Reduced-rank Vector Generalized Linear Models

Thomas W. Yee,

University of Auckland, New Zealand

and Trevor J. Hastie

Stanford University, USA

Summary. Reduced-rank regression is proposed for the class of vector generalized linear and additive models (VGLMs/VGAMs; Yee and Wild (1996)), resulting in the subclasses RR-VGLMs and RR-VGAMs. By focusing particularly on models for categorical data, and especially the multinomial logit model, we show that the reduced-rank idea can be motivated by linear discriminant analysis and neural networks, and are related to other multivariate methods including canonical correspondence analysis and biplots. We provide algorithmic details, and consider its use for classification problems.

Keywords: Canonical correspondence analysis; Linear discriminant analysis; Neural networks; Non-parametric multivariate regression; Projection pursuit regression; Vector generalized additive models.

Running title: Reduced-rank Vector Generalized Linear Models

†*Address for correspondence*: Thomas Yee, Department of Statistics, University of Auckland, Private Bag 92019, Auckland, New Zealand.

E-mail: yee@stat.auckland.ac.nz

1 Introduction

Reduced-rank regression (Anderson, 1951; Izenman, 1975) has a significant niche in the classical theory of multivariate analysis. In this setting it has been restricted to data where the response variable is continuous; even recent monographs (Schmidli 1995, Reinsel and Velu 1998) do not mention its use outside the Gaussian family. Consequently, this may be the major reason why reduced-rank regression has found few applications (ter Braak and Looman, 1994).

To circumvent this limitation, we consider the class of vector generalized linear models (VGLMs) and its nonparametric extension, vector generalized additive models (VGAMs; Yee and Wild (1996)). These classes are very large and encompass a wide range of multivariate response types and models. For example, it includes univariate and multivariate distributions, categorical data analysis, time series, survival analysis, generalized estimating equations, correlated binary data, bioassay data and nonlinear least-squares problems. In this article we extend the reduced-rank idea to the VGLM/VGAM classes to obtain subclasses which we term RR-VGLMs and RR-VGAMs. The multinomial logit model (MLM; Nerlove and Press, 1973) for categorical data is used as the main example to bring out some of the characteristics of the RR-subclasses, and investigate its use to regression and classification problems. Recently, Srivastava (1997) considered the problem of reduced-rank regression for classification or discrimination, but only for the Gaussian model. Hastie and Tibshirani (1996) also discuss the ideas of reduced rank regression to discrimination problems, but in a larger framework involving mixture models. Gabriel (1998) and Aldrin (2000) are also recent works.

One model where the reduced-rank regression idea has been applied to non-Gaussian errors is the MLM. This was proposed and referred to as the *stereotype* model by Anderson (1984). However, in that paper and in subsequent papers by others, the reduced-rank regression idea was not explicitly stated in the framework presented below.

The aim of this paper is twofold. Firstly, we extend the reduced-rank concept to the VGLM and VGAM class. Secondly, we describe and motivate the reduced-rank idea applied to regression models for categorical data analysis, especially the MLM. We do this by elaborating on its connections to other statistical models such as neural networks, projection pursuit regression, linear discriminant analysis, canonical correspondence analysis and biplots.

An outline of this paper is as follows. In the remainder of this section we briefly review

VGLMs and VGAMs—further details can be found in Yee and Wild (1996). In Section 2 we propose reduced-rank regression for the VGLM class. In Section 3 we focus on the RR-MLM, and show how it relates to several other statistical methods. Section 4 gives two examples and Section 5 extends the reduced-rank concept to the nonparametric case. The article finishes with some brief notes on estimation and a discussion.

1.1 VGLMs and VGAMs

Suppose that for each of n individuals under study, a q-dimensional response vector $\boldsymbol{y}$ ($q \geq 1$) and a p-dimensional covariate vector $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is observed. VGLMs are defined as a model for which the conditional distribution of $\boldsymbol{Y}$ given $\boldsymbol{x}$ is of the form

$$f(\boldsymbol{y}|\boldsymbol{x}; \mathbf{B}) \;=\; h(\boldsymbol{y}, \eta_1, \ldots, \eta_M)$$

for some known function $h(\cdot)$, where $\mathbf{B} = (\boldsymbol{\beta}_1\, \boldsymbol{\beta}_2 \, \cdots \, \boldsymbol{\beta}_M)$ is $p \times M$ and

$$\eta_j \;=\; \eta_j(\boldsymbol{x}) \;=\; \boldsymbol{\beta}_j^T \boldsymbol{x} \;=\; \beta_{(j)0} + \beta_{(j)1}\, x_1 + \cdots + \beta_{(j)p}\, x_p \tag{1}$$

is the jth linear predictor. GLMs are a special case having only $M = 1$ linear predictor, and frequently M does not coincide with the dimension of $\boldsymbol{y}$. We have

$$\boldsymbol{\eta}(\boldsymbol{x}_i) \;=\; \begin{pmatrix} \eta_1(\boldsymbol{x}_i) \\ \vdots \\ \eta_M(\boldsymbol{x}_i) \end{pmatrix} \;=\; \boldsymbol{\eta}_i \;=\; \boldsymbol{\eta}_0 + \mathbf{B}^T \boldsymbol{x}_i \;=\; \boldsymbol{\eta}_0 + \begin{pmatrix} \boldsymbol{\beta}_1^T \boldsymbol{x}_i \\ \vdots \\ \boldsymbol{\beta}_M^T \boldsymbol{x}_i \end{pmatrix} \tag{2}$$

where $\boldsymbol{\eta}_0 = (\beta_{(1)0}, \cdots, \beta_{(M)0})^T$ is a vector of intercepts.

VGAMs provide additive-model extensions to VGLMs, that is, (1) is generalized to

$$\eta_j(\boldsymbol{x}) \;=\; \beta_{(j)0} + f_{(j)1}(x_1) + \cdots + f_{(j)p}(x_p), \qquad j = 1, \ldots, M, \tag{3}$$

a sum of smooth functions of the individual covariates, just as with ordinary GAMs (Hastie and Tibshirani, 1990). The η_j in (3) are referred to as additive predictors. For identifiability, each component function is centered, i.e., $E[f_{(j)k}] = 0$.

In practice, it is very useful to consider constraints-on-the-functions. For VGAMs, we have

$$\begin{aligned} \boldsymbol{\eta}(\boldsymbol{x}) \;&=\; \boldsymbol{\beta}_0 + \boldsymbol{f}_1(x_1) + \cdots + \boldsymbol{f}_p(x_p) \\ &=\; \mathbf{H}_0\, \boldsymbol{\beta}_0^* + \mathbf{H}_1\, \boldsymbol{f}_1^*(x_1) + \cdots + \mathbf{H}_p\, \boldsymbol{f}_p^*(x_p) \end{aligned} \tag{4}$$

where $\mathbf{H}_0, \mathbf{H}_1, \ldots, \mathbf{H}_p$ are known full-column rank *constraint matrices*, $\boldsymbol{f}_k^*$ is a vector containing a possibly reduced set of component functions and $\boldsymbol{\beta}_0^*$ is a vector of unknown intercepts. With no constraints at all, $\mathbf{H}_0 = \mathbf{H}_1 = \cdots = \mathbf{H}_p = \mathbf{I}_M$ and $\boldsymbol{\beta}_0^* = \boldsymbol{\beta}_0$ $(= \boldsymbol{\eta}_0)$. Like the $\boldsymbol{f}_k$, the $\boldsymbol{f}_k^*$ are centered. For VGLMs, the $\boldsymbol{f}_k$ are linear so that

$$\mathbf{B}^T \;=\; \left(\mathbf{H}_0\boldsymbol{\beta}_0^* \;\; \mathbf{H}_1\boldsymbol{\beta}_1^* \;\; \cdots \;\; \mathbf{H}_p\boldsymbol{\beta}_p^*\right).$$

VGLMs are usually estimated by maximum likelihood estimation by Fisher scoring or Newton-Raphson. At its heart, the quantity

$$Q \;=\; \sum_{i=1}^{n} \left(\boldsymbol{z}_i - \mathbf{H}_0\,\boldsymbol{\beta}_0^* - \sum_{k=1}^{p} \mathbf{H}_k\,\boldsymbol{\beta}_k^*\,x_{ik}\right)^T \mathbf{W}_i \left(\boldsymbol{z}_i - \mathbf{H}_0\,\boldsymbol{\beta}_0^* - \sum_{k=1}^{p} \mathbf{H}_k\,\boldsymbol{\beta}_k^*\,x_{ik}\right) \tag{5}$$

is minimized at each iteration, where $\boldsymbol{z}_i$ is an adjusted dependent vector and $\mathrm{Var}(\boldsymbol{z}_i) = \mathbf{W}_i^{-1}$ is a positive definite weight matrix. VGAMs can be estimated by a modified vector backfitting algorithm involving vector smoothing.

2 Reduced-rank Regression and VGLMs

For various VGLM models, the matrix of regression coefficients (2) may be very large for the data on hand, for example, the MLM for categorical data specifies the probability of a $(M+1)$-level factor Y falling into class j as

$$\Pr(Y = j|\boldsymbol{x}) \;=\; \frac{\exp\{\eta_j(\boldsymbol{x})\}}{\sum_{t=1}^{M+1} \exp\{\eta_t(\boldsymbol{x})\}}, \quad \eta_{M+1} \equiv 0, \quad j = 1, \ldots, M+1. \tag{6}$$

For many applications M and p may be large, for example, in the vowel data set considered later, there are $M+1 = 11$ symbols (Table 3) and $p = 10$ explanatory variables. This produces $10 \times (1+10) = 110$ regression coefficients for only 528 training observations. Clearly, it is advantageous to reduce the number of parameters. There have been several proposals to attack this problem, for example, Boser and Guyon (1992) used support vector machines to fit optimal margin hyperplanes without using an exorbitant number of parameters. In this paper we propose using reduced-rank regression. The concept is simply to replace $\mathbf{B}$ by the approximation

$$\mathbf{B} \;=\; \mathbf{C}\,\mathbf{A}^T \tag{7}$$

where $\mathbf{C} = (\boldsymbol{c}_{(1)}\,\boldsymbol{c}_{(2)} \cdots \boldsymbol{c}_{(r)})$ is $p \times r$, $\mathbf{A} = (\boldsymbol{a}_{(1)}\,\boldsymbol{a}_{(2)} \cdots \boldsymbol{a}_{(r)})$ is $M \times r$ and r $(\leq \min(M,p))$ is known as the rank ($\mathbf{A}$ and $\mathbf{C}$ are both of full-column rank). When applied to the VGLM class we call the result a *reduced-rank VGLM* (RR-VGLM).

There are a number of reasons why the reduced-rank idea is important. Firstly, if $r \ll p$ then a more parsimonious model results. The resulting number of parameters is often much less than the full model (the difference is $(M-r)(p-r)$, which is substantial when $r \ll \min(M,p)$. For example, for the vowel example, this is a saving of $(10-2)\times(10-2) = 64$ parameters for a rank-2 model.) Secondly, the reduced-rank approximation provides a vehicle for a low dimensional view of the data—see later. Thirdly, it allows for the flexible nonparametric generalizations proposed in Section 5. A fourth advantage of the reduced-rank concept is that it is readily interpretable. One can think of $\boldsymbol{\nu}_i = (\boldsymbol{c}_{(1)}^T \boldsymbol{x}_i, \ldots, \boldsymbol{c}_{(r)}^T \boldsymbol{x}_i)^T$ as a vector of latent derived variables—linear combinations of the original predictor variables that give more explanatory power. They often can be thought of as a proxy for some underlying variable behind the mechanism of the process generating the data. For some models, such as the cumulative logit model (McCullagh, 1980), this argument is natural and well-known. In fields such as plant ecology the idea is an important one. One can think of the role of $\mathbf{C}$ as choosing the 'best' predictors from a linear combination of the original predictors, and $\mathbf{A}$ as regression coefficients of these new predictors. Other motivations for reduced-rank regression, for the MLM specifically, are given in the next section.

