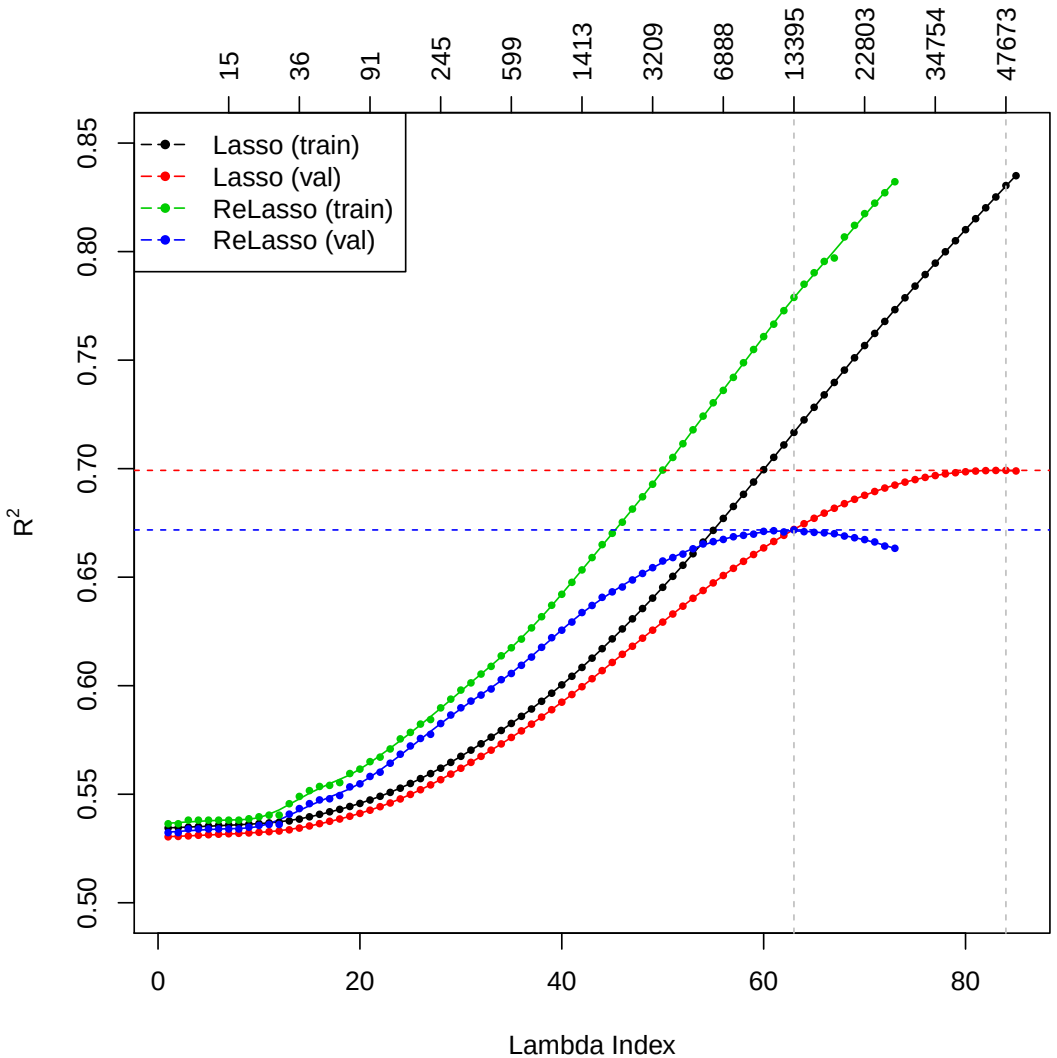




- Work with PhD students Junyang Qian, Yosuke Tanigawa, Rob Tibshirani and the Manny Rivas group at Stanford DBDS
- Large cohort of 500K British adults (Bycroft et al 2018)
- Each individual genotyped at 805K locations (AA, Aa, aa or NA)
- 100s of phenotypes measured on each subject
- We looked at white British subset of 337K, and illustrate with `height`, `asthma`, and `MI` phenotypes.
- Divided the data 60% training, 20% validation, 20% test.

Package `SNPNET` available on Github (link on my website, and to 2019 report)

