Statistica Sinica **20** (2010), 927-1010

ENVELOPE MODELS FOR PARSIMONIOUS AND EFFICIENT MULTIVARIATE LINEAR REGRESSION

R. Dennis Cook[1], Bing Li[2] and Francesca Chiaromonte[2]

[1] *University of Minnesota and* [2] *Pennsylvania State University*

Abstract: We propose a new parsimonious version of the classical multivariate normal linear model, yielding a maximum likelihood estimator (MLE) that is asymptotically less variable than the MLE based on the usual model. Our approach is based on the construction of a link between the mean function and the covariance matrix, using the minimal reducing subspace of the latter that accommodates the former. This leads to a multivariate regression model that we call the *envelope model*, where the number of parameters is maximally reduced. The MLE from the envelope model can be *substantially* less variable than the usual MLE, especially when the mean function varies in directions that are orthogonal to the directions of maximum variation for the covariance matrix.

Key words and phrases: Discriminant analysis, functional data analysis, grassmann manifolds, invariant subspaces, principal components, reduced rank regression, reducing subspaces, sufficient dimension reduction.

1. Introduction

A cornerstone of multivariate analysis is the multivariate linear regression model

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta}\mathbf{X} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $\mathbf{Y} \in \mathbb{R}^r$ is the random response vector, $\mathbf{X} \in \mathbb{R}^p$ is a non-stochastic vector of predictors, and the error vector $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is normally distributed with mean $\mathbf{0}$ and unknown covariance matrix $\boldsymbol{\Sigma} \geq \mathbf{0}$ (see Christensen (2001) for background). If $\mathbf{X}$ is random during sampling then the model is conditional on the observed values of $\mathbf{X}$. This conditioning, common practice in regression, was discussed by Aldrich (2005) from an historical perspective. The intercept $\boldsymbol{\alpha} \in \mathbb{R}^r$ is an unknown parameter vector and $\boldsymbol{\beta}$ is an unknown parameter matrix of dimensions $r \times p$. Model (1.1) has a total of $r+pr+r(r+1)/2$ unknown real parameters when $\boldsymbol{\Sigma} > \mathbf{0}$, and it may be a rather coarse tool if this number is large. Variations have been developed to sharpen its abilities. Notable among them is the class of reduced-rank regressions, which allow for the possibility that $\mathrm{rank}(\boldsymbol{\beta}) < \min(p, r)$ (Reinsel and Velu (1998)). In this article we propose a new version of model (1.1) that yields a maximum likelihood estimator (MLE) of $\boldsymbol{\beta}$ with the potential to be substantially less variable asymptotically than the usual MLE. In the remainder of this section we discuss our motivation and describe its implications informally, outline the rest of the article, and establish notation for the technical developments that begin in Section 2.

1.1. Motivation

Our primary motivation comes from the simple observation that some characteristics of the response vector could be unaffected by changes in the predictors. Multiple responses are incorporated in many regressions in an effort to encapsulate changes in the distribution of an experimental or sampling unit as the predictors vary. For example, several anatomical measurements might be taken on individual skulls to compare populations, milk production might be measured on dairy cows at several points during the lactation cycle, hematological measures might be taken on patients at several times following a drug treatment, or spectral readings might be taken on samples at several wavelengths. In the same vein, multiple distances and angular measurements are used to model human motion in ergonomic studies (e.g., Faraway and Reed (2007)), and multiple biomarkers are used as responses when studying dietary patterns that affect coronary artery disease (Hoffmann et al. (2004)). In these types of multivariate regression it may be reasonable to allow for the possibility that aspects of the response vector are stochastically constant as the predictors vary.

Assuming model (1.1), suppose that we can find an orthogonal matrix $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ that satisfies the conditions: (i) span$(\boldsymbol{\beta}) \subseteq$ span$(\mathbf{\Gamma})$, and (ii) $\mathbf{\Gamma}^T\mathbf{Y}$ is conditionally independent of $\mathbf{\Gamma}_0^T\mathbf{Y}$ given $\mathbf{X}$. Condition (i) is not restrictive by itself, since at least one, and typically infinitely many semi-orthogonal matrices $\mathbf{\Gamma}$ exist with a span containing span$(\boldsymbol{\beta})$. Under this condition the marginal distribution of $\mathbf{\Gamma}_0^T\mathbf{Y}$ does not depend on $\mathbf{X}$. However, $\mathbf{\Gamma}_0^T\mathbf{Y}$ may still provide information about the regression through its association with $\mathbf{\Gamma}^T\mathbf{Y}$. This possibility is ruled out by condition (ii). Together (i) and (ii) imply that $\mathbf{\Gamma}_0^T\mathbf{Y}$ is marginally independent of $\mathbf{X}$ *and* conditionally independent of $\mathbf{X}$ given $\mathbf{\Gamma}^T\mathbf{Y}$. If $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)$ were known the analysis could be facilitated by using the transformed response $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)^T\mathbf{Y}$, and then backtransforming to the original scale after estimation. In practice we would not normally know a suitable transformation; nevertheless the possibility that such a transformation exists has important implications for the analysis. In this setting it can be verified that

$$\mathbf{\Sigma} = \mathbf{P}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Gamma}} + \mathbf{Q}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{Q}_{\mathbf{\Gamma}}, \tag{1.2}$$

where $\mathbf{P}_{\mathbf{\Gamma}}$ is the projection onto span$(\mathbf{\Gamma})$ in the usual inner product, and $\mathbf{Q}_{\mathbf{\Gamma}} = \mathbf{I}_r - \mathbf{P}_{\mathbf{\Gamma}}$. More precisely, given condition (i), condition (ii) is equivalent to (1.2). The crucial point here is that conditions (i) and (1.2) establish a *parametric link* between $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ that is the key for the new methodology proposed in this article. However, this link is not now well-defined because there may still be infinitely many subspaces span($\boldsymbol{\Gamma}$) that satisfy the conditions. Section 2 is devoted to the algebraic background necessary to construct the unique smallest subspace span($\boldsymbol{\Gamma}$) that satisfies (1.2) and contains span($\boldsymbol{\beta}$). This minimal subspace, which we call the $\boldsymbol{\Sigma}$-*envelope of* span($\boldsymbol{\beta}$) in full, and the *envelope* for brevity, is then used as a parameter in the *envelope model for multivariate linear regression* defined in Section 3. For now we proceed as if span($\boldsymbol{\Gamma}$) were the envelope.

The full space $\mathbb{R}^r = \text{span}(\mathbf{I}_r)$ trivially contains span($\boldsymbol{\beta}$) and satisfies the decomposition (1.2). If $\mathbb{R}^r$ is the envelope, then the entire response vector $\mathbf{Y}$ is relevant to the regression, a finding that could be useful in its own right. We expect $\mathbb{R}^r$ to be the envelope when r is small and the responses are carefully chosen to reflect distinct aspects of the sampling units. However, we also expect that redundant or irrelevant information is present in the kinds of applications we have in mind, particularly when many responses are measured in an effort to capture characteristics of the sampling units that vary with the predictors.

Instances of this may occur as a consequence of reasoning about underlying processes. This is the case, for example, in the context of large-scale gene expression data from microarrays. Our argument is tantamount to that used by Leek and Storey (2007) when proposing their method of surrogate variable analysis. Suppose we would like to regress a vector $\mathbf{Y}$ of many (perhaps thousands) gene expression readings on a set of covariates $\mathbf{C}$ (these may comprise environmental factors, treatments or clinical outcomes). Assume that there is an "ideal" vector $\boldsymbol{\nu} \in \mathbb{R}^d$ of latent variables connecting these covariates and the expression levels, so that $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu} + \boldsymbol{\epsilon}_0$ – where $\boldsymbol{\Gamma}$ is a semi-orthogonal matrix and $\text{Var}(\boldsymbol{\epsilon}_0) = \sigma^2\mathbf{I}_r$, as argued by Leek and Storey. Since $\boldsymbol{\nu}$ is unobserved, we write $\boldsymbol{\nu} = \text{E}(\boldsymbol{\nu}|\mathbf{C}) + \boldsymbol{\epsilon}$ and then substitute into the model to obtain $\mathbf{Y} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\text{E}(\boldsymbol{\nu}|\mathbf{C}) + \boldsymbol{\Gamma}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_0$. The covariates $\mathbf{C}$ might provide only partial information on $\boldsymbol{\nu}$, so some coordinates of $\text{E}(\boldsymbol{\nu}|\mathbf{C})$ could be constant, with the consequence that $\text{E}(\boldsymbol{\nu}|\mathbf{C})$ varies in fewer than d dimensions. The modeling process can be viewed as providing a representation for the unknown conditional mean $\text{E}(\boldsymbol{\nu}|\mathbf{C}) = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}\mathbf{X}(\mathbf{C})$, where $\mathbf{X}$ is the vector of predictors included in the model. As represented, $\mathbf{X}$ is a function of $\mathbf{C}$ and might contain transformations of the measured covariates, or interactions among them. Assuming that $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\epsilon}_0$ leads to the multivariate linear model (1.1) with $\boldsymbol{\alpha} = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\gamma}_0$, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\gamma}$, $\boldsymbol{\varepsilon} = \boldsymbol{\Gamma}\boldsymbol{\epsilon} + \boldsymbol{\epsilon}_0$, and

$$\begin{aligned}\boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\text{Var}(\boldsymbol{\epsilon})\boldsymbol{\Gamma}^T + \sigma^2\mathbf{I}_r \\ &= \boldsymbol{\Gamma}(\text{Var}(\boldsymbol{\epsilon}) + \sigma^2\mathbf{I}_d)\boldsymbol{\Gamma}^T + \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T. \qquad (1.3)\end{aligned}$$

Since span($\boldsymbol{\beta}$) $\subseteq$ span($\boldsymbol{\Gamma}$) we have an instance of (1.2) with $\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}(\text{Var}(\boldsymbol{\epsilon}) + \sigma^2\mathbf{I}_d)\boldsymbol{\Gamma}^T$ and $\mathbf{Q}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{Q}_{\boldsymbol{\Gamma}} = \sigma^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T$. The same essential reasoning can be applied in the context of multivariate calibration, where $\mathbf{Y}$ is the vector of spectral readings and $\boldsymbol{\nu}$ depends on the concentrations of interest and all other characteristics of the sample that affect the readings.

Decomposition (1.2) implies that the eigenvectors of $\boldsymbol{\Sigma}$ fall in either the envelope span($\boldsymbol{\Gamma}$) or its orthogonal complement span($\boldsymbol{\Gamma}_0$). The corresponding eigenvalues of $\boldsymbol{\Sigma}$ need not be partitioned in any particular order, since (1.2) does not presume any relation between the magnitudes of the two terms comprising $\boldsymbol{\Sigma}$. The greatest gains in efficiency occur when the first term on the right of (1.2), $\mathbf{P}_{\boldsymbol{\Gamma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Gamma}}$, is associated with the smaller eigenvalues of $\boldsymbol{\Sigma}$. However, efficiency gains can also occur under (1.3), where the envelope captures the leading eigenvectors of $\boldsymbol{\Sigma}$. Relatedly, the estimated error covariance matrix $\widehat{\boldsymbol{\Sigma}}$ for these regressions often contains a few large eigenvalues followed by a large "tail space" of relatively small eigenvalues of similar size. One can think of this as the sample counterpart of a population error variability structure with a few leading directions, and a large tail space of approximately spherical spread. This structure is a useful descriptor not just for microarray data, but also for other large-scale genomic data; we recently described it for frequencies of short alignment patterns in a comparative genomic study of regulatory elements (sections of nuclear DNA that determine the activation of genes; Cook, Li and Chiaromonte (2007, Figure 2)).