Unfortunately, much of the standard theory of reduced-rank regression (see, for example, Reinsel and Valu (1998)) and its ramifications are not applicable to RR-VGLMs in general. This is mainly due to the $\mathbf{W}_i$ in (5) being unequal. If they actually were equal then $\mathbf{A}$ and $\mathbf{C}$ could simply be estimated by a generalized singular value decomposition (SVD), i.e., the Eckart and Young (1936) theorem in the metric $\mathbf{W}_i^{-1}$. Another complication in inference of RR-VGLMs is that the solution to a lower rank problem is not nested within a higher rank problem. That is, the range space of $\hat{\mathbf{B}} = \sum_{s=1}^{t} \hat{\boldsymbol{c}}_{(s)} \hat{\boldsymbol{a}}_{(s)}^T$ is not in the subspace of $\hat{\mathbf{B}} = \sum_{s=1}^{r} \hat{\boldsymbol{c}}_{(s)} \hat{\boldsymbol{a}}_{(s)}^T$ for $r > t$. See Anderson (1984) for more details. In this article, we find it convenient to write

$$\boldsymbol{\eta}_i \;=\; \boldsymbol{\eta}_0 + \mathbf{A}\,\mathbf{C}^T \boldsymbol{x}_i \;=\; \boldsymbol{\eta}_0 + \mathbf{A}\,\boldsymbol{\nu}_i. \tag{8}$$

The vector $\boldsymbol{\eta}_0$ containing the intercepts is usually left alone.

The factorization (7) is not unique because $\boldsymbol{\eta}_i = \boldsymbol{\eta}_0 + \mathbf{A}\,\mathbf{M}\,\mathbf{M}^{-1}\,\boldsymbol{\nu}_i$ for any nonsingular matrix $\mathbf{M}$. The following lists some common uniqueness constraints.

1. Restrict $\mathbf{A}$ to the form

$$\mathbf{A} \;=\; \begin{pmatrix} \mathbf{I}_r \\ \widetilde{\mathbf{A}} \end{pmatrix}, \quad \text{say.} \tag{9}$$

(Actually, it may be necessary to represent $\mathbf{I}_r$ in rows other than the first r.)

2. Another normalization of $\mathbf{A}$, which makes direct comparisons with other statistical methods possible, is based on the SVD

$$\mathbf{A}\mathbf{C}^T = (\mathbf{U}\mathbf{D}^{\alpha})(\mathbf{D}^{1-\alpha}\mathbf{V}^T) \tag{10}$$

for some specified $0 \leq \alpha \leq 1$. In this paper we follow Gabriel (1998) and choose $\alpha = \frac{1}{2}$ as it scales both sides symmetrically.

3. Choose $\mathbf{M}$ so that $\widehat{\mathrm{Var}}(\mathbf{M}\,\hat{\boldsymbol{\nu}})$ is diagonal, i.e., the latent variables are uncorrelated.

4. For the stereotype model described in the next section, we can choose $\mathbf{M}$ so that the columns of $\mathbf{C}$, $\boldsymbol{c}_{(j)}$, are orthogonal with respect to the within-group covariance matrix $\mathbf{W}$—this type of normalization is similar to linear discriminant analysis (LDA).

3 The Stereotype Model

Regression models for categorical responses naturally produce a large number of parameters, and therefore are prime candidates for reduced-rank regression. Indeed, special cases of this have already been made, for example, the nonproportional odds model can be written

$$\text{logit}\ \Pr(Y \leq j|\boldsymbol{x}) \ = \ \eta_j \ = \ \beta_{(j)0} + \boldsymbol{\beta}_j^T\boldsymbol{x}, \quad j = 1, \ldots, M.$$

Under the parallelism assumption $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_M$ ($= \boldsymbol{\beta}$, say) this becomes the well-known proportional odds model (McCullagh, 1980). The parallelism assumption corresponds to the reduced-rank model (7) but with a *known* $\mathbf{A} = \mathbf{1}_M$ and (unknown) $\mathbf{C} = \boldsymbol{\beta}$, hence $r = 1$. The parallelism assumption may be applied to other categorical data models such as the adjacent-categories and continuation-ratio models. In the rest of this section we focus on the MLM and its relationships with other statistical methods.

Following from (6), we write the MLM as

$$\text{logit}\ \boldsymbol{p}(\boldsymbol{x}) \ = \ \begin{pmatrix} \log(p_1/p_{M+1}) \\ \vdots \\ \log(p_M/p_{M+1}) \end{pmatrix} \ = \ \boldsymbol{\eta}(\boldsymbol{x}) \ = \ \boldsymbol{\eta}_0 + \mathbf{B}^T\boldsymbol{x}. \tag{11}$$

Replacing $\mathbf{B}$ by $\mathbf{CA}^T$ gives the RR-MLM. Anderson (1984) described it as being suitable for *all* (ordered or unordered) categorical response variables and his one-dimensional stereotype model constrained $\boldsymbol{\beta}_j = \phi_j \boldsymbol{\beta}$, $j = 1, \ldots, M$, so that

$$\Pr(Y = j|\boldsymbol{x}) = \frac{\exp\{\alpha_j + \phi_j \boldsymbol{\beta}^T \boldsymbol{x}\}}{\sum_{k=1}^{M+1} \exp\{\alpha_k + \phi_k \boldsymbol{\beta}^T \boldsymbol{x}\}}, \quad \alpha_{M+1} = \phi_{M+1} = 0, \quad j = 1, \ldots, M+1. \tag{12}$$

Under this model $\beta_k \phi_j$ represents the log odds ratio for $Y = j$ versus $Y = M + 1$ per unit increase in x_k. The parameters ϕ_j, which may be known or unknown, can be thought of as the score attached to outcome j. If unknown then they are estimated with $\phi_1 = 1$ being a common identifiability constraint. It can easily be seen that (12) corresponds to the rank–1 RR-MLM as

$$\mathbf{B} \equiv (\boldsymbol{\beta}_1 \boldsymbol{\beta}_2 \cdots \boldsymbol{\beta}_M) = (\boldsymbol{\beta} \;\; \phi_2 \boldsymbol{\beta} \;\; \cdots \;\; \phi_M \boldsymbol{\beta}) = \boldsymbol{\beta} \cdot (1, \phi_2, \ldots, \phi_M)$$

which is of the form $\mathbf{CA}^T$. Anderson also suggested higher dimensional stereotype models ($r > 1$), in particular, his two-dimensional stereotype model used the constrained form (9).

Unfortunately, Anderson died before his 1984 paper appeared and little or no subsequent work has appeared since. The untapped potential of the RR-MLM was recognized by Greenland (1994) who attempted to promulgate its use in medical statistics, and argued for it to have equal coverage with other ordinal regression models. Greenland considered only the $r = 1$ case, and fitted stereotype models by performing a series of GLM fits in which $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\phi}} = (\hat{\phi}_2, \ldots, \hat{\phi}_M)^T$ were alternately held fixed while the other was estimated. He call this the *alternating algorithm*, and noted that his experience was that it converged rapidly and reliably when started with scores derived from the unordered MLM. We adopt this algorithm too—see Section 6.

3.1 Connection With Neural Networks

RR-VGLMs can be motivated by their connection with neural networks. As an example, Figure 1 shows the network diagram representation of a RR-MLM. One has

$$\begin{aligned}
\nu_t &= \boldsymbol{c}_{(t)}^T \boldsymbol{x}, \quad t = 1, \ldots, r, \\
\eta_j &= \eta_{(j)0} + \boldsymbol{a}_{(j)}^T \boldsymbol{\nu}, \quad j = 1, \ldots, M, \\
p_j &= \sigma_j(\boldsymbol{\eta}), \quad j = 1, \ldots, M,
\end{aligned}$$

where $\sigma_j(\boldsymbol{\eta}) = \exp(\eta_j)/\{1 + \sum_s \exp(\eta_s)\}$ is the multiple inverse logit activation function. The hidden layer of only r ν_t's can be thought of as a restricted bottle-neck through the network. The classical reduced-rank model comes about by replacing $\sigma_j(\boldsymbol{\eta})$ by the identity function.

Reinsel and Valu (1999, p.17) note that an appealing physical interpretation of the reduced-rank model was offered by Brillinger (1969). This is in regard to a situation where the p component vector $\boldsymbol{x}_i$ represents information which is to be used to send a message $\boldsymbol{y}_i$ having M components ($M \leq p$) but such a message can only be transmitted through r channels ($r \leq M$). Thus, $\mathbf{C}^T\boldsymbol{x}_i$ acts as a code and on receipt of the code, forms the vector $\mathbf{A}\mathbf{C}^T\boldsymbol{x}_i$ by premultiplying the code, which it is hoped, would be as close as possible to the desired $\boldsymbol{y}_i$.

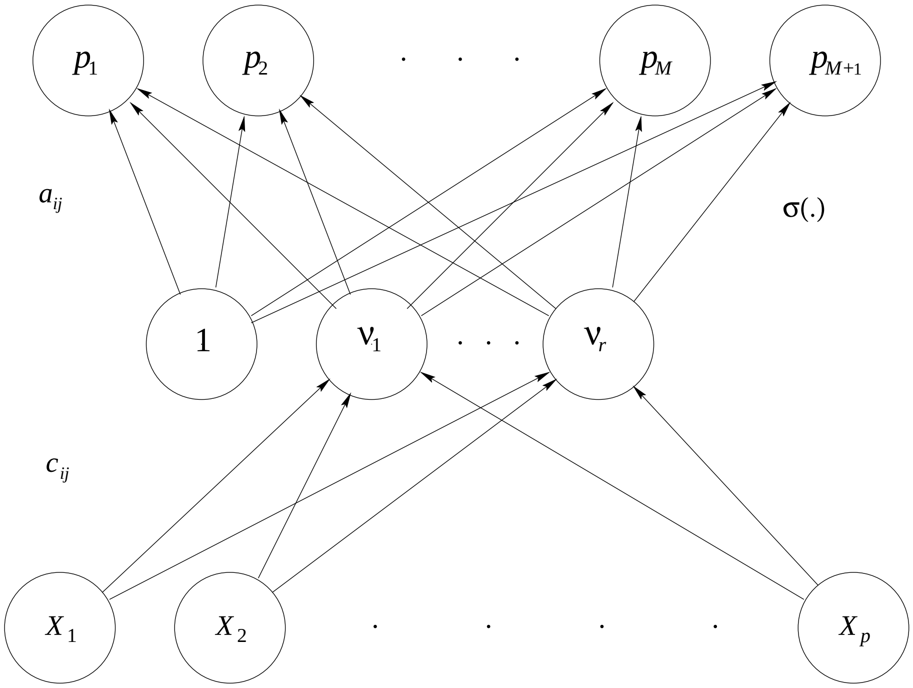

Figure 1: Network diagram of a RR-MLM. The "1" denotes a bias or intercept term.