Lasso fit to `height`

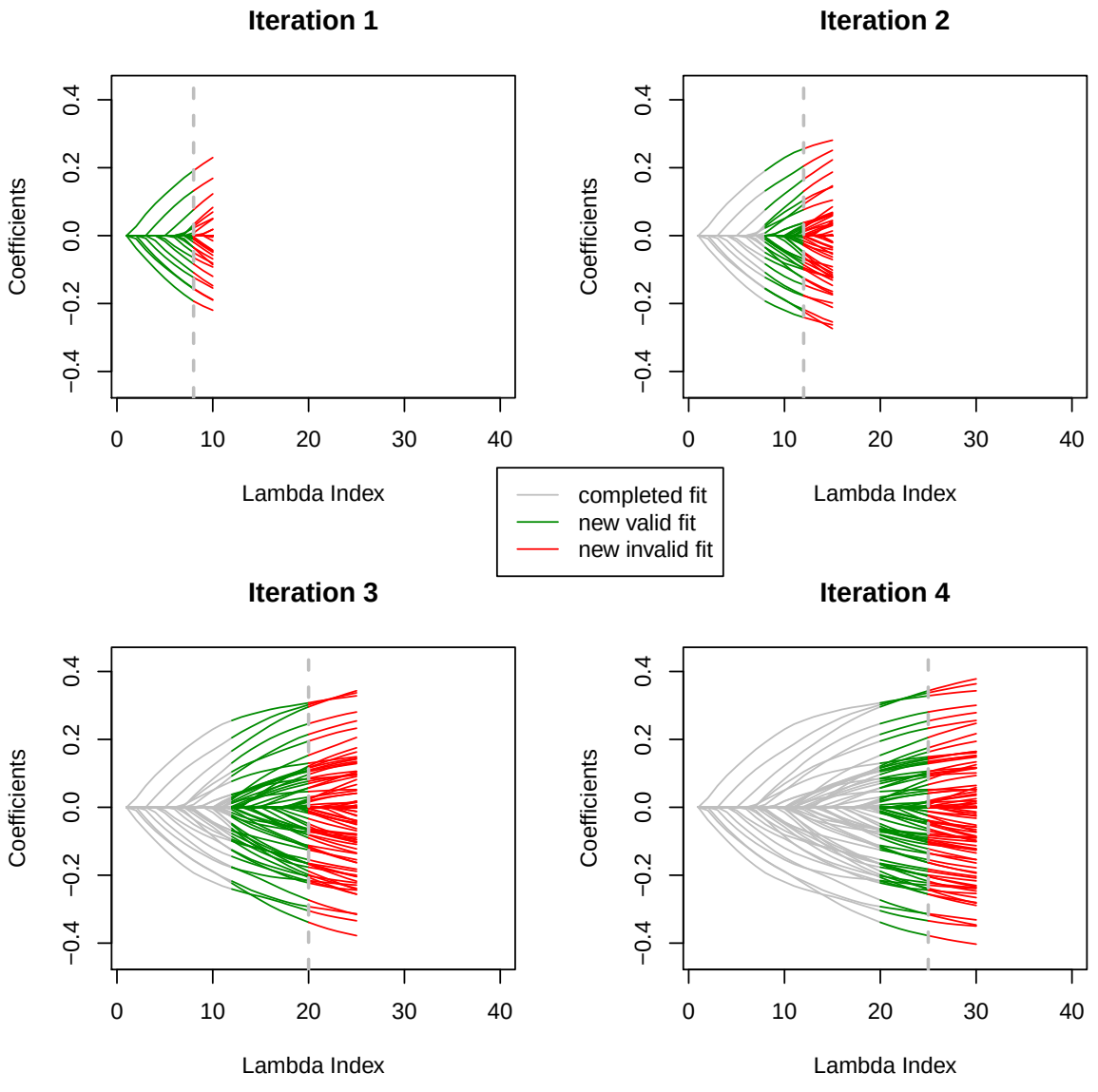


R-squared or heritability shown for regularization paths

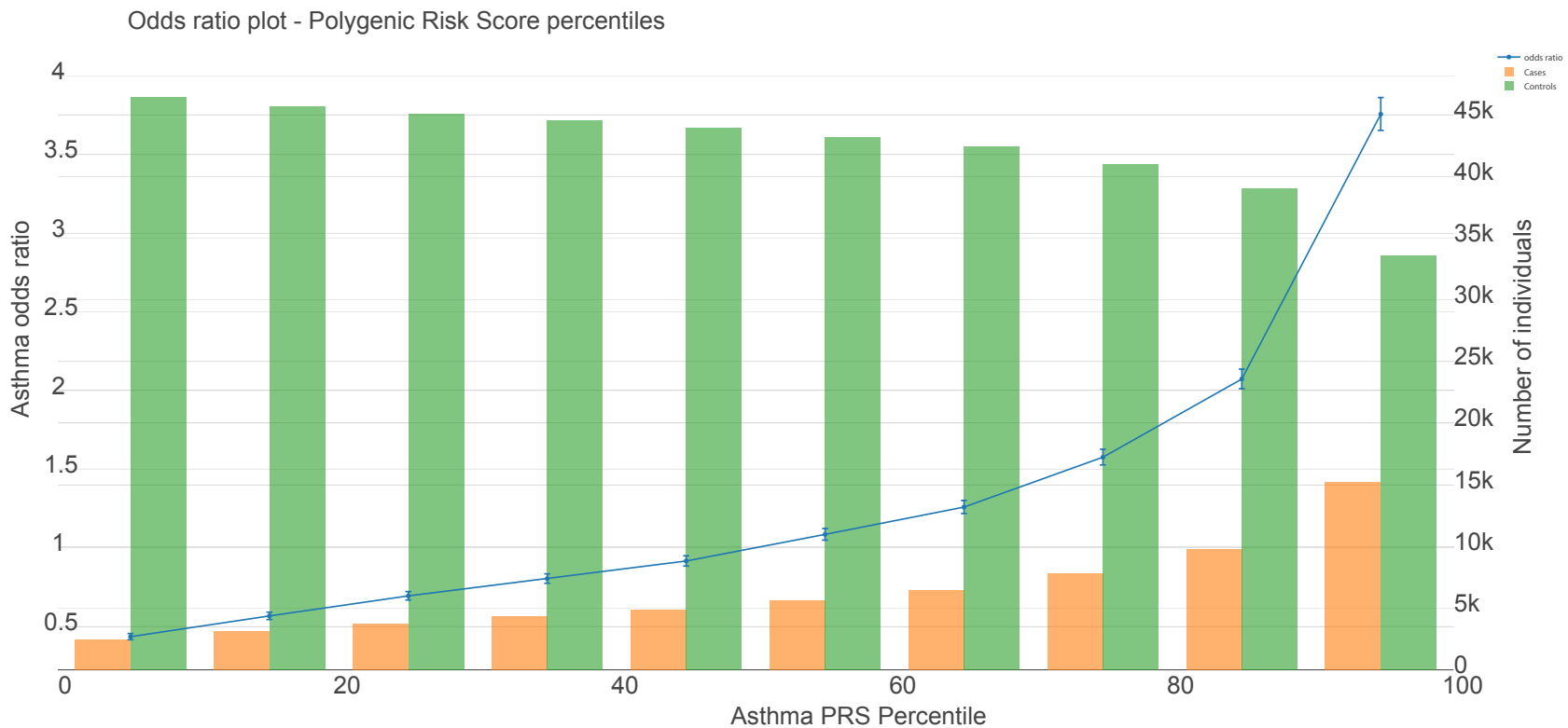
Detailed results

Model	Form	R^2_{train}	R^2_{val}	R^2_{test}	Size
(1)	Age + Sex	0.5300	0.5260	0.5288	2
(2)	Age + Sex + 10 PCs	0.5344	0.5304	0.5336	12
(3)	Strong Single SNP	0.5364	0.5323	0.5355	13
(4)	10K Combined	0.5482	0.5408	0.5444	10,012
(5)	100K Combined	0.5833	0.5515	0.5551	100,012
(6)	Sequential LR	0.7416	0.6596	0.6601	17,012
(7)	Lasso	0.8304	0.6992	**0.6999**	47,673
(8)	Relaxed Lasso	0.7789	0.6718	0.6727	13,395