The connection with the eigenstructure of $\boldsymbol{\Sigma}$ can be used to provide some intuition about the mechanisms that produce efficiency gains in our approach. Consider a regression in which $p = 1$, and $\boldsymbol{\Sigma} > \mathbf{0}$ is known, and has distinct eigenvalues. Knowledge of $\boldsymbol{\Sigma}$ alone does not alter the MLE of $\boldsymbol{\beta}$. However, if we also know that $\boldsymbol{\beta}$ falls in the span of, say, the last eigenvector $\mathbf{v}_r$ of $\boldsymbol{\Sigma}$, then span($\mathbf{v}_r$) is the envelope and we can use a simple univariate linear regression model with response $\mathbf{v}_r^T\mathbf{Y}$ to estimate the direction and length of $\boldsymbol{\beta}$. If the eigenvalue of $\boldsymbol{\Sigma}$ corresponding to $\mathbf{v}_r$ is substantially smaller than the largest eigenvalue, then the MLE based on $\mathbf{v}_r^T\mathbf{Y}$ will have substantially smaller variation than the usual MLE. Gains can also be realized when $\boldsymbol{\Sigma}$ is unknown, but we can infer that the envelope is contained in a subspace spanned by a proper subset of the eigenvectors of $\boldsymbol{\Sigma}$. In full generality, our envelope models are not limited to regressions with $p = 1$, and do not constrain the rank of $\boldsymbol{\beta}$. They do not require $\boldsymbol{\Sigma}$ to have distinct eigenvalues, or even to be positive definite. However, to focus on the main ideas, we assume throughout that $\boldsymbol{\Sigma} > 0$.

Next, we use a data example to demonstrate the efficiency gains that are possible with our approach. Consider data on $r = 6$ responses, the logarithms of near infrared reflectance at six wavelengths across the range 1,680-2,310 nm, measured on samples from two populations of ground wheat with low and high protein content (24 and 26 samples, respectively). The mean difference $\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2$

corresponds to the parameter vector $\boldsymbol{\beta}$ in model (1.1), with $\mathbf{X}$ representing a binary indicator: $\mathbf{X} = 0$ for high protein wheat, and $\mathbf{X} = 1$ for low protein wheat. For these data, the standard errors of the six estimated mean differences based on the usual normal-theory analysis under (1.1) range between 6.4 and 65.8 times the standard errors of the corresponding estimates based on the envelope model. In other words, to achieve comparable standard errors, normal-theory estimates might have to use as many as $65^2 \times 50$ samples where envelope estimates use 50. This example is revisited in Section 7.2.

Reducing redundancy in large data sets has become paramount in an era of high-throughput technologies and fast computing. In many applications, costs are accrued when increasing the number of units, while hundreds or thousands of variables can be recorded on each unit at relatively low expense – which is often done without articulating a specific design at the outset. The resulting data may contain a considerable amount of information that is either irrelevant or redundant for a given purpose. Contemporary statistical theories and methodologies are quickly evolving to adapt to this new reality, with rapid advances in areas such as dimension reduction, sparse variable selection via regularization, and "large-p-small-n" hypotheses testing. The envelope model we introduce uses the error variability structure to create a minimal enclosing of the mean signal in mutivariate data. If these constraints correspond to physical mechanisms, enveloping is a natural way to reflect them; if not, it can still be used as a means of regularization. In either case, controlling the dimension of the envelope can achieve a degree of "eigen sparsity" for the first two moments – arguably the most important descriptors for a broad range of data analyses.

1.2. Outline

Envelopes, which arise from the concepts of invariant and reducing subspaces, are introduced in Section 2. The results in this section, although technical in nature, are immediately relevant to the core developments of the paper. Envelope models for multivariate linear regression are described in Section 3, and maximum likelihood estimation of their parameters is developed in Section 4. Selected asymptotic results are presented in Section 5, and a discussion to aid their interpretation is given in Section 6. Section 7 contains simulation and data analysis results. The envelope theory and methods described in Sections 3−7 make use of the error covariance matrix associated with model (1.1), i.e., the intra-population covariance matrix $\boldsymbol{\Sigma} = \mathrm{Var}\,(\mathbf{Y}|\mathbf{X})$. They do not involve the marginal covariances $\boldsymbol{\Sigma}_{\mathbf{Y}} = \mathrm{Var}\,(\mathbf{Y})$ and $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathrm{Var}\,(\mathbf{X})$. In Section 3.2 we consider some connections among envelopes based on different matrices, and in Section 8 we discuss other contexts in which envelopes might be useful, including reduced rank multivariate models, discriminant analysis, sufficient dimension reduction, and some multivariate methods that involve either $\mathbf{\Sigma_Y}$ or $\mathbf{\Sigma_X}$. Section 9 contains some concluding remarks. An on-line supplement to this article with proofs and other technical details is available at `http://www.stat.sinica.edu.tw/statistica`.

1.3. Notation and definitions

The following notation and basic definitions are used repeatedly in our exposition. For positive integers r and p, $\mathbb{R}^{r\times p}$ stands for the class of all matrices of dimension $r \times p$, and $\mathbb{S}^{r\times r}$ denotes the class of all symmetric $r \times r$ matrices. For $\mathbf{A} \in \mathbb{R}^{r\times r}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^r$, $\mathbf{A}\mathcal{S} \equiv \{\mathbf{A}\mathbf{x} : \mathbf{x} \in \mathcal{S}\}$. For $\mathbf{B} \in \mathbb{R}^{r\times p}$, span($\mathbf{B}$) denotes the subspace of $\mathbb{R}^r$ spanned by the columns of $\mathbf{B}$. A *basis matrix* for a subspace $\mathcal{S}$ is any matrix whose columns form a basis for $\mathcal{S}$. A *semi-orthogonal matrix* $\mathbf{A} \in \mathbb{R}^{r\times p}$ has orthogonal columns, $\mathbf{A}^T\mathbf{A} = \mathbf{I}_p$. A sum of subspaces of $\mathbb{R}^r$ is indicated with the notation '$\oplus$': $\mathcal{S}_1 \oplus \mathcal{S}_2 = \{\mathbf{x}_1 + \mathbf{x}_2 : \mathbf{x}_1 \in \mathcal{S}_1, \mathbf{x}_2 \in \mathcal{S}_2\}$. For a positive definite matrix $\mathbf{\Sigma} \in \mathbb{S}^{r\times r}$, the inner product in $\mathbb{R}^r$ defined by $\langle \mathbf{x}_1, \mathbf{x}_2\rangle_{\mathbf{\Sigma}} = \mathbf{x}_1^T\mathbf{\Sigma}\mathbf{x}_2$ is referred to as the $\mathbf{\Sigma}$ inner product; when $\mathbf{\Sigma} = \mathbf{I}_r$, the r by r identity matrix, this inner product is called the usual inner product. A projection relative to the $\mathbf{\Sigma}$ inner product is the projection operator in the inner product space $\{\mathbb{R}^r, \langle\cdot,\cdot\rangle_{\mathbf{\Sigma}}\}$; that is, if $\mathbf{B} \in \mathbb{R}^{r\times p}$, then the projection onto span($\mathbf{B}$) relative to $\mathbf{\Sigma}$ has the matrix representation $\mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})} \equiv \mathbf{B}(\mathbf{B}^T\mathbf{\Sigma}\mathbf{B})^{\dagger}\mathbf{B}^T\mathbf{\Sigma}$, where † indicates the Moore-Penrose inverse. The projection onto the orthogonal complement of span($\mathbf{B}$) relative to the $\mathbf{\Sigma}$ inner product, $\mathbf{I}_r - \mathbf{P}_{\mathbf{B}(\mathbf{\Sigma})}$, is denoted by $\mathbf{Q}_{\mathbf{B}(\mathbf{\Sigma})}$. Projection operators employing the usual inner product are written with a single subscript argument $\mathbf{P}_{(\cdot)}$, where the subscript describes the subspace, and $\mathbf{Q}_{(\cdot)} = \mathbf{I}_r - \mathbf{P}_{(\cdot)}$. The orthogonal complement $\mathcal{S}^{\perp}$ of a subspace $\mathcal{S}$ is constructed with respect to the usual inner product, unless indicated otherwise.

2. Envelopes

The discussion revolves around the parameterization of a covariance matrix in reference to a subspace that contains a conditional mean vector. Specifically, as we saw in (1.2), this is achieved by decomposing the covariance matrix into the sum of two matrices, each of whose column spaces either contains or is orthogonal to the subspace containing the mean. The only way to do this is to create a split based on the eigenvectors of the covariance. This leads us naturally to invariant and reducing subspaces of a matrix, from which the concept of an envelope arises.

2.1. Invariant and reducing subspaces

Recall that a subspace $\mathcal{R}$ of $\mathbb{R}^r$ is an *invariant subspace* of $\mathbf{M} \in \mathbb{R}^{r\times r}$ if $\mathbf{M}\mathcal{R} \subseteq \mathcal{R}$; so $\mathbf{M}$ maps $\mathcal{R}$ to a subset of itself. $\mathcal{R}$ is a *reducing subspace* of $\mathbf{M}$ if, in addition, $\mathbf{M}\mathcal{R}^{\perp} \subseteq \mathcal{R}^{\perp}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$

reduces $\mathbf{M}$. Some intuition may be provided here by describing how invariant subspaces arise in Zyskind's (1967) pioneering work on linear models. Consider n observations on a univariate linear model written in terms of the $n \times 1$ response vector $\mathbf{W} = \mathbf{F}\boldsymbol{\alpha} + \boldsymbol{\epsilon}$, where $\mathbf{F} \in \mathbb{R}^{n\times p}$ is known, $\boldsymbol{\alpha} \in \mathbb{R}^p$ is the vector we would like to estimate, and $\mathbf{V} = \mathrm{Var}(\boldsymbol{\epsilon}) \in \mathbb{R}^{n\times n}$ denotes the error covariance matrix. The rank of $\mathbf{F}$ may be less than p and $\mathbf{V}$ may be singular. Let $\mathbf{a}^T\boldsymbol{\alpha}$ be an estimable linear combination of the coefficients $\boldsymbol{\alpha}$. Zyskind (1967) showed that the ordinary least squares estimator of $\mathbf{a}^T\boldsymbol{\alpha}$ is equal to the corresponding generalized least squares estimator for every $\mathbf{a} \in \mathbb{R}^p$ if and only if $\mathrm{span}(\mathbf{F})$ is an invariant subspace of $\mathbf{V}$. Our approach is distinct from Zyskind's since we are working with multivariate models and have quite different goals. Additionally, Zyskind's dimensions grow with n, while ours remain fixed.

The next proposition characterizes a matrix $\mathbf{M}$ in terms of projections on its reducing subspaces, and gives exactly the kind of decomposition we seek.