3.2 Connection With Fisher-Rao Linear Discriminant Analysis

In this section we consider the consequences of the two assumptions

A. $\boldsymbol{X} \sim N_p(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$ in class j, and

B. $\dim\left(\text{span}\left\{\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_{M+1}\right\}\right) = r$, i.e., the group means lie in a r-dimensional linear subspace.

Bayes formula applied to a MLM gives

$$\log \frac{p_j(\boldsymbol{x})}{p_{M+1}(\boldsymbol{x})} = \log \frac{g_j(\boldsymbol{x})\, \pi_j}{g_{M+1}(\boldsymbol{x})\, \pi_{M+1}} = c_j + \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{M+1}\right)^T \boldsymbol{\Sigma}^{-1} \boldsymbol{x} \qquad (13)$$

where $c_j = \log(\pi_j/\pi_{M+1}) - \frac{1}{2}\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{M+1}\right)^T \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\mu}_j + \boldsymbol{\mu}_{M+1}\right)$, π_j are prior probabilities and g_j are density functions, therefore

$$\operatorname{logit} \boldsymbol{p}(\boldsymbol{x}) = \boldsymbol{\eta}(\boldsymbol{x}) \;=\; \begin{pmatrix} c_1 \\ \vdots \\ c_M \end{pmatrix} + \begin{pmatrix} \boldsymbol{\mu}_1^T - \boldsymbol{\mu}_{M+1}^T \\ \vdots \\ \boldsymbol{\mu}_M^T - \boldsymbol{\mu}_{M+1}^T \end{pmatrix} \boldsymbol{\Sigma}^{-1}\boldsymbol{x}. \tag{14}$$

The matrix containing the $\boldsymbol{\mu}_j$'s is of rank r, therefore (14) is a RR-MLM. Anderson (1984) was the first to note this feature. From a practical point of view, Assumption A. is often unrealistic, for example, when some of the predictors are discrete. However, this assumption is unnecessary for the RR-MLM! This analogy is akin to the comparison between LDA with normal populations and logistic regression (Efron, 1975).

In LDA with $M+1$ classes, we can choose a subspace of rank $r < M$ that maximally separates the class centroids. In fact, LDA can be derived as the maximum likelihood method for normal populations with different means and common covariance matrix (A.). Campbell (1984) (see also Hastie and Tibshirani (1996)) show that reduced-rank LDA can be viewed as a restricted Gaussian maximum likelihood solution (A. and B.). These results suggest that, providing A. and B. hold, fitting a MLM to the discriminant variables from a LDA should give similar results to a RR-MLM applied to the original covariates $\boldsymbol{x}_i$. That is, choosing $\mathbf{C}$ from a LDA and $\mathbf{A}$ from a MLM should be similar to a RR-MLM.

However, even if the two assumptions do not hold, this approach can still be recommended because it allows a method of choosing r. This is because the MLM allows one to check for the significance of each regression coefficient via Wald t-statistics. One fits a MLM to the first r linear discriminant variables and check for the statistical significance of the regression coefficients. Increase r until the regression coefficients of the $(r+1)$th linear discriminant variables are not statistically significant. This approach can also be advantageous over the RR-MLM for classification problems. Here, maximizing the likelihood does not always lead to the best model—a procedure such as LDA which separates out the groups first may provide better performance. The hybrid method has the advantage of both a regression and classification tool.

3.2.1 Canonical Regression Models

McCullagh, in the discussion section of Anderson (1984), called the stereotype model a "canonical regression model". He showed that the rank-1 stereotype model may be written

$$\log p_{ij} \;=\; \alpha_i + \beta^*_{0j} - \phi_j \, \boldsymbol{\beta}^T \boldsymbol{x}_i, \quad j = 1, \ldots, M; \; i = 1, \ldots, n,$$

and because α_i is nuisance, the analogous multivariate linear model is

$$E(Y_{ij}) \;=\; \mu_{ij} \;=\; \beta^*_{0j} - \phi_j \, \boldsymbol{\beta}^T \boldsymbol{x}_i. \tag{15}$$

The linear combination of the Y's having maximal regression on $\mathbf{X}$ is $\sum \phi_j \boldsymbol{Y}_j$, also called the linear discriminant function. Thus (15) implies that there is a single non-zero canonical root with vectors $\boldsymbol{\phi}$ and $\boldsymbol{\beta}$. In a similar way, models with several canonical roots can be constructed, corresponding to higher-rank stereotype models.

3.3 Connection With Canonical Correspondence Analysis

The idea of approximating a table of counts ($\mathbf{Y}$) by a lower-rank approximation is the basis of correspondence analysis (CA). When $\mathbf{Y}$ is supplemented by covariate information $\mathbf{X}$, the technique is referred to as canonical correspondence analysis (CCA). CCA as proposed by ter Braak (1986) has become very popular in plant ecology through its implementation in the software package CANOCO. CCA is CA subject to the latent variable being a linear combination of the environmental variables—constrained ordination.

In the terminology of ter Braak (1986), the RR-MLM is a redundancy analysis method for compositional (relative abundance) data based on GLMs which is fitted by maximum likelihood estimation. It has both a 'linear' and 'unimodal' face because, for Species j's optimum u_j, $\eta_{ij} = \eta^*_{(j)0} - \frac{1}{2}(\nu_i - u_j)^2$ in (6) [also known as Ihm and van Groenewoud's (1984) Model B] is equivalent to $\eta_{ij} = \eta^{**}_{(j)0} + \beta_j \, \nu_i$ where $\beta_j = u_j$, i.e., is linear in the latent variable ν_i.

That the RR-MLM is *a* method for CCA is easy to see from Equations (7) and (11), and ter Braak and Smilauer (1998, p.58) earlier noted this. The classical reduced-rank regression problem is a RR-VGLM having Gaussian errors (identity link for the $\boldsymbol{\eta}$), and we have written a VGAM family function called `rrGaussian()` implementing this.

Hastie and Little (1987), who proposed a new method for modelling compositional data, solved a similar model to the RR-MLM. It differed by having $\mathbf{C} = \mathbf{I}_r$, and *unknown* $\boldsymbol{x}_i$ which were estimated. Their work showed that the logit transformation (11) of the MLM, which is unconstrained, did not have the "arch" or "horseshoe" effect that is present in such methods such as correspondence analysis. We can infer from this that the RR-MLM, when estimated by maximum likelihood, does not suffer (or suffers less) from the arch effect, which plagues other methods such as CCA.

Sabatier *et al.* (1989) and Lebreton *et al.* (1991) showed that CCA is a weighted form of reduced-rank regression (also known as redundancy analysis and principal component analysis with respect to instrumental variables (Rao, 1964)). For a rank-r CCA, the quantity

$$\sum_{i=1}^{n} \sum_{j=1}^{M+1} y_{i+}\, y_{+k} \left\{ \frac{y_{ij}\, y_{++}}{y_{i+}\, y_{+k}} - 1 - \sum_{k=1}^{p} \beta_{kj}\, x_{ik} \right\}^2 \tag{16}$$

is minimized, subject to the constraint $\beta_{kj} = a_{k1} c_{j1} + \cdots + a_{kr} c_{jr}$. It is very easy to see that this equation fits neatly within Equation (22). We have fitted CCA's by preprocessing the response and applying our S-PLUS family function `rrGaussian()`.

3.3.1 CCA and LDA

Ter Braak and Verdonschot (1995) review the relationship between CCA and Fisher's LDA, and this section is heavily based on that. Chessel *et al.* (1987) and Lebreton *et al.* (1988) recognized the formal equivalence between CCA and LDA on reformatted data (see also Takane *et al.* (1991)), in fact, CCA is a generalization of LDA. The derivation of CCA is very similar to that of LDA because it finds canonical variates, linear combinations of the features, that show maximum discrimination among groups, or equivalently, that maximally separate the groups. Replacement of "groups" by "niches of species" in the above yields similar definitions for LDA ad CCA.

The main difference between CCA and LDA is that the unit of statistical analysis in LDA is the individual, whereas it is the site in CA. However, if the species data for CCA are counts of individuals at sites, then one can create a new inflated data matrix based on each individual counted. LDA carried out on this new type of data is identical to CCA. There are minor differences in the default output, for example, if the eigenvalue of CCA is λ, then the corresponding eigenvalue in LDA is $\lambda/(1-\lambda)$, and the scores of the species and sites are linearly related. The scaling of the scores as used in LDA is a variant of Hill's scaling in (C)CA.

Recent work by Mu Zhu, a graduate student at Stanford University, has also established the equivalence of CCA and LDA, and generalized it to allow for non-Gaussian densities for the species as classes.

4 Two Examples

4.1 A Car Data Example

To illustrate the basic RR-MLM we consider the S-PLUS data frame `car.all` and model the country of manufacture with respect to the car's length, width, weight, engine horsepower, engine displacement and price. We have $M = 3$ and $p = 6$ variables, which were centred and scaled prior to analysis. The subset of four countries considered were those providing the most counts; these were $Y = 1 =$ Germany, $Y = 2 =$ Japan, $Y = 3 =$ Japan/USA, $Y = 4 =$ USA. There were 10, 31, 9 and 39 ($n = 89$) vehicles in these groups respectively after missing values were deleted. The last category, American vehicles, was chosen as baseline. Table 2 provides the estimated regression coefficients for each of the RR-MLMs. The crux of the interpretation lies in the fact that $\beta_{(j)k}$ in (11) represents the log odds ratio for $Y = j$ versus $Y = M + 1$ per unit increase in x_k. It may be seen that from the full MLM that length, engine displacement and price are the most significant. The model suggests:

1. Keeping all other covariates fixed, increasing the price raises the likelihood of the car being German the most, followed by Japanese and Japanese/American, relative to American vehicles.

2. Keeping all other covariates fixed, increasing the engine displacement increases the likelihood of the car being American, relative to the other countries.

The first is not surprising as overseas cars (especially European ones) tend to be more expensive in USA. The second is not surprising considering the price is fixed. Of the 21 regression coefficients, 9 are 'approximately statistically significant' as indicated by a t value with absolute value greater than 2.

Table 1 gives the estimates of the rank-1 and rank-2 RR-MLMs. The $\hat{\mathbf{A}}$ and $\hat{\mathbf{C}}$ have been normalized by the SVD method with $\alpha = \frac{1}{2}$ and the $\hat{\boldsymbol{\nu}}$ uncorrelated. Several interesting features can be seen from the rank-1 model. Firstly, it can be seen that the latent variable $\hat{\nu}_1$ is mainly a contrast between price and engine displacement. Secondly, the elements of $\mathbf{A}$ are monotonic and positive. Thus as the difference between price and engine displacement increases so does the likelihood of the car being German relative to USA. This is followed by Japan and Japan/USA. This essentially captures the interpretation of the full model described above.

For the rank-2 RR-MLM, the interpretation of $\hat{\nu}_1$ varies only slightly from the rank-1 model because the first columns of $\hat{\mathbf{A}}$ and $\hat{\mathbf{C}}$ have changed little. The second latent variable is mainly a contrast between engine horsepower with price. The order of magnitude of values in the second column of $\mathbf{A}$ is much smaller than the first column, implying the much larger effect of the first latent variable compared to the second.

Another feature that can be seen is how the RR-MLMs approximate the full MLM $\mathbf{B}$ more as the rank increases. These concentrate on the most significant variables first.