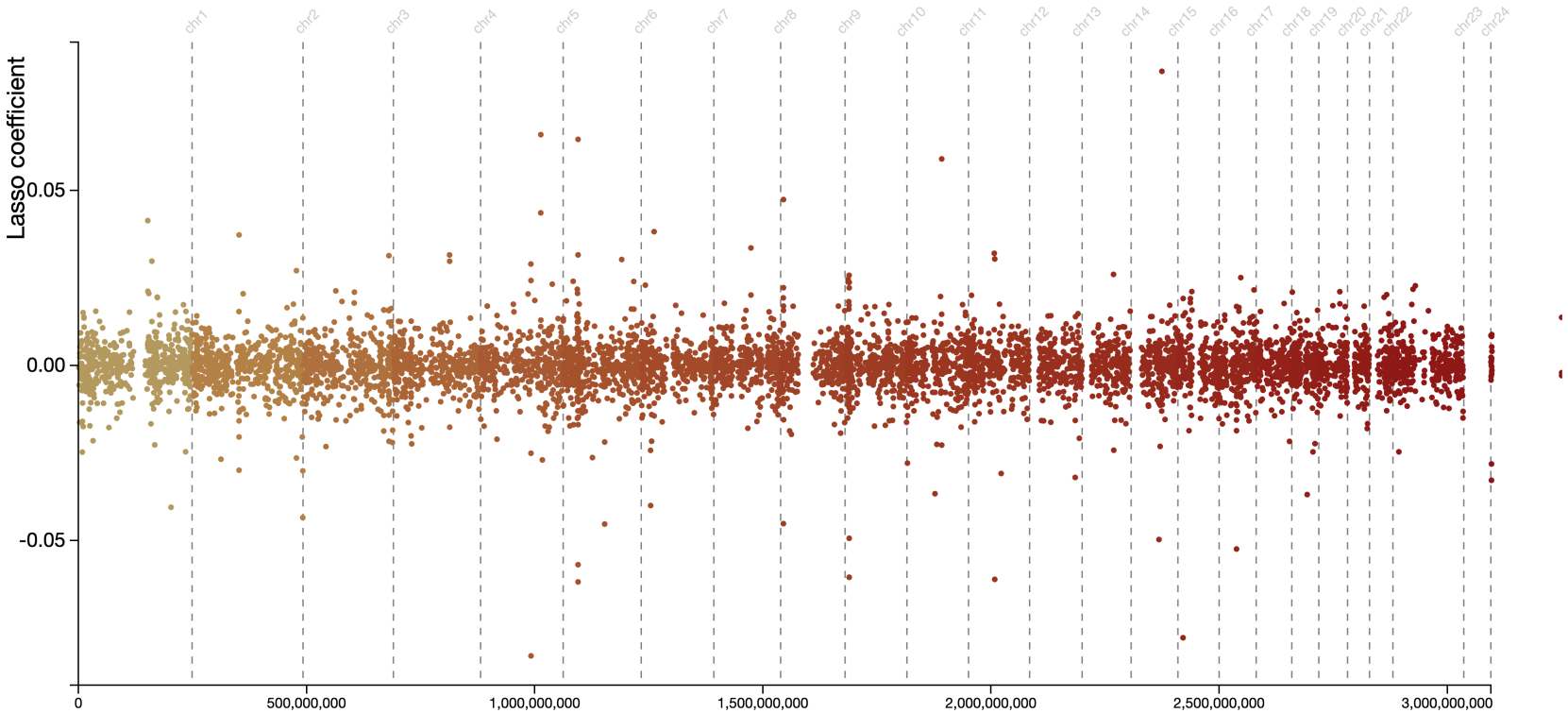
Algorithmic details in pictures



UK Biobank — Asthma

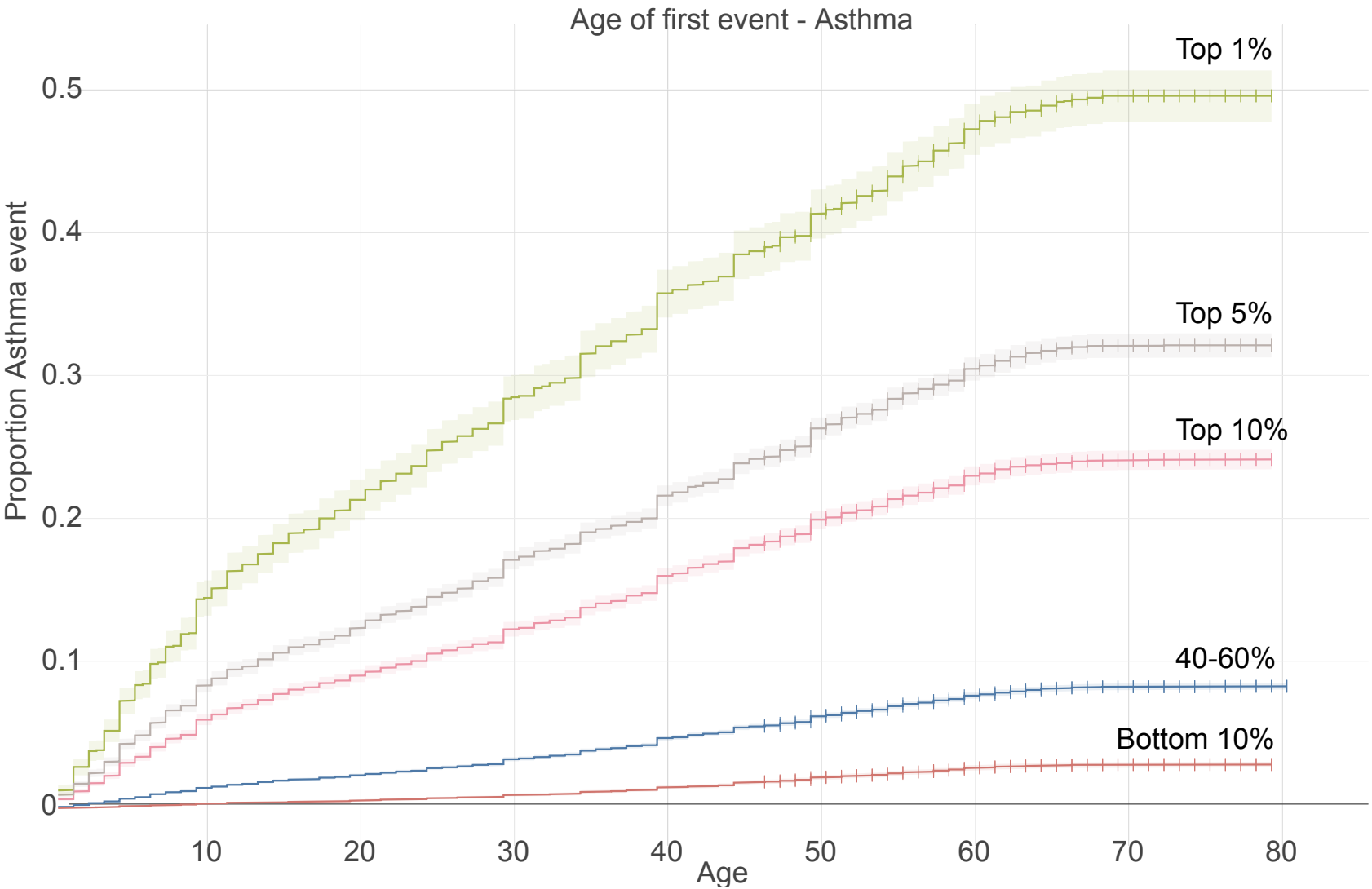


UK Biobank — Asthma



Note: only 6347 variants in active path for Lasso are included in th coefficient plot.