Proposition 2.1. *$\mathcal{R}$ reduces $\mathbf{M} \in \mathbb{R}^{r\times r}$ if and only if $\mathbf{M}$ can be written in the form*

$$\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}. \tag{2.1}$$

Corollary 2.1 describes the consequences of Proposition 2.1 (and Lemma A.1 reported in the Supplement), including a relationship between reducing subspaces of $\mathbf{M}$ and $\mathbf{M}^{-1}$, when $\mathbf{M}$ is non-singular.

Corollary 2.1. *Let $\mathcal{R}$ reduce $\mathbf{M} \in \mathbb{R}^{r\times r}$, let $\mathbf{A} \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for $\mathcal{R}$, and let $\mathbf{A}_0$ be a semi-orthogonal basis matrix for $\mathcal{R}^{\perp}$. Then*

1. *$\mathbf{M}$ and $\mathbf{P}_{\mathcal{R}}$, and $\mathbf{M}$ and $\mathbf{Q}_{\mathcal{R}}$ commute.*
2. *$\mathcal{R} \subseteq \mathrm{span}(\mathbf{M})$ if and only if $\mathbf{A}^T\mathbf{M}\mathbf{A}$ is full rank.*
3. *If $\mathbf{M}$ is full rank, then*

$$\mathbf{M}^{-1} = \mathbf{A}(\mathbf{A}^T\mathbf{M}\mathbf{A})^{-1}\mathbf{A}^T + \mathbf{A}_0(\mathbf{A}_0^T\mathbf{M}\mathbf{A}_0)^{-1}\mathbf{A}_0^T. \tag{2.2}$$

As mentioned in the preamble to this section, there is a connection between the eigenstructure of a symmetric matrix $\mathbf{M}$ and its reducing subspaces. By definition, any invariant subspace of $\mathbf{M} \in \mathbb{S}^{r\times r}$ is also a reducing subspace of $\mathbf{M}$. In particular, it follows from Proposition 2.1 that the subspace spanned by any set of eigenvectors of $\mathbf{M}$ is a reducing subspace of $\mathbf{M}$. This connection is formalized as follows.

Proposition 2.2. *Let $\mathcal{R}$ be a subspace of $\mathbb{R}^r$ and let $\mathbf{M} \in \mathbb{S}^{r\times r}$. Assume that $\mathbf{M}$ has $q \leq r$ distinct eigenvalues, and let $\mathbf{P}_i$, $i = 1, \ldots, q$, indicate the projections on the corresponding eigenspaces. Then the following statements are equivalent:*

1. $\mathcal{R}$ *reduces* $\mathbf{M}$,
2. $\mathcal{R} = \oplus_{i=1}^{q} \mathbf{P}_i \mathcal{R}$,
3. $\mathbf{P}_{\mathcal{R}} = \sum_{i=1}^{q} \mathbf{P}_i \mathbf{P}_{\mathcal{R}} \mathbf{P}_i$,
4. $\mathbf{M}$ *and* $\mathbf{P}_{\mathcal{R}}$ *commute.*

2.2. M-envelopes

Since the intersection of two reducing subspaces of a matrix $\mathbf{M} \in \mathbb{S}^{r\times r}$ is itself a reducing subspace, it makes sense to talk about the smallest reducing subspace of $\mathbf{M}$ that contains a certain subspace $\mathcal{S}$.

Definition 2.1. Let $\mathbf{M} \in \mathbb{S}^{r\times r}$ and let $\mathcal{S} \subseteq \text{span}(\mathbf{M})$. The $\mathbf{M}$-envelope of $\mathcal{S}$, to be written as $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that contain $\mathcal{S}$.

This definition requires that $\mathcal{S} \subseteq \text{span}(\mathbf{M})$. Since the column space of $\mathbf{M}$ is itself a reducing subspace of $\mathbf{M}$, this containment guarantees existence of the $\mathbf{M}$-envelope, and is assumed throughout. Note that the containment holds trivially if $\mathbf{M}$ is full rank, i.e, if $\text{span}(\mathbf{M}) = \mathbb{R}^r$. Moreover, closure under intersection guarantees that the $\mathbf{M}$-envelope is in fact a reducing subspace of $\mathbf{M}$. Thus the $\mathbf{M}$-envelope of $\mathcal{S}$ can be interpreted as the unique smallest reducing subspace of $\mathbf{M}$ that contains $\mathcal{S}$, and represents a well-defined parameter in some statistical problems.

To develop some intuition on $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$, consider the case where all the r eigenvalues of $\mathbf{M}$ are distinct. Then, among the 2^r ways of dividing the eigenvectors of $\mathbf{M}$ into two groups, there is one and only one way in which one of the two groups spans a subspace of minimal dimension that contains $\mathcal{S}$. This minimal subspace is $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$. Thus, in this case, $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ is the smallest subspace that contains $\mathcal{S}$ and that is aligned with the eigenstructure of $\mathbf{M}$. Of course, the situation becomes more complicated if $\mathbf{M}$ has less than r distinct eigenvalues, and that is why we use reducing subspaces in the general definition of $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$.

The $\mathbf{M}$-envelope of any reducing subspace is the reducing subspace itself; that is, $\mathcal{E}_{\mathbf{M}}(\mathcal{R}) = \mathcal{R}$ if $\mathcal{R}$ reduces $\mathbf{M}$. A special case of this statement is that, for any subspace $\mathcal{S}$ of $\text{span}(\mathbf{M})$, $\mathcal{E}_{\mathbf{M}}(\mathcal{E}_{\mathbf{M}}(\mathcal{S})) = \mathcal{E}_{\mathbf{M}}(\mathcal{S})$. Thus, as an operator, $\mathcal{E}_{\mathbf{M}}(\cdot)$ is idempotent. Additionally, since an envelope is a reducing subspace, the results in Section 2.1 are applicable.

The following proposition, derived from Proposition 2.2 and Definition 2.1, gives a characterization of $\mathbf{M}$-envelopes.

Proposition 2.3. *Let* $\mathbf{M} \in \mathbb{S}^{r\times r}$, *let* $\mathbf{P}_i$, $i = 1, \ldots, q$, *be the projections onto the eigenspaces of* $\mathbf{M}$, *and let* $\mathcal{S}$ *be a subspace of* $\text{span}(\mathbf{M})$. *Then* $\mathcal{E}_{\mathbf{M}}(\mathcal{S}) = \oplus_{i=1}^{q} \mathbf{P}_i \mathcal{S}$.

We next investigate how the $\mathbf{M}$-envelope is modified by linear transformations of $\mathcal{S}$. While an envelope does not transform equivariantly for all linear transformations, it does so for symmetric linear transformations that commute with $\mathbf{M}$.

Proposition 2.4. *Let* $\mathbf{K} \in \mathbb{S}^{r\times r}$ *commute with* $\mathbf{M} \in \mathbb{S}^{r\times r}$*, and let* $\mathcal{S}$ *be a subspace of* $\operatorname{span}(\mathbf{M})$*. Then* $\mathbf{K}\mathcal{S} \subseteq \operatorname{span}(\mathbf{M})$ *and*

$$\mathcal{E}_{\mathbf{M}}(\mathbf{K}\mathcal{S}) = \mathbf{K}\mathcal{E}_{\mathbf{M}}(\mathcal{S}). \tag{2.3}$$

If, in addition, $\mathcal{S} \subseteq \operatorname{span}(\mathbf{K})$ *and* $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ *reduces* $\mathbf{K}$*, then*

$$\mathcal{E}_{\mathbf{M}}(\mathbf{K}\mathcal{S}) = \mathcal{E}_{\mathbf{M}}(\mathcal{S}). \tag{2.4}$$

We conclude this section by exploring a useful consequence of (2.4). Starting with any function $f : \mathbb{R} \to \mathbb{R}$, we can create $f^* : \mathbb{S}^{r\times r} \to \mathbb{S}^{r\times r}$ as follows. Let m_i and $\mathbf{P}_i$, $i = 1, \ldots, q$, indicate the distinct eigenvalues and the projections on the corresponding eigenspaces for a matrix $\mathbf{M} \in \mathbb{S}^{r\times r}$, and let $f^*(\mathbf{M}) = \sum_{i=1}^{q} f(m_i)\mathbf{P}_i$. If $f(\cdot)$ is such that $f(0) = 0$ and $f(x) \neq 0$ whenever $x \neq 0$, then it is easy to verify that (i) $f^*(\mathbf{M})$ commutes with $\mathbf{M}$, (ii) any subspace $\mathcal{S} \subseteq \operatorname{span}(\mathbf{M})$ satisfies $\mathcal{S} \subseteq \operatorname{span}\{f^*(\mathbf{M})\}$, and (iii) $\mathcal{E}_{\mathbf{M}}(\mathcal{S})$ reduces $f^*(\mathbf{M})$. Hence, by Proposition 2.4 we have $\mathcal{E}_{\mathbf{M}}(f^*(\mathbf{M})\mathcal{S}) = \mathcal{E}_{\mathbf{M}}(\mathcal{S})$. In particular, this guarantees invariance for any power of $\mathbf{M}$:

$$\mathcal{E}_{\mathbf{M}}(\mathbf{M}^k\mathcal{S}) = \mathcal{E}_{\mathbf{M}}(\mathcal{S}) \text{ for all } k \in \mathbb{R}. \tag{2.5}$$

3. Envelope Models

3.1. Theoretical formulation of envelope models

We are now in a position to refine model (1.1) by using an envelope to connect $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$. Let $\mathcal{B} = \operatorname{span}(\boldsymbol{\beta})$, $d = \dim(\mathcal{B})$ and, to exclude the trivial case, assume $d > 0$. Consider the $\boldsymbol{\Sigma}$-envelope of $\mathcal{B}$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, of dimension u, so that $0 < d \leq u \leq r$. We use this envelope as a well-defined parameter to link the mean and variance structures of the multivariate linear model. Since $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is unknown, it needs to be estimated, and this is facilitated by writing formal model statements that incorporate it as a parameter. We give two such statements: a coordinate-free version that uses $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ as the parameter, and a coordinate version that uses a semi-orthogonal basis matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{r\times u}$ for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. Each has advantages, depending on the phase of the analysis. For instance, the coordinate version is necessary for computation. Our use of "coordinate-free" and "coordinate" terminology applies only to the representation of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, and not to the rest of the model.

Since $\mathbf{\Sigma}$ is a positive definite matrix reduced by $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, all of the results in Section 2 apply. In particular, $\mathbf{\Sigma}$ can be written in the form given by Proposition 2.1 with $\mathcal{R} = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, its inverse can be expressed as in part 3 of Corollary 2.1, and $\mathbf{\Sigma}^k\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathbf{\Sigma}^k\mathcal{B}) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ for all $k \in \mathbb{R}$, because of Proposition 2.4. The following corollary gives a coordinate-free version of Proposition 2.1, making use of the additional properties characterizing a covariance matrix.