Figure 2 plots the two uncorrelated latent variables $\hat{\nu}_1$ and $\hat{\nu}_2$. By visually fitting a curve through the centroids of the four groups, there appears to be the ordering: Germany, Japan, Japan/USA, and USA. The plot thus successfully captures a relationship between the groups on a two-dimensional subspace, and agrees with intuition. In particular, Japan/USA lies in a tight cluster between Japan and USA. Manual inspection of all the pairwise scatterplots (15 in total) of the covariates is laborious and none seem to provide such a satisfactory ordination as Fig. 2.

A biplot based on the coefficents in Table 1 shows the first axis predominates. Fig. 3 shows a more presentable (scaled) biplot where $\hat{\mathbf{A}}$ and $\hat{\mathbf{C}}$ have been multiplied and divided by some scalar respectively. It is easy to project each country to the variables to get an approximation to the coefficients in the large $\mathbf{B}$ matrix, for example, the coefficients of price for the group Germany is large and positive.

What r is best? We also fitted MLMs to a sequence of linear discriminant variables (RR-MLM-LDAs, say). For this, inspection of the Wald t-statistics of the coefficients for the rank-r RR-MLM-LDAs suggest $r = 2$ to be best: for $r = 2$, five of the six regression coefficients are significant ($|t| > 2$), whereas none for the third linear discriminant variable are so.

In summary, this simple example illustrates how RR-MLMs can provide substantial insight to a data set and that the output of a RR-MLM can be very interpretable.

Table 1: Rank-1 and 2 RR-MLMs fitted to the standardized car data: $\hat{\mathbf{C}}$ is above and $\hat{\mathbf{A}}$ below. The standardization is $\alpha = \frac{1}{2}$ and uncorrelated $\hat{\boldsymbol{\nu}}$.

Rank	1	2	
Length	−0.373	−0.388	−0.535
Width	−0.200	−0.250	−0.554
Weight	0.253	0.308	0.293
HP	0.108	0.014	1.203
Disp.	−0.946	−0.896	0.409
Price	0.905	0.938	−1.208
$\log(p_1/p_4)$	9.568	10.432	−0.824
$\log(p_2/p_4)$	7.415	7.669	0.919
$\log(p_3/p_4)$	5.189	5.400	0.453
$\log(p_4/p_4)$	0.000	0.000	0.000

Table 2: Estimated RR-MLM regression coefficients $\hat{\mathbf{B}}$ for the standardized car data. The rank-3 corresponds to an ordinary multinomial logit model. Deviances are 94.91, 105.179, 114.906 respectively. Asterisks in the rank-3 model are coefficients whose $|t|$-values are greater than the number of asterisks.

Rank	Variable	$\log(p_1/p_4)$	$\log(p_2/p_4)$	$\log(p_3/p_4)$
3	(Intercept)	-3.284^{**}	0.206	-3.344^{**}
	Length	-3.681^{**}	-3.378^{**}	0.896
	Width	-1.954^{*}	-2.246^{*}	-3.090^{*}
	Weight	2.777^{*}	2.420^{*}	2.961^{*}
	HP	-0.979	0.997	0.940
	Disp.	-8.654^{***}	-5.739^{**}	-9.066^{**}
	Price	10.146^{***}	5.599^{**}	2.232
2	(Intercept)	-3.371	0.169	-0.343
	Length	-3.607	-3.467	-2.337
	Width	-2.153	-2.428	-1.602
	Weight	2.972	2.631	1.796
	HP	-0.847	1.211	0.620
	Disp.	-9.684	-6.495	-4.653
	Price	10.780	6.084	4.519
1	(Intercept)	-2.674	0.334	-0.299
	Length	-3.570	-2.766	-1.936
	Width	-1.910	-1.480	-1.036
	Weight	2.418	1.874	1.311
	HP	1.032	0.799	0.559
	Disp.	-9.054	-7.016	-4.910
	Price	8.661	6.712	4.697

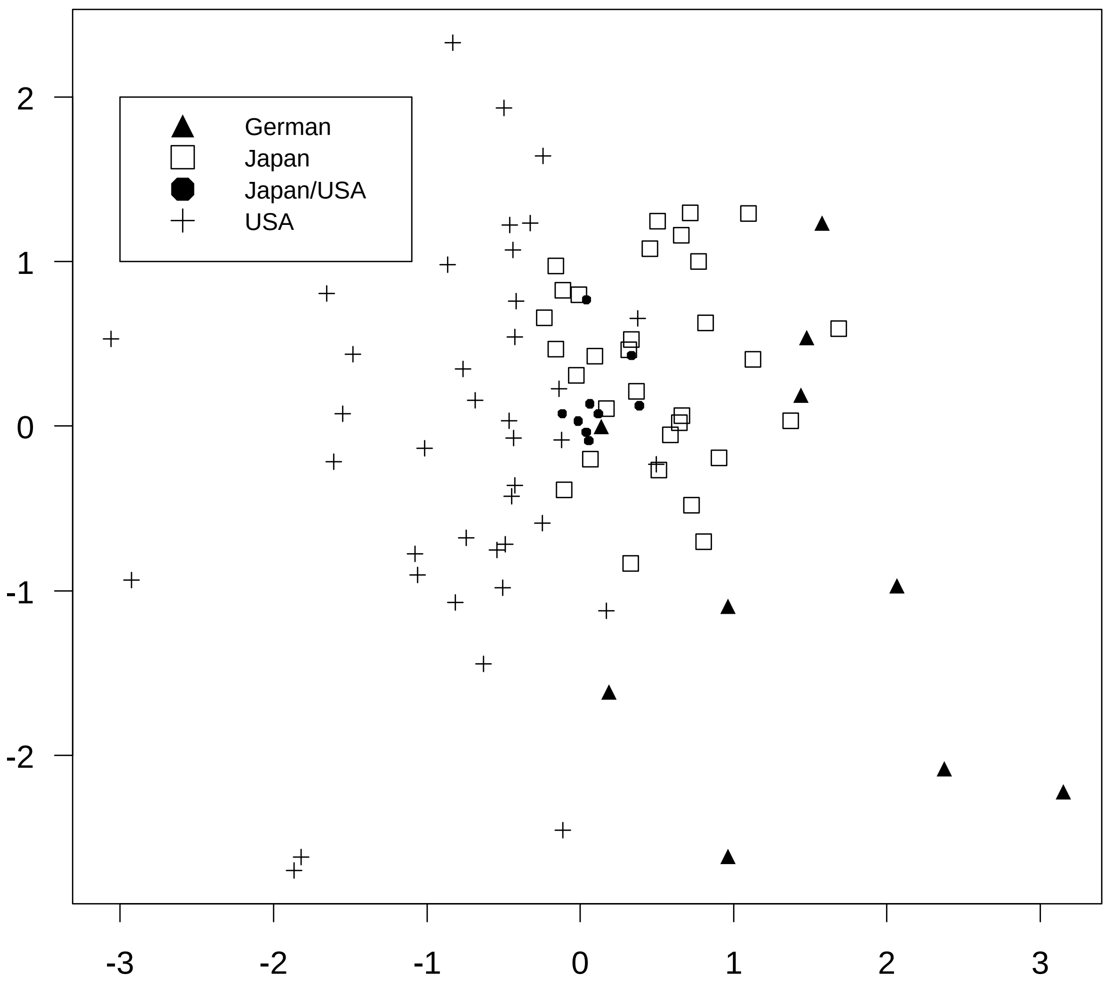

Figure 2: Scatterplot of $\hat{\nu}_1$ (horizontal axis) and $\hat{\nu}_2$ for the car data (Rank-2 RR-MLM).

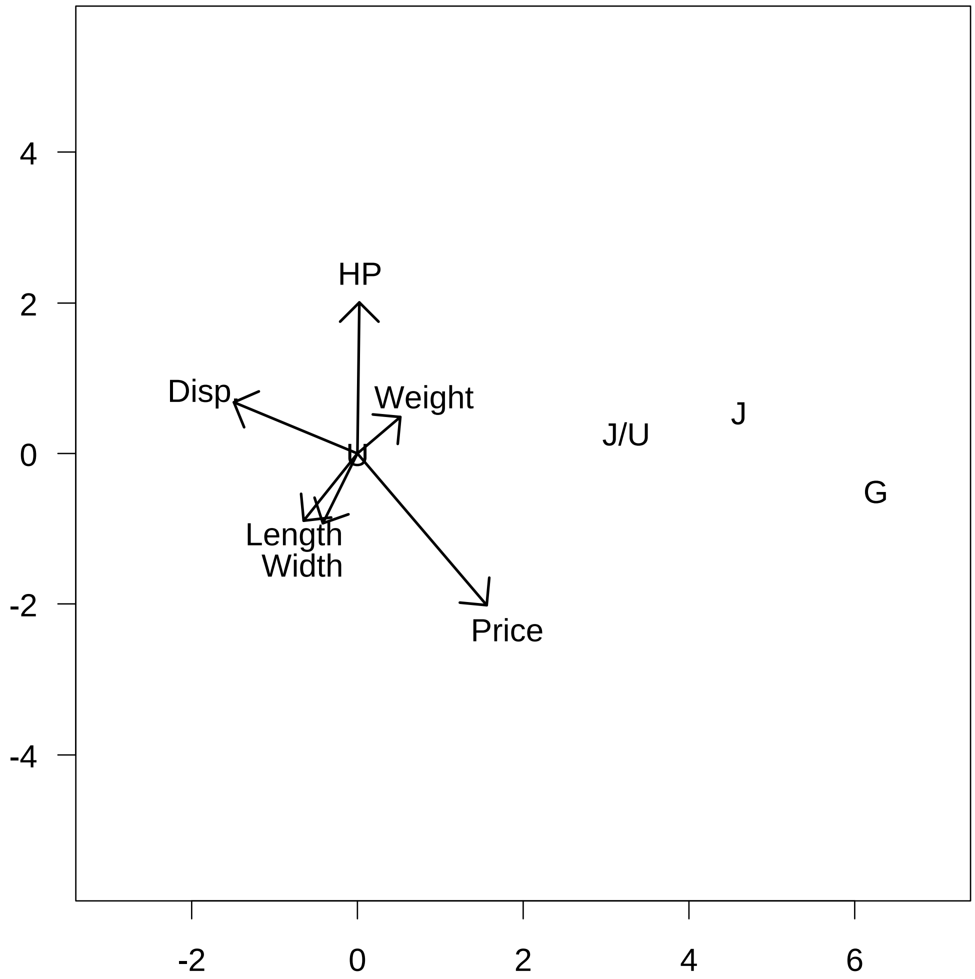

Figure 3: Biplot for the cars data: `G` = Germany, `J` = Japan, `J/U` = Japan/USA, `U` = USA. The arrows and points have been scaled for presentation.

4.2 A Vowel Recognition Data Example

Hastie *et al.* (1994) consider a vowel recognition data set involving 11 symbols (Table 3) produced from 8 speakers who had 6 replications each. The training data comprises 528 observations and 10 input features based on digitized utterances, and the test data consists of 462 observations. We scaled the covariates prior to the analysis.

The sequence of reduced-rank models for $r = 1$ up to the full model are plotted in Fig. 6. The last symbol (`3:`) was used as the baseline cell. Without having formal inference procedures, it appears a rank-2 model looks quite good. For this model a scatterplot of the $\hat{\nu}_1$ versus $\hat{\nu}_2$, with convex hulls and symbols distinguishing the vowels, is given in Fig. 7. While there is a large degree of overlap between the groups, the groups are nevertheless well-defined, and the amount of overlap explains the large error rates obtained.