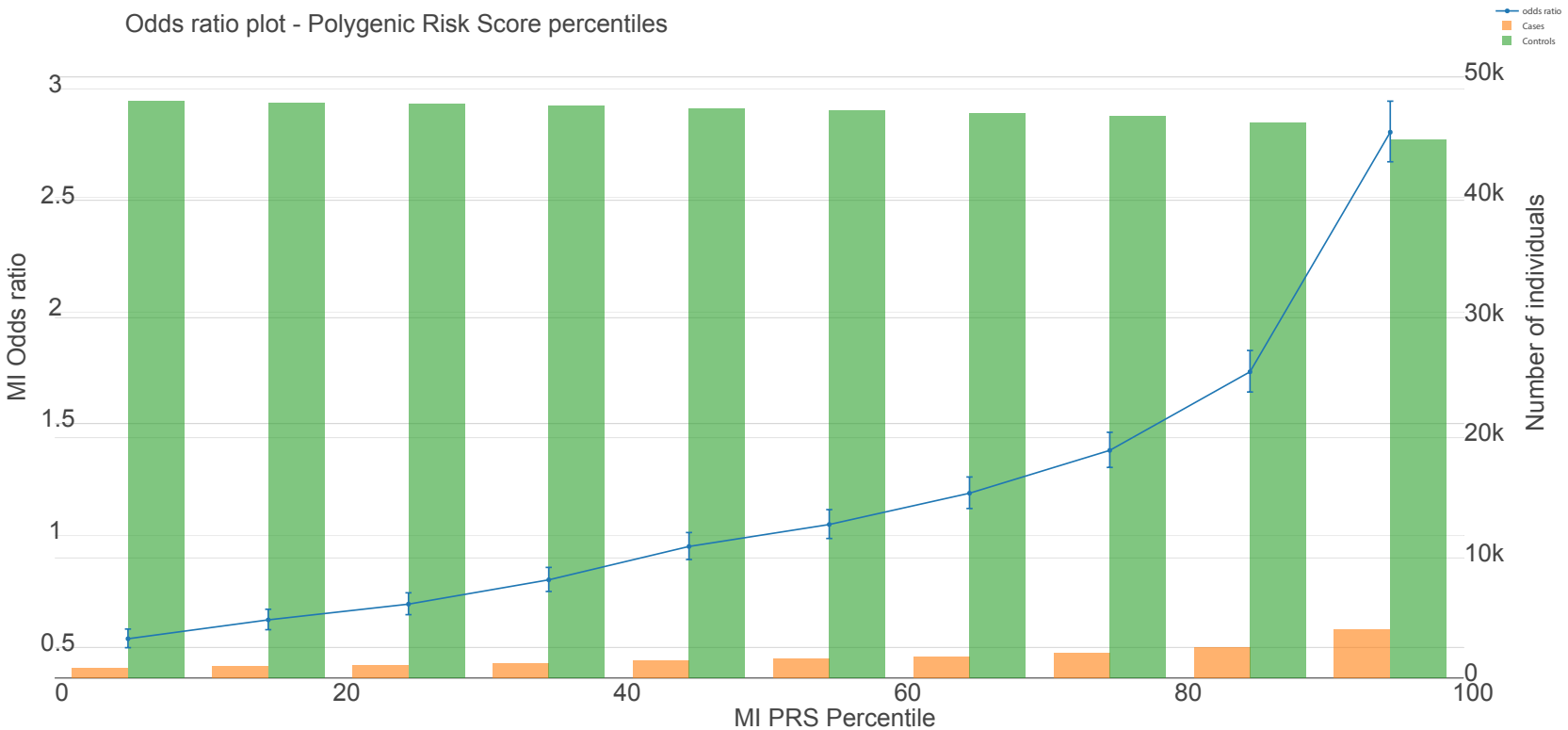
UK Biobank — Asthma



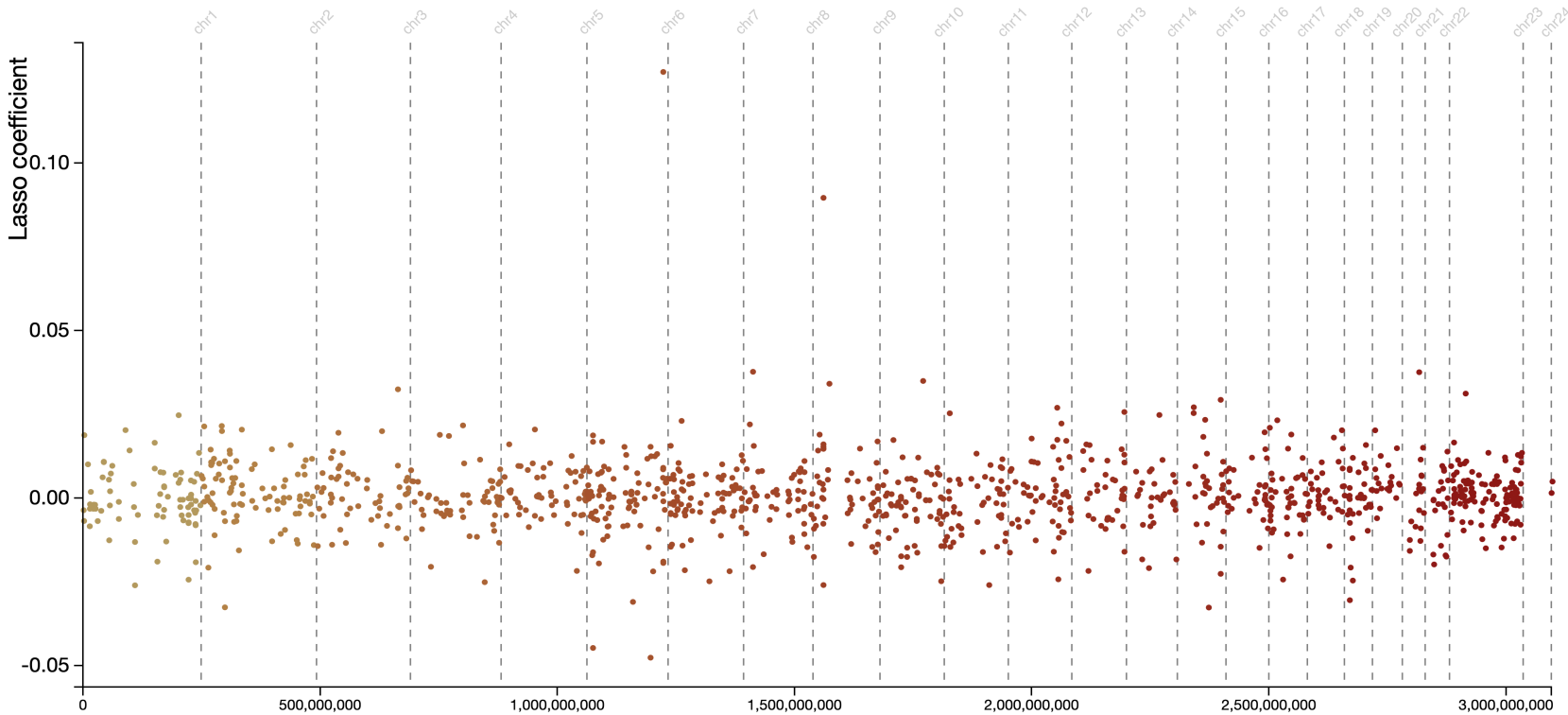
UK Biobank — Asthma

Variable	Level	N	Events	Hazard Ratio	*P* value
Polygenic risk score	40-60%	62139	5017	Reference	
	Bottom 10%	31036	896	0.36 (0.34, 0.39)	<0.001
	Top 1%	3085	1494	7.31 (6.90, 7.74)	<0.001
	Top 10%	15480	3636	3.06 (2.93, 3.20)	<0.001
	Top 5%	12403	3894	4.27 (4.09, 4.45)	<0.001
Eosinophil threshold	below .18	73655	6392	Reference	
	above .18	50488	8545	1.61 (1.56, 1.67)	<0.001