Corollary 3.1. *A subspace $\mathcal{R}$ of $\mathbb{R}^r$ reduces $\mathbf{\Sigma}$ if and only if $\mathbf{\Sigma}$ can be written in the form $\mathbf{\Sigma} = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2$, where $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ are symmetric positive semi-definite matrices such that $\mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{0}$ and $\mathcal{R} = \mathrm{span}(\mathbf{\Sigma}_1)$.*

The coordinate-free representation of the envelope model is model (1.1) with error covariance matrix satisfying

$$\mathbf{\Sigma} = \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2,\ \mathbf{\Sigma}_1\mathbf{\Sigma}_2 = \mathbf{0},\ \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathrm{span}(\mathbf{\Sigma}_1). \tag{3.1}$$

Since reducing subspaces are specified by this decomposition of $\mathbf{\Sigma}$, we could equivalently replace the requirement $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathrm{span}(\mathbf{\Sigma}_1)$ with the condition that $\mathrm{span}(\mathbf{\Sigma}_1)$ has minimal dimension under the constraint $\mathcal{B} \subseteq \mathrm{span}(\mathbf{\Sigma}_1)$. However, it is important to note that (3.1), *per se*, does not restrict the scope of model (1.1). If $u = r$, then we must have $\mathbf{\Sigma}_1 = \mathbf{\Sigma}$ and $\mathbf{\Sigma}_2 = \mathbf{0}$. If $r \le p$ and $d = r$, then the envelope model coincides with the standard multivariate linear model, since there are evidently no linear redundancies in (1.1), and thus no reduction is possible with the new parameterization. On the other hand, if $u < r$ there is a potential for the envelope model expressed through (3.1) to yield substantial gains. As an extension of the ideas presented here, alternative uses of envelopes that allow reduction when $r \le p$ and $d = r$ are described in Section 8.4.

To write the coordinate version of the envelope model, let $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, and let $(\mathbf{\Gamma}, \mathbf{\Gamma}_0) \in \mathbb{R}^{r\times r}$ be an orthogonal matrix. Then there is an $\boldsymbol{\eta} \in \mathbb{R}^{u\times p}$ such that $\boldsymbol{\beta} = \mathbf{\Gamma}\boldsymbol{\eta}$. Additionally, let $\mathbf{\Omega} = \mathbf{\Gamma}^T\mathbf{\Sigma}\mathbf{\Gamma} \in \mathbb{S}^{u\times u}$ and let $\mathbf{\Omega}_0 = \mathbf{\Gamma}_0^T\mathbf{\Sigma}\mathbf{\Gamma}_0 \in \mathbb{S}^{(r-u)\times(r-u)}$. Then, using Proposition 2.1 and Corollary 3.1, we can write

$$\begin{aligned}\mathbf{Y} &= \boldsymbol{\alpha} + \mathbf{\Gamma}\boldsymbol{\eta}\mathbf{X} + \boldsymbol{\varepsilon},\\ \mathbf{\Sigma} &= \mathbf{\Sigma}_1 + \mathbf{\Sigma}_2 = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T,\end{aligned} \tag{3.2}$$

where $\boldsymbol{\varepsilon}$ is normally distributed with mean $\mathbf{0}$ and variance $\mathbf{\Sigma}$. The matrices $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ can be thought of as *coordinate matrices*, since they carry the coordinates of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ relative to $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$, just as $\boldsymbol{\eta}$ contains the coordinates of $\boldsymbol{\beta}$ relative to $\mathbf{\Gamma}$.

The total number N of parameters needed to estimate (3.2) is

$$N = r + pu + u(r-u) + \frac{u(u+1)}{2} + \frac{(r-u)(r-u+1)}{2}.$$

The first term on the right hand side corresponds to the intercept $\boldsymbol{\alpha} \in \mathbb{R}^r$; the second term corresponds to the unconstrained coordinate matrix $\boldsymbol{\eta} \in \mathbb{R}^{u \times p}$; the last two terms correspond to $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$. Their parameter counts arise because, for any integer $k > 0$, it takes $k(k+1)/2$ numbers to specify a nonsingular matrix in $\mathbb{S}^{k \times k}$. The third term, $u(r-u)$, which corresponds roughly to $\boldsymbol{\Gamma}$, arises as follows. The matrix $\boldsymbol{\Gamma}$ is not identified, since, for any orthogonal matrix $\mathbf{A}$, replacing $\boldsymbol{\Gamma}$ with $\boldsymbol{\Gamma}\mathbf{A}$ results in an equivalent model. However, $\operatorname{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is identified and estimable. The parameter space for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is a Grassmann manifold $\mathbb{G}^{r \times u}$ of dimension u in $\mathbb{R}^r$; that is, the collection of all u-dimensional subspaces of $\mathbb{R}^r$. From basic properties of Grassmann manifolds it is known that $u(r-u)$ parameters are needed to specify an element of $\mathbb{G}^{r \times u}$ (Edelman, Tomás and Smith (1998)). Once $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is determined, so is its orthogonal complement $\operatorname{span}(\boldsymbol{\Gamma}_0)$, and no additional free parameters are required.

Simplifying the above expression for N, we obtain $N = r + pu + r(r+1)/2$. The difference between the total parameter count for the full model (1.1) with $r = u$ and the envelope model (3.2) with $u < r$ is therefore $p(r-u)$.

Note that a specific envelope model is identified by the value of u, with the full model (1.1) occurring when $u = r$. All envelope models are nested within the full model, but two envelope models with different values of u are not necessarily nested. To see this, it is enough to realize that the number of free parameters needed to specify an element of $\mathbb{G}^{r \times u}$ is the same for $u = 1$ and $u = r - 1$. In full generality, u is a model selection parameter that can be chosen using traditional reasoning, as discussed in Section 7.1.

3.2. Alternative envelopes for random designs

The models introduced so far are parameterized in terms of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, the $\boldsymbol{\Sigma}$-envelope of $\mathcal{B}$, in coordinate-free and coordinate versions. While this seems to be the natural route when $\mathbf{X}$ is chosen by design, other choices are available when $\mathbf{X}$ is random. For instance, we might create a parameterization in terms of $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}}}(\mathcal{B})$, the envelope of $\mathcal{B}$ based on the marginal response covariance matrix $\boldsymbol{\Sigma}_{\mathbf{Y}} = \operatorname{Var}(\mathbf{Y})$. The next proposition states the equality of several envelopes. The first equality shows an important equivalence between enveloping in reference to the error variability $\boldsymbol{\Sigma}$ and the response variability $\boldsymbol{\Sigma}_{\mathbf{Y}}$. The other equalities will be relevant in Section 8.

Proposition 3.1. *Assume model (1.1). Then* $\boldsymbol{\Sigma}^{-1}\mathcal{B} = \boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathcal{B}$, *and*

$$\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}}}(\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}^{-1}\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}}}(\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{Y}}}(\boldsymbol{\Sigma}^{-1}\mathcal{B}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\boldsymbol{\Sigma}_{\mathbf{Y}}^{-1}\mathcal{B}).$$

4. Maximum Likelihood Estimation

Before deriving the MLEs for the envelope model, we give a few preliminary results in Section 4.1. These are intended primarily to facilitate derivations in Section 4.2 but, like the results in Section 2, may have wider applicability. The calculations necessary to obtain the estimates are summarized in Section 4.3.

4.1. Preliminary results

Lemma 4.1. *Let* $\mathbf{U} \in \mathbb{R}^{n\times p}$*,* $\mathbf{V} \in \mathbb{R}^{n\times r}$*, and* $\mathbf{W} \in \mathbb{R}^{p\times d}$ *be known matrices. Let* $\mathbf{\Lambda}$ *be a positive semi-definite matrix in* $\mathbb{R}^{p\times p}$ *such that* $\mathrm{span}(\mathbf{W}) \subseteq \mathrm{span}(\mathbf{\Lambda})$*. Then the minimizer of*

$$\mathrm{tr}\left[(\mathbf{U}-\mathbf{A})\mathbf{\Lambda}(\mathbf{U}-\mathbf{A})^T\right] \tag{4.1}$$

over the set of matrices $\mathcal{A} = \{\mathbf{A} :\ \mathrm{span}(\mathbf{A}) \subseteq \mathrm{span}(\mathbf{V}),\ \mathrm{span}(\mathbf{A}^T) \subseteq \mathrm{span}(\mathbf{W})\}$ *is* $\mathbf{A}^* = \mathbf{P}_{\mathbf{V}}\mathbf{U}\mathbf{P}^T_{\mathbf{W}(\mathbf{\Lambda})}$*, and the corresponding minimum of (4.1) is*

$$\mathrm{tr}(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T) - \mathrm{tr}(\mathbf{P}_{\mathbf{V}}\mathbf{U}\mathbf{P}^T_{\mathbf{W}(\mathbf{\Lambda})}\mathbf{\Lambda}\mathbf{P}_{\mathbf{W}(\mathbf{\Lambda})}\mathbf{U}^T\mathbf{P}_{\mathbf{V}}).$$

For a nonzero $\mathbf{A} \in \mathbb{S}^{r\times r}$ (i.e., an $r \times r$ symmetric matrix whose entries are not all equal to 0), we denote by $\mathrm{det}_0(\mathbf{A})$ the product of its non-zero eigenvalues. Note that, for any constant c, $\mathrm{det}_0(c\mathbf{A}) = c^k\mathrm{det}_0(\mathbf{A})$, where k is the rank of $\mathbf{A}$. The next lemma facilitates analysis with the structure introduced in Corollary 3.1.

Lemma 4.2. *If* $\mathbf{A}_1$ *and* $\mathbf{A}_2$ *are nonzero symmetric matrices such that* $\mathbf{A}_1\mathbf{A}_2 = \mathbf{0}$*, then*

1. $\mathrm{det}_0(\mathbf{A}_1 + \mathbf{A}_2) = \mathrm{det}_0(\mathbf{A}_1) \times \mathrm{det}_0(\mathbf{A}_2)$,
2. $(\mathbf{A}_1 + \mathbf{A}_2)^\dagger = \mathbf{A}_1^\dagger + \mathbf{A}_2^\dagger$*, and*
3. $(\mathbf{A}_1 + \mathbf{A}_2)^r = \mathbf{A}_1^r + \mathbf{A}_2^r$*, for any* $r > 0$*.*

Finally, we introduce a lemma that gives an explicit expression for the MLE of the covariance matrix in a multivariate normal likelihood when the column space of the covariance is fixed and the mean is known.

Lemma 4.3. *Let* $\mathcal{A}$ *be a class of* $p \times p$ *positive semi-definite matrices having the same column space of dimension* k*,* $0 < k \le p$*, and let* $\mathbf{P}$ *be the projection onto the common column space. Let* $\mathbf{U}$ *be a matrix in* $\mathbb{R}^{n\times p}$*, and let* $L(\mathbf{A}) = [\mathrm{det}_0(\mathbf{A})]^{-n/2}e^{-1/2\,\mathrm{tr}(\mathbf{U}\mathbf{A}^\dagger\mathbf{U}^T)}$*. Then the maximizer of* $L(\mathbf{A})$ *over* $\mathcal{A}$ *is the matrix* $n^{-1}\mathbf{P}\mathbf{U}^T\mathbf{U}\mathbf{P}$*, and the maximum value of* $L(\mathbf{A})$ *is* $n^{nk/2}e^{-nk/2}[\mathrm{det}_0(\mathbf{P}\mathbf{U}^T\mathbf{U}\mathbf{P})]^{-n/2}$*.*

4.2. Coordinate-free representation of the MLE

Derivation of the MLE is easier using the coordinate-free representation of the envelope model, as given by (1.1) and (3.1). We assume that the observations $\mathbf{Y}_i$, $i = 1, \ldots, n$, are independent, and that $\mathbf{Y}_i$ is sampled from the conditional distribution of $\mathbf{Y}|\mathbf{X}_i$, $i = 1, \ldots, n$, with $\bar{\mathbf{X}} = 0$. We assume also that $n > r + p$. Let $\mathbf{G}$ be the $n \times r$ matrix whose ith row is $\mathbf{Y}_i^T$, $\mathbf{F}$ be the $n \times p$ matrix whose ith row is $\mathbf{X}_i^T$, and $\mathbf{1}_n$ be the $n \times 1$ vector with all entries 1.