Figure 5 shows a biplot of the Rank-2 fit, obtained by superimposing the rows of $\mathbf{A}$ as groups and $\mathbf{C}$ as arrows together. It may be seen that the general orientation of the biplot matches that of Fig. 7; the position of the vowels looks like the sample mean of the $\hat{\nu}_j$ of each vowel. It may be seen that `c:`, `U` and `u:` form a cluster of vowels, which suggests a similarity of effects of the variables on these responses. The variables, represented by arrows, appear to have two clusters: Variables 3 to 6 and Variables 7 to 10. One can apply the usual inner-product interpretation, for example, `A` and Variable 2 are approximately perpendicular, therefore the corresponding coefficient of $\mathbf{B}$ should be approximately zero.

Biplots for the RR-MLM require some special treatment. Writing $\mathbf{C}^T = (\boldsymbol{c}_1 \, \boldsymbol{c}_2 \cdots \boldsymbol{c}_p)$ and $\mathbf{A}^T = (\boldsymbol{a}_1 \, \boldsymbol{a}_2 \cdots \boldsymbol{a}_M)$, one has effectively $\boldsymbol{a}_{M+1} = \mathbf{0}$, the origin, because the $(M+1)$th level is baseline. To compare two non-baseline levels, one has

$$\log \frac{p_s(\boldsymbol{x})}{p_t(\boldsymbol{x})} = \log \frac{p_s(\boldsymbol{x})}{p_{M+1}(\boldsymbol{x})} - \log \frac{p_t(\boldsymbol{x})}{p_{M+1}(\boldsymbol{x})} = \boldsymbol{\eta}_{(s)0} - \boldsymbol{\eta}_{(t)0} + \sum_{k=1}^{p} (\boldsymbol{a}_s - \boldsymbol{a}_t)^T \, \boldsymbol{c}_k \, x_k.$$

That is, all the $\boldsymbol{a}_j$ coordinates need to be shifted by $-\boldsymbol{a}_t$ so that $\boldsymbol{a}_t$ is positioned at the origin. Then the usual inner product interpretation between $\mathbf{A}$ and $\mathbf{C}$ may be applied. For example, comparing `E` with `Y` shifts `E` near `i` and `I`, which is nearly orthogonal with `x1`. Hence $\log\{P(Y = \mathtt{E}|\boldsymbol{x})/P(Y = \mathtt{Y}|\boldsymbol{x})\}$ should not depend very much on X_1.

The error rates for various rank RR-MLMs and LDAs fitted to both the training and test data are given in Table 4 and plotted in Fig. 8. As expected, for both methods, the error rate generally decreases on the training data as the rank increases. When applied to the test data, the most parsimonious models both seem to be of rank-2, where an error rate of around 50% is achieved. Although error rates in this example appear high, this is not poor compared to the null error rate of $10/11 = 91\%$ due to the large number of classes.

The rank 2 approximation to $\hat{\mathbf{B}}$ (described in Section 6.3) was computed for comparison with the RR-MLM solution. This rank 2 approximation gave $D = 1709.688$ (cf. $D = 1052.04$ for the proper RR-MLM solution) which is substantially inferior.

Table 3: Symbols in the vowel recognition data set.

Class	Vowel	Word	Class	Vowel	Word
1	`i`	heed	7	`O`	hod
2	`I`	hid	8	`C`	hoard
3	`E`	head	9	`U`	hood
4	`A`	had	10	`u`	who'd
5	`a`	hard	11	`3`	heard
6	`Y`	hud			

Table 4: Error rates (%) of various rank RR-MLMs and LDAs fitted to the vowel recognition data.

r	RR-MLM Training data	RR-MLM Test data	LDA Training data	LDA Test data
1	64.0	62.6	61.2	69.9
2	36.6	52.2	35.0	49.1
3	33.7	56.5	33.0	49.6
4	28.6	52.6	33.0	51.1
5	26.3	51.1	31.6	51.5
6	22.9	52.4	30.1	55.4
7	22.7	51.3	31.2	55.4
8	22.7	50.2	31.8	55.6
9	22.2	51.1	31.4	55.2
10	22.3	51.3	31.6	55.6

5 Reduced-rank VGAMs

There is a simple nonparametric extension to the RR-VGLM class described above. Since $\boldsymbol{\eta}$ in (8) can be written

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_0 + \begin{pmatrix} a_{11} \cdot (\boldsymbol{c}_{(1)}^T \boldsymbol{x}_i) \\ \vdots \\ a_{M1} \cdot (\boldsymbol{c}_{(1)}^T \boldsymbol{x}_i) \end{pmatrix} + \cdots + \begin{pmatrix} a_{1r} \cdot (\boldsymbol{c}_{(r)}^T \boldsymbol{x}_i) \\ \vdots \\ a_{Mr} \cdot (\boldsymbol{c}_{(r)}^T \boldsymbol{x}_i) \end{pmatrix},$$

then

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_0 + \begin{pmatrix} f_{(1)1}(\boldsymbol{c}_{(1)}^T \boldsymbol{x}_i) \\ \vdots \\ f_{(M)1}(\boldsymbol{c}_{(1)}^T \boldsymbol{x}_i) \end{pmatrix} + \cdots + \begin{pmatrix} f_{(1)r}(\boldsymbol{c}_{(r)}^T \boldsymbol{x}_i) \\ \vdots \\ f_{(M)r}(\boldsymbol{c}_{(r)}^T \boldsymbol{x}_i) \end{pmatrix} \tag{17}$$

is a nonparametric generalization. That is, instead of $\boldsymbol{\eta}$ being linear functions of additive latent variables, the $\boldsymbol{\eta}$ are additive functions of linear latent variables. This gives rise to the *Reduced-rank VGAMs* class (RR-VGAMs). Equation (17) is related to vector projection pursuit regression (see Section 7.1), therefore a common identifiability constraint is $\|\boldsymbol{c}_{(j)}\|^2 = 1$, $j = 1, \ldots, r$.

It should be noted that, in fact, there are alternative types of nonparametric RR-VGLMs. Here are some of these.

5.1 Variants Thereof

1. One can generalize $\boldsymbol{\nu} = \left(\sum_{k=1}^{p} c_{k1}\, x_k, \ldots, \sum_{k=1}^{p} c_{kr}\, x_k \right)^T$ in (8) to

$$\boldsymbol{\nu} = \sum_{k=1}^{p} \boldsymbol{f}_k^*(x_k), \quad \text{i.e.,} \quad \boldsymbol{\eta} = \boldsymbol{\eta}_0 + \mathbf{A} \sum_{k=1}^{p} \boldsymbol{f}_k^*(x_k), \tag{18}$$

   where $\boldsymbol{f}_k^*(x_k) = (f_{(1)k}^*(x_k), \ldots, f_{(r)k}^*(x_k))$ is an r-vector of smooth arbitrary functions of x_k determined by the data. In this variant, instead of the latent variables being linear in the covariates, they are simply assumed additive in the covariates. Unfortunately, this variant is less flexible and has slow convergence.

2. The following takes linear combinations of a nonlinear transformation of (linear) latent

variables:

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_0 + \mathbf{A} \begin{pmatrix} f_1(\boldsymbol{c}_{(1)}^T \boldsymbol{x}_i) \\ \vdots \\ f_r(\boldsymbol{c}_{(r)}^T \boldsymbol{x}_i) \end{pmatrix}. \tag{19}$$

This is a RR-VGAM subject to $\boldsymbol{f}_t(\boldsymbol{c}_{(t)}^T \boldsymbol{x}_i) = \boldsymbol{a}_{(t)}\, f_t(\boldsymbol{c}_{(t)}^T \boldsymbol{x}_i)$, $t = 1, \ldots, r$.

3. This model was proposed by Hastie and Tibshirani (unpublished mss.)

$$\boldsymbol{\eta}_i = \boldsymbol{\eta}_0 + \mathbf{B}^T \begin{pmatrix} f_1(x_{i1}) \\ \vdots \\ f_p(x_{ip}) \end{pmatrix}. \tag{20}$$

It replaces the covariates x_k by smooth functions of them. The model falls within the constraints-on-the-functions framework in Section 1.1 except with unknown constraint matrices. One estimation process involves alternating between solving for $\mathbf{B}$ and the f_k's. A variant of (20) can be obtained by applying the reduced-rank idea to $\mathbf{B}$.

6 Estimation

In this section we briefly describe algorithms for the maximum likelihood estimation of RR-VGLMs and RR-VGAMs. We have considered/developed several, but only two have been found to be practical for us. These are the alternating algorithm and first-derivative algorithms. Each method is best applied within each Fisher scoring iteration, and the first-derivative algorithm can only be applied to RR-VGLMs. Of course, RR-MLMs could be fitted by neural networks, but this could be rather slow.

6.1 Alternating Algorithm

For RR-VGLMs, the alternating algorithm (also called the criss-cross method by Gabriel (1998)) for minimizing (5) involves fixing $\mathbf{A}$ and solving for $\boldsymbol{\nu}$, then keeping $\boldsymbol{\nu}$ fixed and solving for $\mathbf{A}$; this alternating can be continued until convergence in Q is achieved. It is clear that given $\boldsymbol{\nu}$, solving for $\mathbf{A}$ and $\boldsymbol{\eta}_0$ is easy—one simply uses $\boldsymbol{\nu}_i$ as explanatory variables in fitting a generalized least-squares problem called the *vector linear model*. One can solve for $\boldsymbol{\nu}$ given $\mathbf{A}$ from the relationship

$$\boldsymbol{\eta} \;=\; \boldsymbol{\eta}_0 + \mathbf{A}\,\boldsymbol{\nu} \;=\; \boldsymbol{\eta}_0 + \mathbf{A}\sum_{k=1}^{p} \boldsymbol{f}_k^*(x_k) \;=\; \boldsymbol{\eta}_0 + \sum_{k=1}^{p} \mathbf{A}\,\boldsymbol{f}_k^*(x_k)\;. \qquad (21)$$

This falls within the constraints-on-the-functions framework (Equation (4)). It entails fitting a vector additive model to the $\boldsymbol{z}_i$ using the $\boldsymbol{x}_i$ with each of the p constraint matrices of $\boldsymbol{x}_i$ equalling $\mathbf{A}$.