Sex	Female	66292	8330	Reference	
	Male	57851	6607	0.87 (0.85, 0.90)	<0.001
Body Mass Index	Below 25	40969	4414	Reference	
	25-30	53273	6220	1.04 (1.00, 1.08)	0.06
	30-35	21600	2880	1.12 (1.07, 1.17)	<0.001
	35-40	6010	954	1.26 (1.17, 1.35)	<0.001
	Above 40	2291	469	1.59 (1.44, 1.75)	<0.001

UK Biobank — Myocardial Infarct

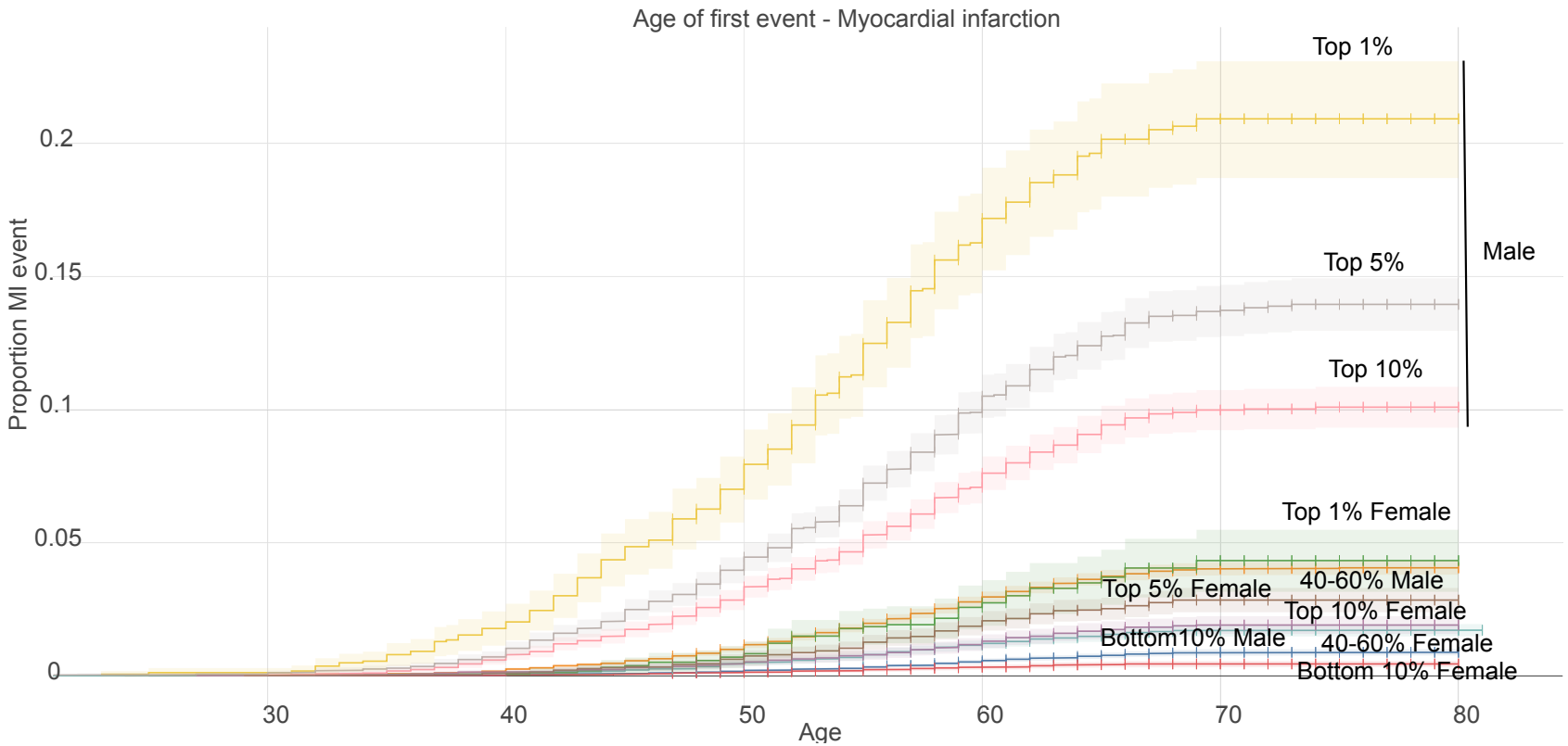


UK Biobank — Myocardial Infarct



Note: only 1197 variants in active path for Lasso are included in th coefficient plot.