For a $\boldsymbol{\Sigma}$-envelope with fixed dimension u, $0 < u < r$, the likelihood based on $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ is

$$\begin{aligned} L^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2) &= [\det(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)]^{-n/2} \\ &\quad \times \operatorname{etr}[-\frac{1}{2}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)(\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2)^{-1}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)^T], \end{aligned} \tag{4.2}$$

where $\operatorname{etr}(\cdot)$ denotes the composite function $\exp \circ \operatorname{tr}(\cdot)$, and $\otimes$ the Kronecker product. This likelihood is to be maximized over $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ subject to the constraints

$$\operatorname{span}(\boldsymbol{\beta}) \subseteq \operatorname{span}(\boldsymbol{\Sigma}_1), \quad \boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2 = \mathbf{0}. \tag{4.3}$$

By Lemma 4.2, and using the relation $\boldsymbol{\Sigma}_2\boldsymbol{\beta} = \mathbf{0}$, the likelihood in (4.2) can be factored as $L_1^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_1) \times L_2^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_2)$, where

$$\begin{aligned} L_1^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}_1) = &\ [\det{}_0(\boldsymbol{\Sigma}_1)]^{-n/2} \\ &\times \operatorname{etr}[-\frac{1}{2}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)\boldsymbol{\Sigma}_1^\dagger(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n - \mathbf{F}\boldsymbol{\beta}^T)^T], \\ L_2^{(u)}(\boldsymbol{\alpha}, \boldsymbol{\Sigma}_2) = &\ [\det{}_0(\boldsymbol{\Sigma}_2)]^{-n/2} \times \operatorname{etr}[-\frac{1}{2}(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n)\boldsymbol{\Sigma}_2^\dagger(\mathbf{G} - \boldsymbol{\alpha}^T \otimes \mathbf{1}_n)^T]. \end{aligned} \tag{4.4}$$

Based on this factorization and the constraints in (4.3), we can decompose the likelihood maximization into the following steps.

1. Fix $\boldsymbol{\Sigma}_1$, $\boldsymbol{\Sigma}_2$, and $\boldsymbol{\beta}$, and maximize $L^{(u)}$ in (4.2) over $\boldsymbol{\alpha}$; substitute the optimal $\boldsymbol{\alpha}$ into $L_1^{(u)}$ and $L_2^{(u)}$ in (4.4) to obtain $L_{11}^{(u)}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1)$ and $L_{21}^{(u)}(\boldsymbol{\Sigma}_2)$. The required maximizer is the sample mean of $\{\mathbf{Y}_i - \boldsymbol{\beta}\mathbf{X}_i : i = 1, \ldots, n\}$ which, because $\mathbf{X}$ has sample mean zero, is simply $\bar{\mathbf{Y}}$. Hence, if we let $\mathbf{U}$ be the $n \times r$ matrix whose ith row is $(\mathbf{Y}_i - \bar{\mathbf{Y}})^T$, the partially maximized $L_1^{(u)}$ and $L_2^{(u)}$ are

$$\begin{aligned} L_{11}^{(u)}(\boldsymbol{\beta}, \boldsymbol{\Sigma}_1) =& [\det{}_0(\boldsymbol{\Sigma}_1)]^{-n/2} \times \operatorname{etr}[-\frac{1}{2}(\mathbf{U} - \mathbf{F}\boldsymbol{\beta}^T)\boldsymbol{\Sigma}_1^\dagger(\mathbf{U} - \mathbf{F}\boldsymbol{\beta}^T)^T], \\ L_{21}^{(u)}(\boldsymbol{\Sigma}_2) =& [\det{}_0(\boldsymbol{\Sigma}_2)]^{-n/2} \times \operatorname{etr}(-\frac{1}{2}\mathbf{U}\boldsymbol{\Sigma}_2^\dagger\mathbf{U}^T). \end{aligned} \tag{4.5}$$

2. Fix $\mathbf{\Sigma}_1$, and further maximize the function $L_{11}^{(u)}$ from step 1 over $\boldsymbol{\beta}$, subject to the first constraint in (4.3), to obtain $L_{12}^{(u)}(\mathbf{\Sigma}_1)$. For this maximization we use Lemma 4.1, with the relevant quadratic form given by
$$\mathrm{tr}[(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T)\mathbf{\Sigma}_1^{\dagger}(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T)^T] \equiv \mathrm{tr}[(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T\mathbf{I}_r)\mathbf{\Sigma}_1^{\dagger}(\mathbf{U}-\mathbf{F}\boldsymbol{\beta}^T\mathbf{I}_r)^T].$$
Thus, the optimal $\mathbf{F}\boldsymbol{\beta}^T\mathbf{I}_r$ is $\mathbf{P_F}\mathbf{U}\mathbf{P}^T_{\mathbf{I}_r(\mathbf{\Sigma}_1^{\dagger})} = \mathbf{P_F}\mathbf{U}\mathbf{P}_{\mathbf{\Sigma}_1}$. This implies that
$$\boldsymbol{\beta} = \mathbf{P}_{\mathbf{\Sigma}_1}\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}, \tag{4.6}$$
where $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}} = \mathbf{U}^T\mathbf{F}(\mathbf{F}^T\mathbf{F})^{-1}$ is the MLE of $\boldsymbol{\beta}$ from the full model (1.1). Consequently, we see that $\widehat{\boldsymbol{\beta}}$ is the projection of $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$ onto the MLE of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$. Substituting this into (4.5), and using the relation $\mathbf{P}_{\mathbf{\Sigma}_1}\mathbf{\Sigma}_1^{\dagger} = \mathbf{\Sigma}_1^{\dagger}$, we see that the maximum of $L_{11}^{(u)}(\boldsymbol{\beta},\mathbf{\Sigma}_1)$ for fixed $\mathbf{\Sigma}_1$ over $\boldsymbol{\beta}$ is
$$\begin{aligned} L_{12}^{(u)}(\mathbf{\Sigma}_1) =& [\det{}_0(\mathbf{\Sigma}_1)]^{-n/2} \times \mathrm{etr}[-\frac{1}{2}(\mathbf{U}-\mathbf{P_F}\mathbf{U})\mathbf{\Sigma}_1^{\dagger}(\mathbf{U}-\mathbf{P_F}\mathbf{U})^T] \\ =& [\det{}_0(\mathbf{\Sigma}_1)]^{-n/2} \times \mathrm{etr}(-\frac{1}{2}\mathbf{Q_F}\mathbf{U}\mathbf{\Sigma}_1^{\dagger}\mathbf{U}^T\mathbf{Q_F}), \end{aligned} \tag{4.7}$$
where $\mathbf{Q_F} = \mathbf{I}_n - \mathbf{P_F}$.

3. Using Lemma 4.3, maximize $L_{12}^{(u)}(\mathbf{\Sigma}_1)$ over all $\mathbf{\Sigma}_1$'s having the same column space to obtain $L_{13}^{(u)}(\mathbf{P}_{\mathbf{\Sigma}_1})$, which is proportional to $[\det_0(\mathbf{P}_{\mathbf{\Sigma}_1}\mathbf{U}^T\mathbf{Q_F}\mathbf{U}\,\mathbf{P}_{\mathbf{\Sigma}_1})]^{-n/2}$. Similarly, maximize $L_{21}^{(u)}(\mathbf{\Sigma}_2)$ over all $\mathbf{\Sigma}_2$'s having the same column space to obtain $L_{22}^{(u)}(\mathbf{P}_{\mathbf{\Sigma}_2})$, which is proportional to $\left[\det_0(\mathbf{P}_{\mathbf{\Sigma}_2}\mathbf{U}^T\mathbf{U}\mathbf{P}_{\mathbf{\Sigma}_2})\right]^{-n/2}$. Note that $L_{13}^{(u)}$ depends only on the column space of $\mathbf{\Sigma}_1$, and $L_{22}^{(u)}$ only on the column space of $\mathbf{\Sigma}_2$.

4. Optimize the partially maximized likelihood $L_{13}^{(u)}(\mathbf{P}_{\mathbf{\Sigma}_1}) \times L_{22}^{(u)}(\mathbf{P}_{\mathbf{\Sigma}_2})$, which is proportional to
$$\begin{aligned} &[\det{}_0(\mathbf{P}_{\mathbf{\Sigma}_1}\mathbf{U}^T\mathbf{Q_F}\mathbf{U}\mathbf{P}_{\mathbf{\Sigma}_1})]^{-n/2} \times [\det{}_0(\mathbf{P}_{\mathbf{\Sigma}_2}\mathbf{U}^T\mathbf{U}\mathbf{P}_{\mathbf{\Sigma}_2})]^{-n/2} \\ &\qquad = [\det{}_0(\mathbf{P}_{\mathbf{\Sigma}_1}\mathbf{U}^T\mathbf{Q_F}\mathbf{U}\mathbf{P}_{\mathbf{\Sigma}_1} + \mathbf{P}_{\mathbf{\Sigma}_2}\mathbf{U}^T\mathbf{U}\mathbf{P}_{\mathbf{\Sigma}_2})]^{-n/2}. \end{aligned} \tag{4.8}$$
Because $\mathbf{P}_{\mathbf{\Sigma}_2} = \mathbf{I}_r - \mathbf{P}_{\mathbf{\Sigma}_1} = \mathbf{Q}_{\mathbf{\Sigma}_1}$, the above depends on $\mathbf{P}_{\mathbf{\Sigma}_1}$ alone. Additionally, $\mathbf{U}^T\mathbf{U}$ is n times the marginal sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ of the responses, and $\mathbf{U}^T\mathbf{Q_F}\mathbf{U}$ is n times the sample covariance matrix $\widehat{\mathbf{\Sigma}}_{\mathrm{res}}$ of the residuals from the fit of the full model (1.1). Since we have assumed that $n > r + p$, it follows that $\mathrm{rank}(\widehat{\mathbf{\Sigma}}_{\mathrm{res}}) = \mathrm{rank}(\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}) = r$ with probability 1. Therefore $\det_0(\cdot)$ in (4.8) can be replaced by $\det(\cdot)$, the usual determinant, and we need to minimize the function
$$D = D(\mathrm{span}(\mathbf{\Sigma}_1)) \equiv \det(\mathbf{P}_{\mathbf{\Sigma}_1}\widehat{\mathbf{\Sigma}}_{\mathrm{res}}\mathbf{P}_{\mathbf{\Sigma}_1} + \mathbf{Q}_{\mathbf{\Sigma}_1}\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}\mathbf{Q}_{\mathbf{\Sigma}_1}) \tag{4.9}$$

over the Grassmann manifold $\mathbb{G}^{r\times u}$, subject to the constraint that rank $(\mathbf{P}_{\mathbf{\Sigma}_1}\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbf{P}_{\mathbf{\Sigma}_1}) = u$ – which arises because $\text{rank}(\mathbf{\Sigma}_1) = u < r$.