6.2 Derivatives

Writing (5) as

$$Q \;=\; \sum_{i=1}^{n} \{\boldsymbol{z}_i - \boldsymbol{\eta}_0 - \mathbf{A}\,\boldsymbol{\nu}_i\}^T\,\mathbf{W}_i\,\{\boldsymbol{z}_i - \boldsymbol{\eta}_0 - \mathbf{A}\,\boldsymbol{\nu}_i\}\,, \qquad (22)$$

the derivative of Q with respect to the elements of $\mathbf{A}$ can be derived. This comes about because, for fixed $\boldsymbol{\eta}_0$ and $\mathbf{A}$, the solution is

$$\hat{\mathbf{c}} \;=\; \left\{\sum_{i=1}^{n}\left(\mathbf{A}^T\mathbf{W}_i\mathbf{A}\right)\otimes\left(\boldsymbol{x}_i\,\boldsymbol{x}_i^T\right)\right\}^{-1}\left\{\sum_{i=1}^{n}\left[\mathbf{A}^T\mathbf{W}_i(\boldsymbol{z}_i-\boldsymbol{\eta}_0)\right]\otimes\boldsymbol{x}_i\right\} \qquad (23)$$

where $\boldsymbol{c} = (\boldsymbol{c}_{(1)}^T, \ldots, \boldsymbol{c}_{(r)}^T)^T = \mathrm{vec}(\mathbf{C})$. Thus, treating it as a profile likelihood,

$$\begin{aligned}
\frac{\partial Q}{\partial \boldsymbol{a}_{(s)}} &= -2\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\eta}_0^T}{\partial \boldsymbol{a}_{(s)}} + \nu_{is}\,\mathbf{I}_M + \frac{\partial \boldsymbol{\nu}_i^T}{\partial \boldsymbol{a}_{(s)}}\,\mathbf{A}^T\right)\mathbf{W}_i\,(\boldsymbol{z}_i - \boldsymbol{\eta}_0 - \mathbf{A}\,\boldsymbol{\nu}_i)\\
\text{where}\quad \frac{\partial \boldsymbol{\eta}_0^T}{\partial \boldsymbol{a}_{(s)}} &= -\left[\sum_{i=1}^{n}\left\{\nu_{is}\,\mathbf{I}_M + \frac{\partial \boldsymbol{\nu}_i^T}{\partial \boldsymbol{a}_{(s)}}\,\mathbf{A}^T\right\}\mathbf{W}_i\right]\left(\sum_{i=1}^{n}\mathbf{W}_i\right)^{-1},\\
\frac{\partial \nu_{ij}}{\partial \boldsymbol{a}_{(s)}} &= \left(\frac{\partial \boldsymbol{c}_{(j)}^T}{\partial \boldsymbol{a}_{(s)}}\right)\boldsymbol{x}_i \;=\; \sum_{k=1}^{p}\frac{\partial c_{kj}}{\partial \boldsymbol{a}_{(s)}}\,x_{ik}, \qquad j = 1,\ldots,r, \quad \text{and}
\end{aligned}$$

$$\begin{aligned}
\frac{\partial\,\boldsymbol{c}}{\partial\,(\boldsymbol{a}_{(s)})_t} \;=\; &\left\{\sum_{i=1}^{n}\left(\mathbf{A}^T\otimes\boldsymbol{x}_i\right)\mathbf{W}_i\left(\mathbf{A}\otimes\boldsymbol{x}_i^T\right)\right\}^{-1}\left\{\sum_{i=1}^{n}\left(\boldsymbol{e}_t^T\,\mathbf{W}_i\,(\boldsymbol{z}_i-\boldsymbol{\eta}_0)\right)(\boldsymbol{e}_s\otimes\boldsymbol{x}_i) - \right.\\
&\left. 2\left[\sum_{i=1}^{n}\sum_{k=1}^{r}\left(\boldsymbol{e}_t^T\mathbf{W}_i\,\boldsymbol{a}_{(k)}\right)\left(\left(\boldsymbol{e}_s\,\boldsymbol{e}_k^T\right)\otimes(\boldsymbol{x}_i\,\boldsymbol{x}_i^T)\right)\right]\boldsymbol{c}\right\}, \qquad s = 1,\ldots,r, \quad t = r+1,\ldots,M.
\end{aligned}$$

Here, $\boldsymbol{e}_s$ is a vector of 0's but with a 1 in the sth position. We have implemented this algorithm using quasi-Newton minimization (using S-PLUS's `nlminb()`) and found the results quite satisfactory.

6.3 Weighted Reduced-rank Approximation to B

We make note of the special case of (22) with Diag($\mathbf{W}_i$) replacing $\mathbf{W}_i$, setting $\boldsymbol{x}_i = \boldsymbol{e}_i$ and suppressing the intercept. This results in the problem of approximating a general $n \times M$ matrix $\mathbf{Z}$ by a specified lower rank matrix $\hat{\mathbf{Z}} = \mathbf{A}\mathbf{C}^T$ by minimizing

$$\| \boldsymbol{\Omega} * \left(\mathbf{Z} - \hat{\mathbf{Z}}\right) \|^2 \;=\; \sum_{j=1}^{n} \sum_{k=1}^{M} (\boldsymbol{\Omega})_{jk} \left(z_{ij} - \boldsymbol{a}_j^T \boldsymbol{c}_k\right)^2, \tag{24}$$

where $\boldsymbol{\Omega}$ is an associated weight matrix and $*$ is the Hadamard (element-by-element) product. This problem was first considered by Gabriel and Zamir (1979), who developed a criss-cross weighted regression algorithm. Unfortunately, that algorithm can converge to a local optimum.

We also note that an approximate solution to a RR-VGLM can be obtained from a fitted VGLM by solving (24) with $\mathbf{Z} = \hat{\mathbf{B}}$ and setting $(\boldsymbol{\Omega})_{jk} = \mathrm{se}(\hat{\beta}_{(j)k})^{-2}$ as an approximate weight matrix.

6.4 RR-VGAMs

RR-VGAMs can be estimated by minimizing the quantity

$$Q^* \;=\; \sum_{i=1}^{n} \left\{ \boldsymbol{z}_i - \boldsymbol{\eta}_0 - \sum_{t=1}^{r} \boldsymbol{f}_t(\boldsymbol{c}_{(t)}^T \boldsymbol{x}_i) \right\}^T \mathbf{W}_i \left\{ \boldsymbol{z}_i - \boldsymbol{\eta}_0 - \sum_{t=1}^{r} \boldsymbol{f}_t(\boldsymbol{c}_{(t)}^T \boldsymbol{x}_i) \right\}, \tag{25}$$

at each Newton-Raphson iteration. This is the natural objective function arising from (26).

Equation (25) is the multivariate version of the problem considered by Roosen and Hastie (1993), who applied projection pursuit regression to the exponential family, especially logistic regression. Estimation of (25) can be achieved by an alternating algorithm: a Gauss-Newton step to estimate the $\boldsymbol{\nu}$ given $\boldsymbol{f}_t$, and estimating the $\boldsymbol{f}_t$ given the $\boldsymbol{\nu}$ by fitting a vector additive model. The former requires the evaluation of derivatives of the $\boldsymbol{f}_t$, which is easy if splines are used. Most of the details in Roosen and Hastie (1993) are needed for its multivariate case.

7 Discussion

Reduced-rank regression is a potentially powerful tool that has been under-utilized in practice. To address this, this article has extended its applicability to the very large VGLM class, and proposed nonparametric extensions. In this paper we have focussed on the MLM. For that model, it is well-known that some groups can be completely separable from others on a linear projection, leading to some infinite estimates. While this shortcoming can also occur with the reduced-rank MLM, it is not a serious problem (Ripley, 1996).

One data type within the VGLM class that would benefit from reduced-rank regression are vector time series. This has been considered by, for example, Johansen (1995), Velu *et al.* (1986). If formulated for general responses and fitted by iteratively reweighted least squares (Li, 1994) they would then fit under the umbrella of this article, and share in its usefulness.

7.1 Vector Projection Pursuit Regression

It is very interesting to see that each component of $\boldsymbol{\eta}$ in (17) is modelled as a projection pursuit regression (PPR; Friedman & Stuetzle, 1981). This shows that the underlying model behind the RR-VGAM class is an extension of PPR to vector responses, i.e.,

$$\boldsymbol{Y}_i \;=\; \boldsymbol{\beta}_0 + \sum_{t=1}^{r} \boldsymbol{f}_t(\boldsymbol{c}_{(t)}^T \boldsymbol{x}) + \boldsymbol{\varepsilon}_i, \quad E(\boldsymbol{\varepsilon}_i) = \mathbf{0}, \quad \mathrm{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma}_i = \mathbf{W}_i^{-1}, \tag{26}$$

where, as in the usual PPR estimation process, r is determined from the data. We call this the *Vector PPR* (VPPR) model. Roosen and Hastie (1993) have applied PPR to distributions in the exponential family, so RR-VGAMs of the VPPR type are the natural generalization of that work.

The VPPR model is a generalization of SMART (Smooth Multiple Additive Regression Technique; Friedman (1984)), who used it for multiple response regression and classification.

7.2 Software

Software implementing VGAMs and the methods of this paper are packaged in the VGAM library for S-PLUS and R. The VGAM library centers on the `vglm()` and `vgam()` functions, which are multivariate versions of `glm()` and `gam()` respectively. The relevant family functions are `stereotype()`, `rrmultinomial()` and `rrGaussian()`. A typical usage might be

```
fit <- vglm(ymatrix ~ x1 + x2 + x3, family=rrmultinomial(rank=2)).
```

Functions `rrmultinomial()` and `stereotype()` implement the alternating and derivative algorithms respectively. Plotting functions are also supplied. In the above example, `fit$constraints` holds $\widehat{\mathbf{A}}$'s, `coef(fit)` holds $\widehat{\boldsymbol{\eta}}_0$ and $\widehat{\mathbf{C}}^T$, and `coef(fit, matrix=T)[-1,]` is the reduced-rank approximation to $\widehat{\mathbf{B}}$. VGAM is freely available at the first author's home page at `http://www.stat.auckland.ac.nz`.

Acknowledgements

The authors thank Michael Greenacre and Brian McArdle for helpful discussions, and Professor Gabriel for beneficial comments. The first author benefited from a Auckland University Research Committee grant, and thanks the Department of Statistics at Stanford University for hospitality on a number of occasions. This paper is dedicated to the U.S. Immigration officer who interviewed one of the authors trying to enter the United States. After a lengthy interrogation process, the officer asked "And what journal do you plan to send it to?" When I replied "Journal of the Royal Statistical Society, Se..." his boss said "That sounds good enough" and let me into the country.

References

Aldrin, M. (2000) Multivariate prediction using softly shrunk reduced-rank regression. *Amer. Statist.*, **54**, 29–34.

Anderson, J. A. (1984) Regression and ordered categorical variables (with discussion), *J. Roy. Statist. Soc.* **B, 46**, 1–30.

Anderson, T. W. (1951) Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Annals of Math. Statist.*, **22**, 327–351. [Correction, (1980), *Ann. Statist.*, **8**, p. 1400].

Brillinger, D. R. (1969) The canonical analysis of stationary time series. In: *Multivariate Analysis 2*, ed. P. R. Krishnaiah, pp. 331–50. New York: Academic Press.

Boser, B. and Guyon, I. (1992) A training algorithm for optimal margin classifiers. In: *Proceedings of COLT II*. Philedelphia.

Breiman, L. and Friedman, J. H. (1997) Predicting multivariate responses in multiple linear regression (with discussion). *J. Roy. Statist. Soc.* **B, 59**, 3–54.

Campbell, N. (1984) Canonical variate analysis—a general formulation. *Aust. J. Statist.*, **26**, 86–96.

Chessel, D., J.-D. Lebreton and N. Yoccoz (1987) Propriétés de l'analyse canonique des correspondances; une illustration en hydrobiologie. *Rev. Statist. Appl.*, **35**, 55–72.

Eckart, C. and Young, G. (1936) The approximation of one matrix by one of lower rank. *Psychometrika*, **1**, 211–18.

Efron, B. (1975) The efficiency of logistic regression compared to normal discriminant analysis. *J. Am. Statist. Ass.*, **70**, 892–898.

Friedman, J. H. and Stuetzle, W. (1981) Projection pursuit regression. *J. Am. Statist. Ass.*, **76**, 817–823.

Friedman, J. H. (1984) *SMART User's Guide.* Technical Report 1, Department of Statistics and Stanford Linear Accelerator Center, Stanford University.

Gabriel, K. R. and Zamir, S. (1979) Lower rank approximation of matrices by least squares with any choice of weight. *Technomet.*, **21**, 489–98.

Gabriel, K. R. (1998) Generalised bilinear regression. *Biometrika*, **85**, 689–700.