UK Biobank — Myocardial Infarct



UK Biobank — Myocardial Infarct

Variable	Level	N	Events	Hazard Ratio	*P* value
Polygenic risk score	40-60%	63742	1228	Reference	
	Bottom 10%	31918	280	0.48 (0.42, 0.54)	<0.001
	Top 1%	3187	349	4.57 (4.06, 5.15)	<0.001
	Top 10%	15948	771	2.24 (2.05, 2.45)	<0.001
	Top 5%	12738	871	3.05 (2.79, 3.33)	<0.001
Smoke	Not current smoker	114637	3010	Reference	
	Occassionally	3275	106	1.19 (0.98, 1.44)	0.081
	Yes, most days	9621	383	1.35 (1.21, 1.50)	<0.001
Sex	female	67841	644	Reference	
	male	59692	2855	5.12 (4.69, 5.58)	<0.001
Body Mass Index	below 25	42311	604	Reference	
	25-30	54416	1560	1.13 (1.02, 1.24)	0.015
	30-35	22228	945	1.20 (1.08, 1.34)	0.001
	35-40	6195	275	1.22 (1.05, 1.42)	0.008
	above 40	2383	115	1.11 (0.91, 1.37)	0.306
non-LDL cholesterol		127533		0.19 (0.18, 0.20)	<0.001
Triglycerides		127533		1.07 (1.03, 1.11)	<0.001
LDL-c		127533		2.86 (2.71, 3.02)	<0.001
Diabetes	non-diabetic	121750	2855	Reference	
	diabetic	5783	644	1.19 (1.08, 1.30)	<0.001

Coordinate Descent in General

Many problems have the form

$$\min_{\{\beta_j\}_1^p} \left[R(y, \beta) + \lambda \sum_{j=1}^{p} P_j(\beta_j) \right].$$

- If R and P_j are convex, and R is differentiable, then coordinate descent converges to the solution (Tseng, 1988).
- Often each coordinate step is trivial. E.g. for lasso, it amounts to soft-thresholding, with many steps leaving $\hat{\beta}_j = 0$.
- each term $P_j(\beta_j)$ can refer to *sets* of parameters.
- Decreasing λ slowly means not much cycling is needed.
- Coordinate moves can exploit sparsity.

Group Lasso

Yuan and Lin, 2007, Meier, Van de Geer, Bühlmann, 2008.

$$R\Big(y, \sum_{j=1}^{J} \mathbf{X}_j \beta_j\Big) + \lambda \sum_{j=1}^{J} \gamma_j \|\beta_j\|_2.$$

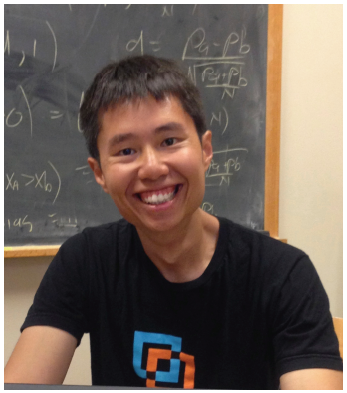
Example: each block $\mathbf{X}_j$ represents the levels for a categorical predictor.

- entire groups are zero, or all elements are nonzero.
- γ_j is penalty modifier for group j; $\gamma_j = \|\mathbf{X}_j\|_F$ is good choice.
- Leads to a block-updating form of coordinate descent.
- Strong rules apply here: $\|\mathbf{X}_j^T \mathbf{r}\|_2 > \gamma_j[\lambda_{\ell+1} - (\lambda_\ell - \lambda_{\ell+1})]$

Sparse interaction models

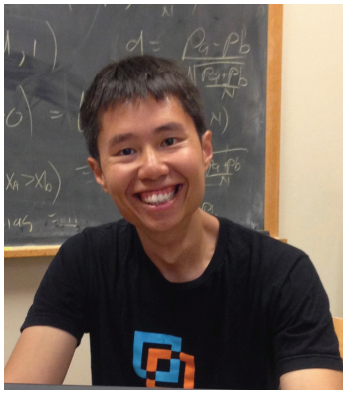
GLINTERNET

With past PhD student Michael Lim (JCGS 2014 and in Chap 4 of SLS book)
Hierarchical linear + first-order interaction models selected using group lasso



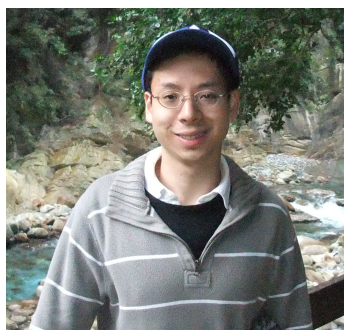
Sparse interaction models

GLINTERNET



With past PhD student Michael Lim (JCGS 2014 and in Chap 4 of SLS book)
Hierarchical linear + first-order interaction models selected using group lasso

Mixed Graphical Models



Allows for both continuous and categorical variables. With past PhD student Jason Lee (JCGS 2015) *(in Wald III)*
General Markov random field representation, with edge and node potentials selected by group lasso

Sparse Generalized Additive Models

$$\eta(x) = \alpha_0 + \sum_{j=1}^{p} f_j(x_j)$$

GAMSEL with Alexandra Chouldechova (Arxiv 2015) using *overlap group lasso* Automatic, *sticky* selection between zero, linear, or nonlinear terms in GAMs.



Sparse Generalized Additive Models

$$\eta(x) = \alpha_0 + \sum_{j=1}^{p} f_j(x_j)$$

GAMSEL with Alexandra Chouldechova (Arxiv 2015) using *overlap group lasso* Automatic, *sticky* selection between zero, linear, or nonlinear terms in GAMs.