4.3. Implementation of the MLE

The MLE described in Section 4.2 hinges on being able to minimize $\log D$ over the Grassmann manifold $\mathbb{G}^{r\times u}$, where D is as defined in (4.9). Available gradient-based algorithms for Grassmann optimization (see Edelman, Tomás and Smith (1998); Liu, Srivastava and Gallivan (2004)) require a coordinate version of the objective function which must have continuous directional derivatives. A coordinate version of objective function (4.9) satisfies this continuity requirement when $\mathbf{\Sigma} > \mathbf{0}$. Recall that $\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0$ are semi-orthogonal basis matrices of $\text{span}(\mathbf{\Sigma}_1) = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ and its orthogonal complement, respectively. Let $\widehat{\mathbf{\Gamma}}$ and $\widehat{\mathbf{\Gamma}}_0$ be semi-orthogonal bases for $\text{span}(\widehat{\mathbf{\Sigma}}_1)$ and its orthogonal complement. Then $\widehat{\boldsymbol{\eta}} = \widehat{\mathbf{\Gamma}}^T\widehat{\boldsymbol{\beta}}_{\text{fm}}$, $\widehat{\mathbf{\Omega}} = \widehat{\mathbf{\Gamma}}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\widehat{\mathbf{\Gamma}}$, and $\widehat{\mathbf{\Omega}}_0 = \widehat{\mathbf{\Gamma}}_0^T\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}\widehat{\mathbf{\Gamma}}_0$. Since $\widehat{\mathbf{\Sigma}}_{\text{res}}$ and $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ have rank r almost surely, the matrices $\mathbf{\Gamma}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbf{\Gamma}$ and $\mathbf{\Gamma}_0^T\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}\mathbf{\Gamma}_0$ are positive definite almost surely. Let $\log\det(\cdot)$ denote the composite function $\log\circ\det(\cdot)$. Then the coordinate form of $\log D$ is

$$\begin{aligned}\log D &= \log\det[\mathbf{\Gamma}\mathbf{\Gamma}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbf{\Gamma}\mathbf{\Gamma}^T + (\mathbf{I}_r - \mathbf{\Gamma}\mathbf{\Gamma}^T)\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}(\mathbf{I}_r - \mathbf{\Gamma}\mathbf{\Gamma}^T)]\\ &= \log\det(\mathbf{\Gamma}^T\widehat{\mathbf{\Sigma}}_{\text{res}}\mathbf{\Gamma}) + \log\det(\mathbf{\Gamma}_0^T\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}\mathbf{\Gamma}_0). \end{aligned} \tag{4.10}$$

In summary, maximum likelihood estimation for the parameters involved in the envelope model can be implemented as follows.

a. Obtain the sample version $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$ of the marginal covariance matrix of $\mathbf{Y}$, and obtain the residual covariance matrix $\widehat{\mathbf{\Sigma}}_{\text{res}}$ and the MLE $\widehat{\boldsymbol{\beta}}_{\text{fm}}$ of $\boldsymbol{\beta}$ from the fit of the full model (1.1).
b. Estimate $\mathbf{P}_{\mathbf{\Sigma}_1}$ by minimizing the objective function (4.10) over the Grassmann manifold $\mathbb{G}^{r\times u}$, and denote the result by $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}$. Estimate $\mathbf{P}_{\mathbf{\Sigma}_2}$ by $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_2} = \mathbf{I}_r - \widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}$.
c. Estimate $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = \widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}\widehat{\boldsymbol{\beta}}_{\text{fm}}$.
d. Estimate $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$ by $\widehat{\mathbf{\Sigma}}_1 = \widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}\widehat{\mathbf{\Sigma}}_{\text{res}}\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}$ and $\widehat{\mathbf{\Sigma}}_2 = (\mathbf{I}_r - \widehat{\mathbf{P}}_{\mathbf{\Sigma}_1})\widehat{\mathbf{\Sigma}}_Y(\mathbf{I}_r - \widehat{\mathbf{P}}_{\mathbf{\Sigma}_1})$.

We assumed at the outset of this derivation that $u < r$. If $u = r$ then $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1} = \mathbf{I}_r$ and $\widehat{\boldsymbol{\beta}}$ reduces to the usual MLE based on (1.1). Generally, objective functions defined on Grassmann manifolds can have multiple local optima, but we have not noticed local minima to be an issue for (4.10).

5. Asymptotic Variances

There is a multitude of approaches for dealing with dimensionality issues in multivariate regression. Many of these, ranging from various versions of principal components to a multivariate implementation of sliced inverse regression (Li et al. (2003)) are algorithmic in nature, making it difficult to determine post-application standard errors and other inference-related quantities. Unlike these approaches, our analysis of envelope models is based entirely on the likelihood. We are therefore able to pursue inference classically, with methodology that inherits optimal properties from general likelihood theory.

5.1. Estimable functions

The parameters in the coordinate representation (3.2) of the envelope model can be combined into the vector

$$\boldsymbol{\phi} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\eta}) \\ \mathrm{vec}(\boldsymbol{\Gamma}) \\ \mathrm{vech}(\boldsymbol{\Omega}) \\ \mathrm{vech}(\boldsymbol{\Omega}_0) \end{pmatrix} \equiv \begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \\ \boldsymbol{\phi}_3 \\ \boldsymbol{\phi}_4 \end{pmatrix}, \tag{5.1}$$

where the "vector" operator vec : $\mathbb{R}^{r\times p} \to \mathbb{R}^{rp}$ stacks the columns of the argument matrix. On the symmetric matrices $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ we use the related "vector half" operator vech : $\mathbb{S}^{r\times r} \to \mathbb{R}^{r(r+1)/2}$, which extracts their unique elements (vech stacks only the unique part of each column that lies on or below the diagonal). vec and vech are related through a "contraction" matrix $\mathbf{C}_r \in \mathbb{R}^{r(r+1)/2\times r^2}$ and an "expansion" matrix $\mathbf{E}_r \in \mathbb{R}^{r^2\times r(r+1)/2}$, which are defined so that $\mathrm{vech}(\mathbf{A}) = \mathbf{C}_r\mathrm{vec}(\mathbf{A})$ and $\mathrm{vec}(\mathbf{A}) = \mathbf{E}_r\mathrm{vech}(\mathbf{A})$ for any $\mathbf{A} \in \mathbb{S}^{r\times r}$. These relations uniquely define $\mathbf{C}_r$ and $\mathbf{E}_r$, and imply $\mathbf{C}_r\mathbf{E}_r = \mathbf{I}_{r(r+1)/2}$. For further background on these operators, see Henderson and Searle (1979).

Selected elements of $\boldsymbol{\phi}$ might be of interest in some applications, but here we focus on some specific estimable functions under the envelope model:

$$\mathbf{h}(\boldsymbol{\phi}) \equiv \begin{pmatrix} \mathrm{vec}(\boldsymbol{\beta}) \\ \mathrm{vech}(\boldsymbol{\Sigma}) \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(\boldsymbol{\Gamma\eta}) \\ \mathrm{vech}(\boldsymbol{\Gamma\Omega\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) \end{pmatrix} \equiv \begin{pmatrix} \mathbf{h}_1(\boldsymbol{\phi}) \\ \mathbf{h}_2(\boldsymbol{\phi}) \end{pmatrix}.$$

We have neglected the intercept $\boldsymbol{\alpha}$ in this setup. This induces no loss of generality because the intercept is not involved in $\mathbf{h}$, and its maximum likelihood estimate is asymptotically independent of the other parameter estimates.

If the gradient matrix

$$\mathbf{H} = \begin{pmatrix} \dfrac{\partial \mathbf{h}_1}{\partial \boldsymbol{\phi}_1^T} \cdots \dfrac{\partial \mathbf{h}_1}{\partial \boldsymbol{\phi}_4^T} \\ \dfrac{\partial \mathbf{h}_2}{\partial \boldsymbol{\phi}_1^T} \cdots \dfrac{\partial \mathbf{h}_2}{\partial \boldsymbol{\phi}_4^T} \end{pmatrix} \tag{5.2}$$

were of full rank when evaluated at the true parameter values, then standard methods could be used to find the asymptotic covariance matrices for $\hat{\mathbf{h}}_1 = \mathbf{h}_1(\hat{\boldsymbol{\phi}})$ and $\hat{\mathbf{h}}_2 = \mathbf{h}_2(\hat{\boldsymbol{\phi}})$. However, because of the over-parameterization in $\boldsymbol{\Gamma}$, $\mathbf{H}$ is not of full rank, and standard methods do not apply directly. Nevertheless, $\mathbf{h}$ is identified and estimable, which enables us to use a result by Shapiro (1986, Proposition 4.1) to derive the asymptotic distribution and efficiency gain of the envelope model, as given by the following theorem.

Theorem 5.1. *Suppose* $\bar{\mathbf{X}} = \mathbf{0}$. *Let* $\mathbf{J}$ *be the Fisher information for* $(\mathrm{vec}^T(\boldsymbol{\beta}), \mathrm{vech}^T(\boldsymbol{\Sigma}))^T$ *in the full model* (1.1)*:*

$$\mathbf{J} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} & \mathbf{0} \\ \mathbf{0} & \frac{1}{2}\mathbf{E}_r^T(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{\Sigma}^{-1})\mathbf{E}_r \end{pmatrix},$$

where $\boldsymbol{\Sigma}_{\mathbf{X}} = \lim_{n\to\infty} \sum_{i=1}^n \mathbf{X}_i\mathbf{X}_i^T/n$, *and let* $\mathbf{V} = \mathbf{J}^{-1}$ *be the asymptotic variance of the MLE under the full model. Then*

$$\sqrt{n}(\hat{\mathbf{h}} - \mathbf{h}) \xrightarrow{\mathcal{D}} N(\mathbf{0}, \mathbf{V}_0), \tag{5.3}$$

where $\mathbf{V}_0 = \mathbf{H}(\mathbf{H}^T\mathbf{J}\mathbf{H})^\dagger\mathbf{H}^T$ *and* $\mathbf{H}$ *is given by*

$$\begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_r & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \mathbf{I}_r - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T) & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{(r-u)} \end{pmatrix}. \tag{5.4}$$

Moreover, $\mathbf{V}^{-1/2}(\mathbf{V} - \mathbf{V}_0)\mathbf{V}^{-1/2} = \mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}} \geq 0$, *so the envelope model decreases the asymptotic variance by the fraction* $\mathbf{Q}_{\mathbf{J}^{1/2}\mathbf{H}}$.