Greenland, S. (1994) Alternative models for ordinal logistic regression. *Statist. Med.*, **13**, 1665–77.

Hastie, T. and Little, F. (1987) Principal profiles. (Draft manuscript).

Hastie, T. J. and Tibshirani, R. J. (in preparation) Multi-response additive models for regression and classification.

Hastie, T. and Tibshirani, R. (1996) Discriminant analysis by Gaussian mixtures. *J. Roy. Statist. Soc.* **B, 58**, 155–176.

Hastie, T., Tibshirani, R. and Buja, A. (1994) Flexible discriminant analysis by optimal scoring. *J. Am. Statist. Ass.*, **89**, 1255–1270.

Ihm, P. and H. van Groenewoud (1984) Correspondence analysis and Gaussian ordination. COMPSTAT lectures, **3**, 5–60.

Izenman, A. J. (1975) Reduced-rank regression for the multivariate linear model. *J. Multivar. Anal.*, **5**, 248–264.

Johansen, S. (1995) *Likelihood-based Inference in Cointegrated Vector Autoregressive Models.* Oxford: Oxford University Press.

Lebreton, J.-D., D. Chessel, R. Prodon and N. Yoccoz (1988) L'analyse des relations espécies-milieu par l'analyses canonique des correspondances. I. Variables de milieu quantitatives. *Acta Oecol. Gen.*, **9**, 53–67.

Lebreton, J.-D., R. Sabatier, G. Banco and A. M. Bacou (1991) Principal component and correspondence analysis with respect to instrumental variables: an overview of their role in studies of structure-activity and species-environment relationships. In: Devillers, J. and W. Karcher (eds.), *Applied Multivariate Analysis in SAR and Environmental Studies*, pp. 85–114. Dordrecht: Kluwer.

Li, W. K. (1994) Time series models based on generalized linear models: some further results. *Biometrics*, **50**, 506–511.

McCullagh, P. (1980) Regression models for ordinal data. *J. Roy. Statist. Soc.* **B, 42**, 109–142.

Nerlove, M. and Press, S. J. (1973) Univariate and multivariate log-linear and logistic models. R-1306-EDA/NIH, Santa Monica, California: Rand Corporation.

Rao, C. R. (1964) The use and interpretation of principal component analysis in applied research. *Sankhya* **A, 26**, 329–358.

Reinsel, G. C., and Velu, R. P. (1998) *Multivariate Reduced-Rank Regression: Theory and Applications.* New York: Springer.

Ripley, B. D. (1996) *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.

Roosen, C. B. and Hastie, T. J. (1993) *Logistic Response Projection Pursuit.* Tech. Report, AT&T Bell Laboratories, Document number BL011214-930806-09TM.

Sabatier, R., J.-D. Lebreton and D. Chessel (1989) Multivariate analysis of composition data accompanied by qualitative variables describing a structure. In: Coppi, R. and S. Bolasco (eds.), *Multiway Data Tables*, pp. 341–352. Amsterdam: North-Holland.

Schmidli, H. (1995) *Reduced Rank Regression: With Applications to Quantitative Structure-Activity Relationships.* Heidelberg: Physica-Verlag.

Srivastava, M. S. (1997) Reduced rank discrimination. *Scan. J. Statist.*, **24**, 115–124.

Takane, Y. H., Yanai H., and S. Mayekawa (1991) Relationships among several methods of linearly constrained correspondence analysis. *Psychometrika* **56**, 667–684.

ter Braak, C. J. F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67**, 1167–1179.

ter Braak, C. J. F. and Looman, C. W. N. (1994) Biplots in reduced-rank regression. *Biometrical Journal*, **36**, 983–1003.

ter Braak, C. J. F. and Smilauer, P. (1998) *CANOCO Reference Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4).* Microcomputer Power, Ithaca, NY.

ter Braak, C. J. F. and Verdonschot, P. F. M. (1995) Canonical correspondence analysis and related multivariate methods in aquatic ecology. *Aquatic Sci.*, **57**, 255–289.

Velu, R. P., G. C. Reinsel, and D. W. Wichern (1986) Reduced rank models for multiple time series. *Biometrika*, **73**, 105–118.

Yee, T. W. and Wild, C. J. (1996) Vector generalized additive models. *J. Roy. Statist. Soc.* **B, 58**, 481–493.

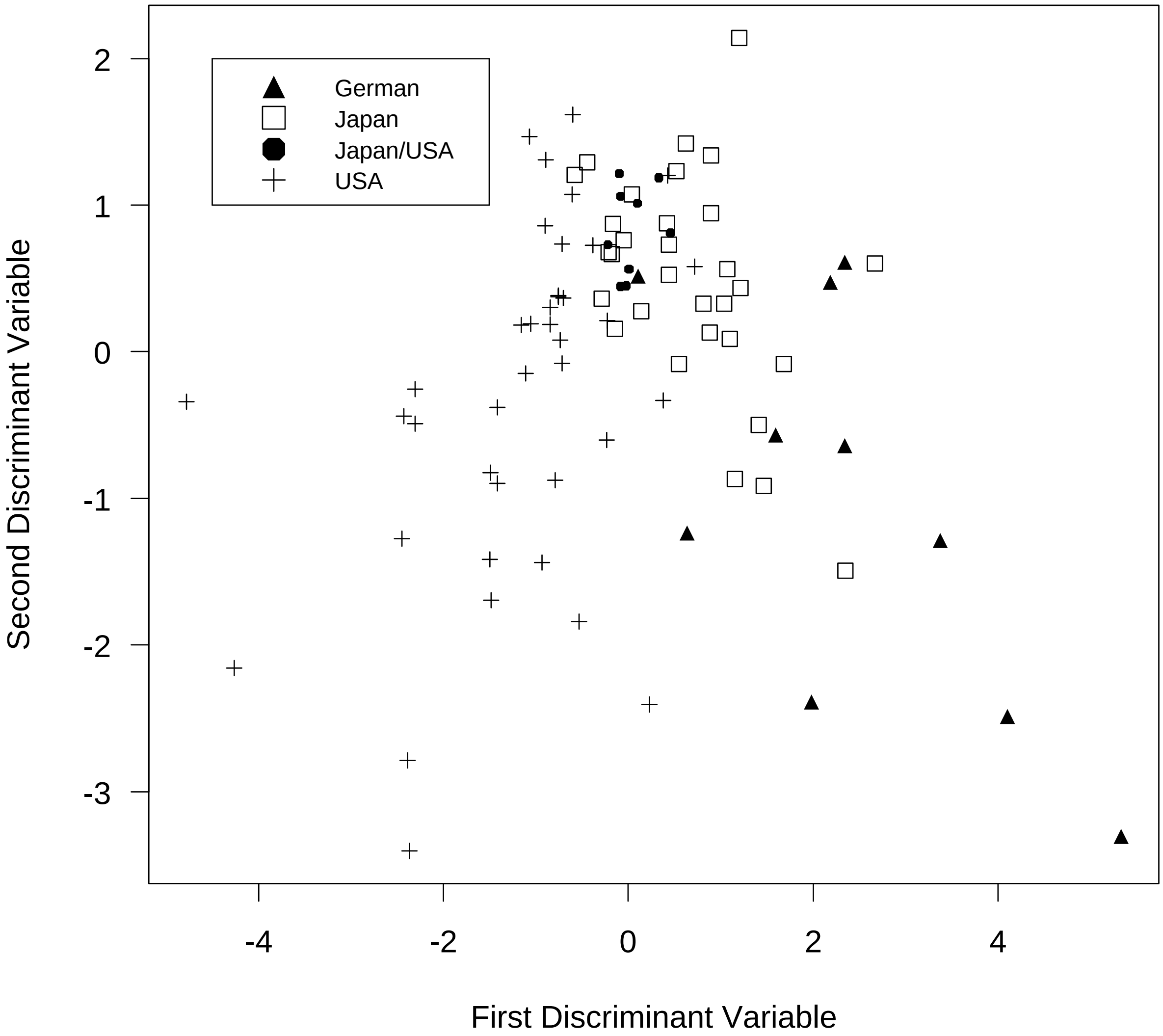

Figure 4: Linear discriminant analysis on the standardized car data.

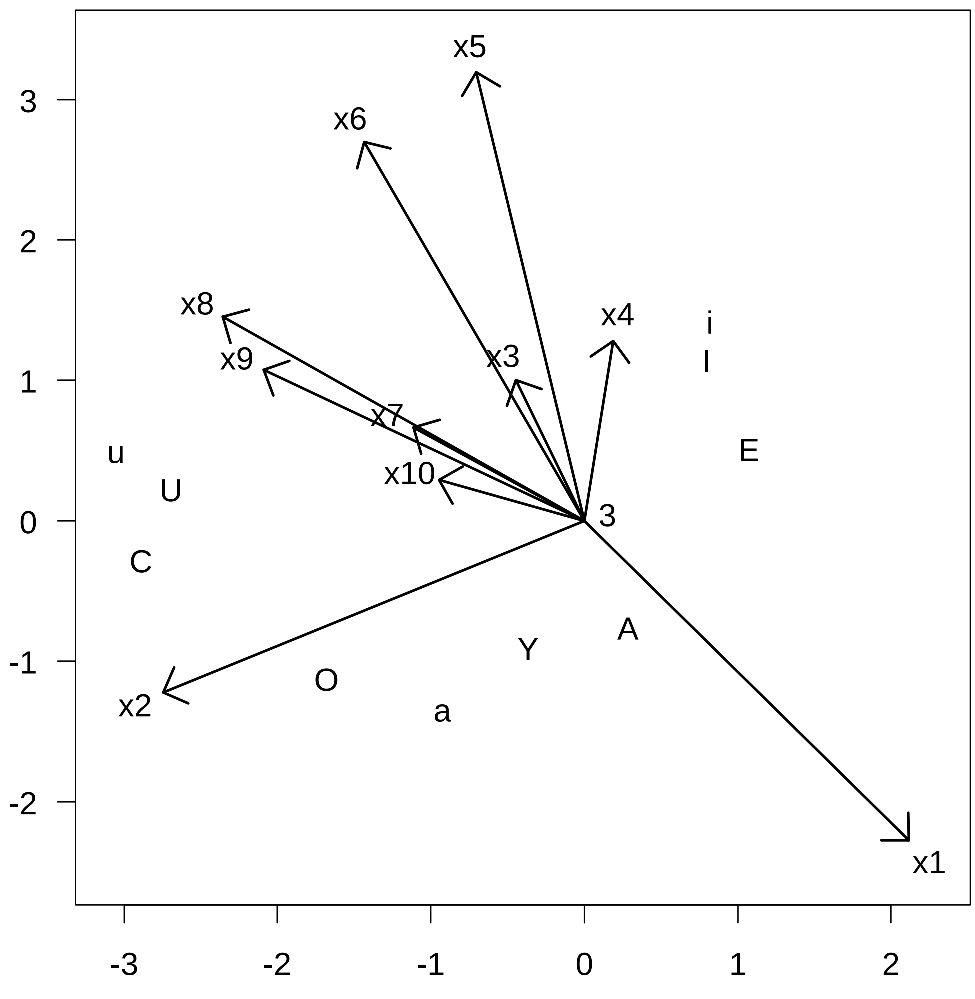

Figure 5: Biplot based on the reduced-rank regression coefficients fitted to the vowel data.