SATURATED SPLINES with Nick Boyd and S. Boyd, Recht and Jordan (JMLR 2018)
Adaptive GAMs using saturated piecewise-linear splines, with automatic feature and knot selection



Matrix Completion

softImpute with past PhD students Rahul Mazumder, Jason Lee, and with Rob Tibshirani and Reza Zadeh (JMLR 2010, 2015)

$$\underset{Z}{\text{minimize}} \sum_{\text{Observed}(i,j)} (X_{ij} - Z_{ij})^2 + \lambda \|Z\|_*$$

where *nuclear norm* $\|Z\|_*$ is the sum of singular values of Z.



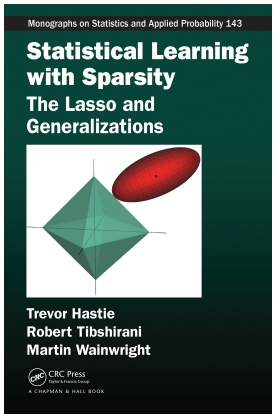
Wald II

Undirected Graphical Models — learning dependence structure via the lasso. Model the inverse covariance $\mathbf{\Theta}$ in the Gaussian family with L_1 penalties applied to elements.

$$\max_{\mathbf{\Theta}} \log\det \mathbf{\Theta} - \mathrm{Tr}(\mathbf{S\Theta}) - \lambda\|\mathbf{\Theta}\|_1$$

GLASSO: modified block-wise lasso algorithm, which we solve by coordinate descent (Friedman, H, Tibshirani 2007). Algorithm is very fast, and solves moderately sparse graphs with 1000 nodes in under a minute.
Topic of *Wald III*, as well as a recent application of graphical models and factor models in anomaly detection.

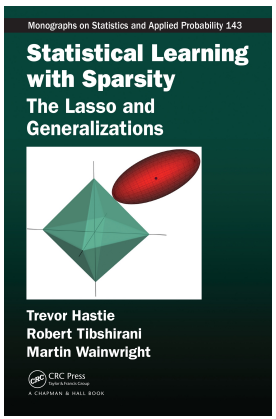
Theory



Theory for lasso: see chapter 11, written by our coauthor Martin Wainwright.

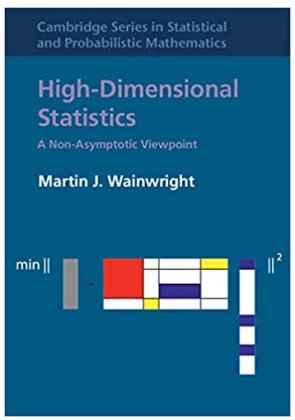


Theory



Theory for lasso: see chapter 11, written by our coauthor Martin Wainwright.





Also new 2019 book by Martin Wainwright (Cambridge) *High-Dimensional Statistics — a non-asymptotic viewpoint*

Statistical Inference

- Can become Bayesian! Lasso penalty corresponds to Laplacian prior. However, need priors for everything, including λ (variance ratio). *Also posterior not sparse!*

Statistical Inference

- Can become Bayesian! Lasso penalty corresponds to Laplacian prior. However, need priors for everything, including λ (variance ratio). *Also posterior not sparse!*
- Bootstrap. Easier to do, with similar results.

Statistical Inference

- Can become Bayesian! Lasso penalty corresponds to Laplacian prior. However, need priors for everything, including λ (variance ratio). *Also posterior not sparse!*
- Bootstrap. Easier to do, with similar results.
- Post-selection inference since circa 2013 ...
  - *A Significance Test for the Lasso* — Richard Lockhart, Jonathan Taylor, Ryan Tibshirani and Rob Tibshirani (AoS, 2014)
  - *Exact Post-Selection Inference, with application to the Lasso* — Jason Lee, Dennis Sun, Yuekai Sun, Jonathan Taylor (AoS, 2016)
  - *Optimal Inference after Model Selection* — Will Fithian, Dennis Sun, Jonathan Taylor (2015, arXiv)

Statistical Inference continued

- *Selective inference with a randomized response* — Xioying Tian and Jonathan Taylor (AoS 2018)
- and many more papers with students of Taylor and Tibshirani, including Keli Liu, Jelena Markovic, Jeremy Taylor, Snighda Panigrahi, Nan Bi.

Statistical Inference continued

  - *Selective inference with a randomized response* — Xioying Tian and Jonathan Taylor (AoS 2018)
  - and many more papers with students of Taylor and Tibshirani, including Keli Liu, Jelena Markovic, Jeremy Taylor, Snighda Panigrahi, Nan Bi.

- *Knockoffs* — false discovery control for sequences of models such as those produced by lasso. Since circa 2015... Emmanuel Candes with collaborators Rina Foygel Barber, Chiara Sabatti, and students including Stephen Bates, Lucas Jansen, Eugene Katsevich, Matteo Sessio, Asaf Weinstein.

Summary

- Sparsity via convex relaxation both effective and computable.
- Although lasso 25 years old, still super cool!
- Good public software is essential.
- Coordinate descent simple and effective, especially when models are sparse.
- Screening rules allow for massive computational savings.
- Many extensions to lasso ideas — I present some in *Wald II* and *Wald III*

Summary

- Sparsity via convex relaxation both effective and computable.
- Although lasso 25 years old, still super cool!
- Good public software is essential.
- Coordinate descent simple and effective, especially when models are sparse.
- Screening rules allow for massive computational savings.
- Many extensions to lasso ideas — I present some in *Wald II* and *Wald III*

One down, two to go — Thank you for your attention!