We next present an alternative form for $\mathbf{V}_0$ that may facilitate computing and that will be helpful in the next section. Since $\mathbf{V}_0$ in (5.3) depends only on the column space of $\mathbf{H}$ we can replace $\mathbf{H}$ by any matrix $\mathbf{H}_1$ that has the same column space as $\mathbf{H}$. The most convenient and interpretable choice of $\mathbf{H}_1$ is one that makes $\mathbf{H}_1^T\mathbf{J}\mathbf{H}_1$ block-diagonal, with blocks corresponding to the parameters in (5.1). We now give such a construction. Let $\mathbf{H}_1$ be the $\{pr + r(r+1)/2\} \times \{pu + r(r+1)/2\}$ matrix

$$\begin{aligned} \mathbf{H}_1 &= \begin{pmatrix} \mathbf{I}_p \otimes \boldsymbol{\Gamma} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_r(\boldsymbol{\Gamma}\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0 - \boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0) & \mathbf{C}_r(\boldsymbol{\Gamma} \otimes \boldsymbol{\Gamma})\mathbf{E}_u & \mathbf{C}_r(\boldsymbol{\Gamma}_0 \otimes \boldsymbol{\Gamma}_0)\mathbf{E}_{r-u} \end{pmatrix} \\ &\equiv \begin{pmatrix} \mathbf{H}_{11} \, \mathbf{H}_{12} \, \mathbf{H}_{13} \, \mathbf{H}_{14} \end{pmatrix}, \end{aligned} \tag{5.5}$$

and let $\mathbf{H}_2$ be the $\{pu + r(r+1)/2\} \times \{pu + r(r+1)/2 + u^2\}$ matrix whose blocks conform to those of $\mathbf{H}_1$:

$$\mathbf{H}_2 = \begin{pmatrix} \mathbf{I}_{pu} & \boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_u \otimes \boldsymbol{\Gamma}_0^T & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & 2\mathbf{C}_u(\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}^T) & \mathbf{I}_{u(u+1)/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}_{(r-u)(r-u+1)/2} \end{pmatrix}.$$

Then, by direct computation using the relation $\mathbf{C}_r(\mathbf{\Gamma}\otimes\mathbf{\Gamma}) = \mathbf{C}_r(\mathbf{\Gamma}\otimes\mathbf{\Gamma})\mathbf{E}_u\mathbf{C}_u$ (Henderson and Searle (1979)), we have $\mathbf{H} = \mathbf{H}_1\mathbf{H}_2$. Because $\mathbf{H}_2$ has full row rank, we have span($\mathbf{H}$) = span($\mathbf{H}_1$). Furthermore, by straightforward multiplication we see that $\mathbf{H}_1^T\mathbf{J}\mathbf{H}_1$ is the desired block-diagonal matrix. Thus, we can now write

$$\mathbf{V}_0 = \mathbf{H}_1(\mathbf{H}_1^T\mathbf{J}\mathbf{H}_1)^{\dagger}\mathbf{H}_1^T = \sum_{j=1}^{4}\mathbf{H}_{1j}(\mathbf{H}_{1j}^T\mathbf{J}\mathbf{H}_{1j})^{\dagger}\mathbf{H}_{1j}^T. \tag{5.6}$$

5.2. Regression coefficients

Henceforth we write an asymptotic covariance matrix as avar$(\cdot)$; that is, if $\sqrt{n}(\mathbf{T}-\boldsymbol{\theta}) \overset{\mathcal{D}}{\rightarrow} N(\mathbf{0},\mathbf{A})$, then avar$(\sqrt{n}\,\mathbf{T}) = \mathbf{A}$. We now focus our attention on the asymptotic covariance matrix avar$[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})]$ of the estimate vec$(\hat{\boldsymbol{\beta}})$ of vec$(\boldsymbol{\beta})$ under the envelope model, since this will likely be of most use in practice. This matrix is the upper $pr \times pr$ block diagonal of $\mathbf{V}_0 = \text{avar}(\sqrt{n}\hat{\mathbf{h}})$. Since the first blocks of $\mathbf{H}_{13}$ and $\mathbf{H}_{14}$ are both $\mathbf{0}$, we have (see Supplement, Section D)

$$\begin{aligned}\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})] &= (\mathbf{I}_p\otimes\mathbf{\Gamma})(\mathbf{H}_{11}^T\mathbf{J}\mathbf{H}_{11})^{\dagger}(\mathbf{I}_p\otimes\mathbf{\Gamma}^T)+(\boldsymbol{\eta}^T\otimes\mathbf{\Gamma}_0)(\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12})^{\dagger}(\boldsymbol{\eta}\otimes\mathbf{\Gamma}_0^T)\\ &= \mathbf{\Sigma}_{\mathbf{X}}^{-1}\otimes\mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + (\boldsymbol{\eta}^T\otimes\mathbf{\Gamma}_0)(\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12})^{\dagger}(\boldsymbol{\eta}\otimes\mathbf{\Gamma}_0^T),\end{aligned} \tag{5.7}$$

where $\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12} = \boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T\otimes\mathbf{\Omega}_0^{-1}+\mathbf{\Omega}\otimes\mathbf{\Omega}_0^{-1}+\mathbf{\Omega}^{-1}\otimes\mathbf{\Omega}_0-2\mathbf{I}_u\otimes\mathbf{I}_{r-u}$. If $u=r$, then $\mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T = \mathbf{\Sigma}$, and the second term on the right hand side of (5.7) does not appear. The first term on the right hand side of (5.7) is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ when $\mathbf{\Gamma}$ is known, and the second term can be interpreted as the "cost" of estimating $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$. The total on the right does not exceed $\mathbf{\Sigma}_{\mathbf{X}}^{-1}\otimes\mathbf{\Sigma}$, which is the asymptotic variance of $\hat{\boldsymbol{\beta}}$ from the full model. A transparent decomposition of this asymptotic variance will be given in the next section.

Although we do not have a full proof, we expect that $\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12}$ will be of full rank, so that regular inverses can be used. This expectation is based on the following reasoning for two extreme cases. Suppose that $\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ have no eigenvalues in common. Then it can be shown that $(\mathbf{\Omega}\otimes\mathbf{\Omega}_0^{-1}+\mathbf{\Omega}^{-1}\otimes\mathbf{\Omega}_0-2\mathbf{I}_u\otimes\mathbf{I}_{r-u}) > \mathbf{0}$. Since $\boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T\otimes\mathbf{\Omega}_0^{-1}\geq\mathbf{0}$, it follows that $\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12} > \mathbf{0}$. On the other extreme, suppose that $\mathbf{\Omega} = \mathbf{I}_u$ and $\mathbf{\Omega}_0 = \mathbf{I}_{r-u}$, so that all their eigenvalues are identical. Then $\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12} = \boldsymbol{\eta}\mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T\otimes\mathbf{I}_{r-u}$, but in this case $\boldsymbol{\eta}$ must have full row rank equal to d, and again $\mathbf{H}_{12}^T\mathbf{J}\mathbf{H}_{12} > \mathbf{0}$.

5.3. Fitted values and predictions

From the above asymptotic results we can derive the asymptotic distribution of the fitted values, as well as the asymptotic prediction variance. In our context the fitted values at a particular $\mathbf{X}$ can be written as $\hat{\mathbf{Y}} = \hat{\boldsymbol{\beta}}\mathbf{X} = (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{vec}(\hat{\boldsymbol{\beta}})$. Hence the fitted value $\hat{\mathbf{Y}}$ has the following asymptotic distribution

$$\sqrt{n}(\hat{\mathbf{Y}} - \mathrm{E}\,(\hat{\mathbf{Y}})) \xrightarrow{\mathcal{L}} N(0, (\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})](\mathbf{X} \otimes \mathbf{I}_r)). \qquad (5.8)$$

The asymptotic mean squared error for prediction at $\mathbf{X}$ can be deduced similarly. Suppose that, at some value of $\mathbf{X}$, we observe a new $\mathbf{Y}$ – independently of the past observations $(\mathbf{X}_1, \mathbf{Y}_1), \ldots, (\mathbf{X}_n, \mathbf{Y}_n)$. Then

$$\begin{aligned}&\mathrm{E}\,[(\hat{\mathbf{Y}} - \mathbf{Y})(\hat{\mathbf{Y}} - \mathbf{Y})^T]\\ &\qquad = \mathrm{E}\,[(\hat{\mathbf{Y}} - \mathrm{E}\,(\hat{\mathbf{Y}}))(\hat{\mathbf{Y}} - \mathrm{E}\,(\hat{\mathbf{Y}}))^T] + \mathrm{E}\,[(\mathrm{E}\,(\hat{\mathbf{Y}}) - \mathbf{Y})(\mathrm{E}\,(\hat{\mathbf{Y}}) - \mathbf{Y})^T],\end{aligned}$$

where the cross-product terms vanish because $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ are independent. Combining this with expression (5.8), we see that the mean squared error of the prediction is approximated by

$$\mathrm{E}\,[(\hat{\mathbf{Y}} - \mathbf{Y})(\hat{\mathbf{Y}} - \mathbf{Y})^T] = n^{-1}(\mathbf{X}^T \otimes \mathbf{I}_r)\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})](\mathbf{X} \otimes \mathbf{I}_r) + \boldsymbol{\Sigma} + o(n^{-1}).$$

6. Interpretations

To gain further insight into the structure of our envelope model for multivariate linear regression, we now provide interpretations for the various quantities in the asymptotic variance of $\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})$ derived in the last section. The key to understanding this variance structure is the special structure of the joint Fisher information for $\boldsymbol{\phi} = (\boldsymbol{\phi}_1^T, \ldots, \boldsymbol{\phi}_4^T)^T$, as defined in (5.1). Let $\ell(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)$ denote the likelihood function for the $\boldsymbol{\phi}$'s. We adopt the notation

$$\mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\eta}} = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)}{\partial\boldsymbol{\phi}_1 \partial\boldsymbol{\phi}_1^T}\right], \quad \mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\Gamma}} = -\mathrm{E}\left[\frac{\partial^2 \ell(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)}{\partial\boldsymbol{\phi}_1 \partial\boldsymbol{\phi}_2^T}\right], \qquad (6.1)$$

and so on. Although it may be more technically correct to use notation such as $\mathbf{J}_{\boldsymbol{\phi}_1\boldsymbol{\phi}_2}$, we nevertheless use (6.1) to keep track of the original parameters. Furthermore, we use notations such as $\mathbf{J}_{(\boldsymbol{\eta},\boldsymbol{\Gamma})(\boldsymbol{\eta},\boldsymbol{\Gamma})}$ to denote the joint information for parameter sub-vectors such as $(\boldsymbol{\phi}_1^T, \boldsymbol{\phi}_2^T)^T$.

From the discussion in Section 5 it can be deduced that the Fisher information for $(\boldsymbol{\phi}_1, \ldots, \boldsymbol{\phi}_4)$, $\mathbf{H}^T\mathbf{J}\mathbf{H}$, has the form

$$\begin{pmatrix} \mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\Gamma}} & \mathbf{0} & \mathbf{0} \\ \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\eta}} & \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Gamma}} & \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Omega}} & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Gamma}} & \mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Omega}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{\boldsymbol{\Omega}_0\boldsymbol{\Omega}_0} \end{pmatrix}, \qquad (6.2)$$

with specific expressions for the non-zero blocks given in the Supplement, Section E. What is special about this form is that, if we cross out the second row and second column, the remaining matrix is block-diagonal with three diagonal blocks; $\mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\eta}}$, $\mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Omega}}$, and $\mathbf{J}_{\boldsymbol{\Omega}_0\boldsymbol{\Omega}_0}$. Similarly, if we cross out the first row and first column, the remaining matrix is block-diagonal with two diagonal blocks; $\mathbf{J}_{(\boldsymbol{\Gamma},\boldsymbol{\Omega})(\boldsymbol{\Gamma},\boldsymbol{\Omega})}$ and $\mathbf{J}_{\boldsymbol{\Omega}_0}$. This implies two important facts.