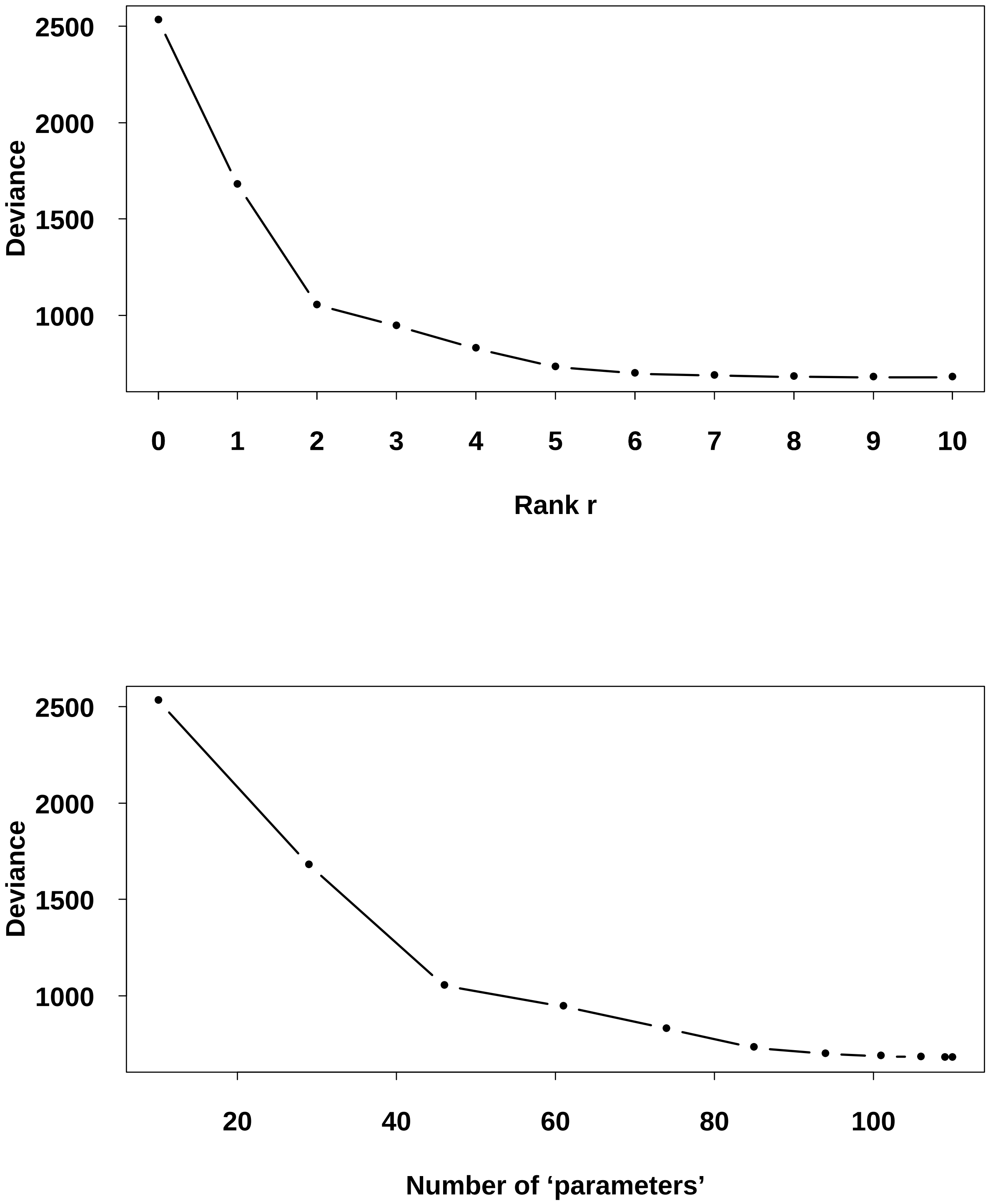

Figure 6: Deviance versus rank r, and the number of unknowns in $\mathbf{A}$ and $\mathbf{C}$ for a sequence of RR-MLMs fitted to the vowel training data.

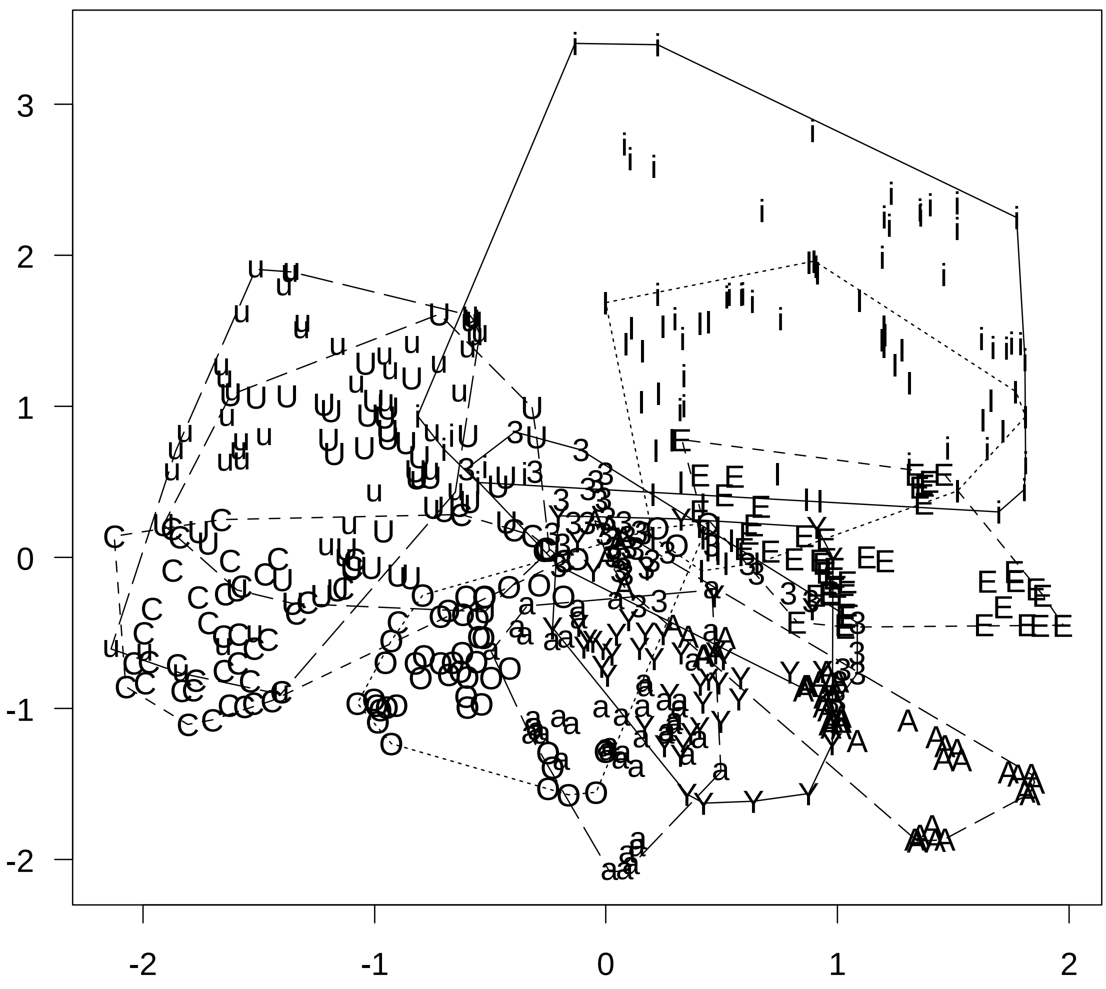

Figure 7: Scatterplot of ν_1 (horizontal axis) and ν_2 for the vowel recognition data (Rank-2 RR-MLM). Convex hulls for each vowel are added.

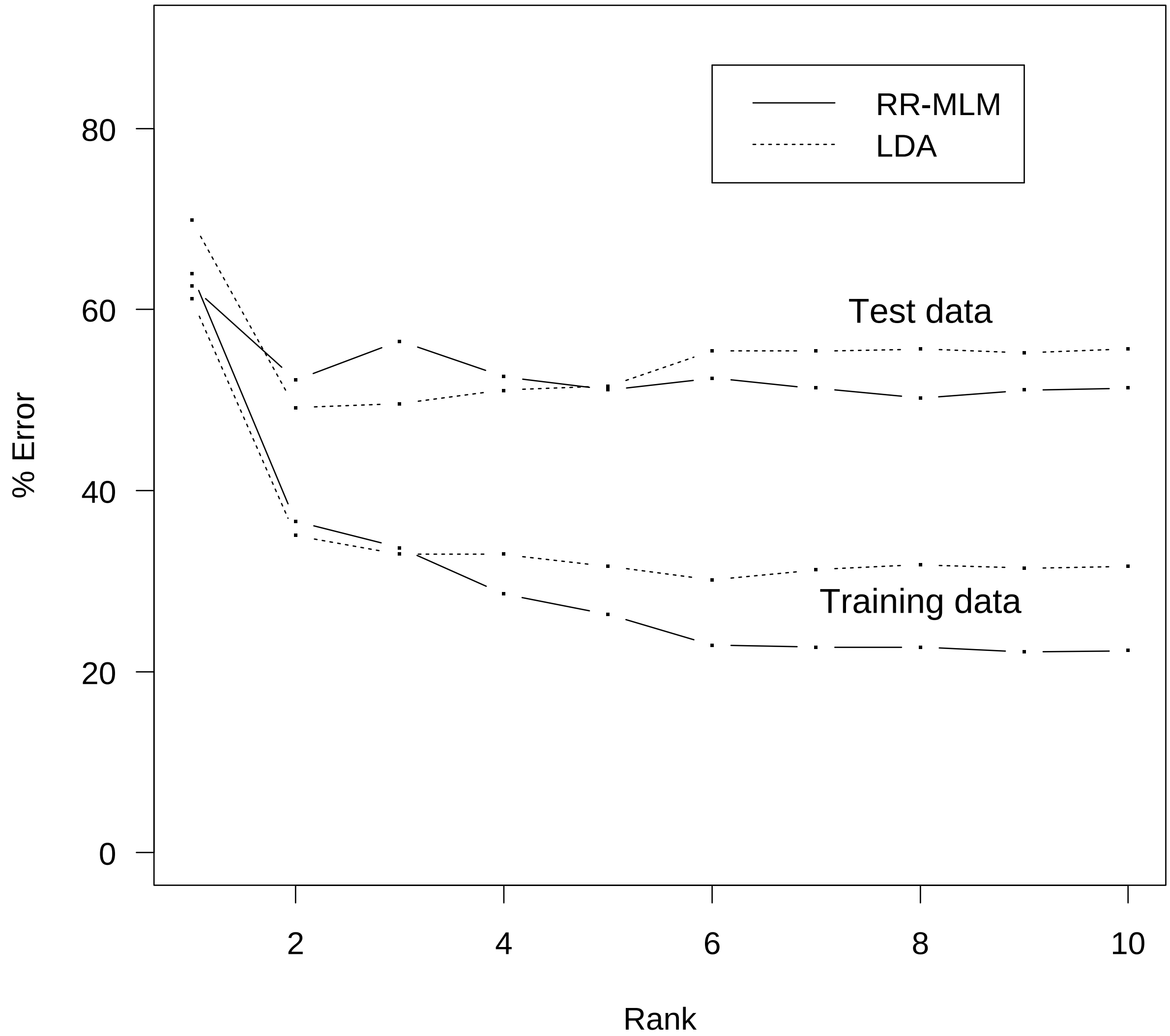

Figure 8: Error rates for various rank RR-MLMs and LDAs fitted to the vowel recognition data.