1. If $\boldsymbol{\Gamma}$ is known, then the asymptotic variance of the MLE of $\boldsymbol{\eta}$, say $\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$, is simply $\mathbf{J}_{\boldsymbol{\eta}\boldsymbol{\eta}}^{-1}$. The other two parameters, $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$, have no plugging-in effect.

2. If $\boldsymbol{\eta}$ is known, then the asymptotic variance of the MLE of $\boldsymbol{\Gamma}$, say $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}}$, is
$$\left(\mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Gamma}} - \mathbf{J}_{\boldsymbol{\Gamma}\boldsymbol{\Omega}}\mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Omega}}^{-1}\mathbf{J}_{\boldsymbol{\Omega}\boldsymbol{\Gamma}}\right)^{-1}, \tag{6.3}$$
that is, $\boldsymbol{\Omega}_0$ has no plugging-in effect on $\hat{\boldsymbol{\Gamma}}_0$.

Interestingly, the asymptotic variance of $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})$ can be written as a simple and transparent linear combination of avar$[\sqrt{n}\text{vec}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})]$ and avar$[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})]$. Explicit forms for these asymptotic variances can be computed from (6.3) and the formulas for the information blocks given in the Supplement, as

$$\begin{aligned}\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})] =& \boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Omega},\\ \text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})] =& [\boldsymbol{\eta}\boldsymbol{\Sigma}_{\mathbf{X}}\boldsymbol{\eta}^T \otimes \boldsymbol{\Sigma}^{-1} + (\boldsymbol{\Omega} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T) - 2(\mathbf{I}_u \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)\\ &+ (\boldsymbol{\Omega}^{-1} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)]^{-1}.\end{aligned} \tag{6.4}$$

The first equality can be obtained straightforwardly from $\mathbf{H}^T\mathbf{J}\mathbf{H}$, but the derivation of the second is quite involved – a detailed proof of (6.4) can be found in the Supplement, Section E. Also in the Supplement is a proof of how the theorem below follows from these equalities.

Theorem 6.1. *The asymptotic variance of* $\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})$ *can be written as*
$$\begin{aligned}\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}})] =& (\mathbf{I}_p \otimes \boldsymbol{\Gamma})\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}})](\mathbf{I}_p \otimes \boldsymbol{\Gamma}^T)\\ &+ (\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})](\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T).\end{aligned} \tag{6.5}$$

This representation can be made even more transparent if we recognize the following: if $\boldsymbol{\Gamma}$ is known, then $\boldsymbol{\Gamma}\hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ is just $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}}$, the maximum likelihood estimate of $\boldsymbol{\beta}$; for the second term in (6.5) we have

$$\begin{aligned}(\boldsymbol{\eta}^T \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)\text{avar}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})](\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T) &= \text{avar}[\sqrt{n}(\boldsymbol{\eta} \otimes \boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T)\text{vec}(\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}})]\\ &= \text{avar}[\sqrt{n}\text{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}}\boldsymbol{\eta})].\end{aligned}$$

However, $\hat{\boldsymbol{\Gamma}}_{\boldsymbol{\eta}}\boldsymbol{\eta}$ is simply the maximum likelihood estimator of $\boldsymbol{\beta}$ when $\boldsymbol{\eta}$ is known.

Corollary 6.1. *The asymptotic variance of* $\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})$ *satisfies* $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})] = \mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})] + \mathrm{avar}[\sqrt{n}\mathrm{vec}(\mathbf{Q}_{\boldsymbol{\Gamma}}\hat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})]$.

Intuitively, the asymptotic variance of $\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})$ comprises those of $\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\Gamma}})$ and $\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}}_{\boldsymbol{\eta}})$; the role played by $\mathbf{Q}_{\boldsymbol{\Gamma}}$ is to orthogonalize these random vectors so that their contributions to the net asymptotic variance are additive.

Finally, to provide some insight on situations in which our estimator can be particularly effective, we compare $\mathrm{avar}[\sqrt{n}\mathrm{vec}(\hat{\boldsymbol{\beta}})]$ and $\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}$, the asymptotic variance of the usual MLE, in a relatively simple setting. Let $p = 1$, $\boldsymbol{\Omega} = \sigma^2\mathbf{I}_u$, and $\boldsymbol{\Omega}_0 = \sigma_0^2\mathbf{I}_{r-u}$. In this case it can be shown that

$$\{\mathrm{avar}(\sqrt{n}\hat{\boldsymbol{\beta}})\}^{-1/2}\{\boldsymbol{\Sigma}/\sigma_X^2\}\{\mathrm{avar}(\sqrt{n}\hat{\boldsymbol{\beta}})\}^{-1/2} = \mathbf{I}_r + \frac{(\sigma_0^2 - \sigma^2)^2}{\sigma_X^2\sigma^2\|\boldsymbol{\beta}\|^2}\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T \tag{6.6}$$

where we have used σ_X^2 in place of $\boldsymbol{\Sigma}_{\mathbf{X}}$ to emphasize that $p = 1$. This result indicates that the difference between our estimator and the standard MLE decreases when the signal ($\|\boldsymbol{\beta}\|$ or σ_X^2) increases, and increases when the variability (σ^2 or σ_0^2) increases. Equation (6.6) says also that the two approaches are equally efficient asymptotically when $\sigma^2 = \sigma_0^2$, a fact that is supported by the simulation results in Section 7. In full generality, (6.6) suggests that our estimator will provide the most gains in efficiency when the envelope $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be constructed from eigenspaces of $\boldsymbol{\Sigma}$ with relatively small eigenvalues (cf. Proposition 2.3). In particular, the size of u seems less important than the relative sizes of these eigenvalues, provided $u < r$.

7. Comparing Two Normal Means: Simulation and Data Analysis Results

We use the classic setting of comparing the means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ of two multivariate normal populations to illustrate the potential benefits of envelope models, and to verify our asymptotic calculations. In terms of model (1.1), the two-means comparison can be represented by taking $\boldsymbol{\alpha} = \boldsymbol{\mu}_1$, $\boldsymbol{\beta} = \boldsymbol{\mu}_2 - \boldsymbol{\mu}_1$, and $\mathbf{X} \in \{0, 1\}$. Since $p = 1$, we again use σ_X^2 in place of $\boldsymbol{\Sigma}_{\mathbf{X}}$ when describing various results.

In our simulations we tracked both small sample bias and variability. Since no appreciable bias was detected, Section 7.1 reports only variability comparisons, summarized using versions of the generalized standard deviation ratio

$$T = \{\mathrm{tr}(\boldsymbol{\Delta}_{\mathrm{em}}^{-1/2}\boldsymbol{\Delta}_{\mathrm{fm}}\boldsymbol{\Delta}_{\mathrm{em}}^{-1/2})/r\}^{1/2}, \tag{7.1}$$

where $\boldsymbol{\Delta}_{\mathrm{fm}}$ represents the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$, the estimator of $\boldsymbol{\beta}$ from the full model (1.1), and $\boldsymbol{\Delta}_{\mathrm{em}}$ represents the covariance matrix of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$, the estimator of $\boldsymbol{\beta}$ from the envelope model (3.2) (for consistency of notation we use $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ instead of $\widehat{\boldsymbol{\beta}}$ to denote the envelope estimator in this section). T^2 can be interpreted as the average variance $\mathrm{E}\,(\boldsymbol{\ell}^T\boldsymbol{\Delta}_{\mathrm{fm}}\boldsymbol{\ell})$, where the average is computed over all $\boldsymbol{\ell} \in \mathbb{R}^r$ subject to the constraint that $\boldsymbol{\ell}^T\boldsymbol{\Delta}_{\mathrm{em}}\boldsymbol{\ell} = 1$. Values of $T > 1$ indicate that the envelope model (3.2) produces smaller standard deviations on average than the full model.

7.1. Simulation results

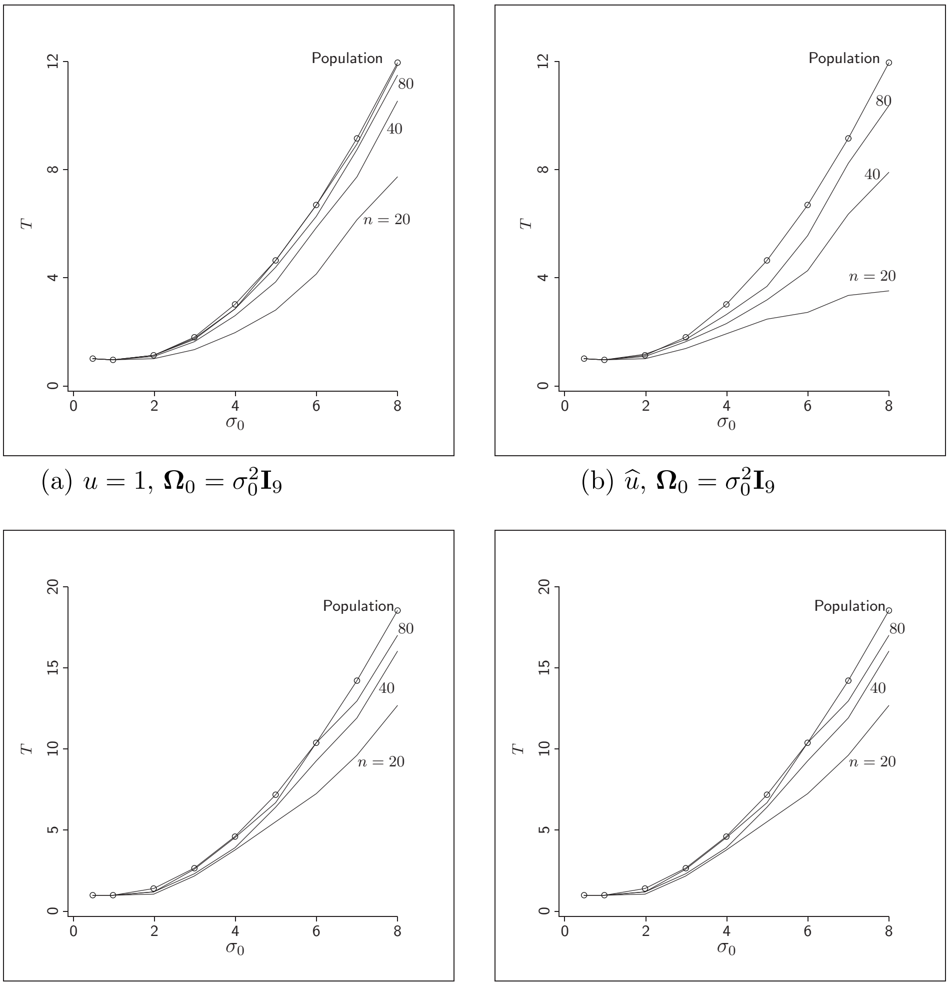
All results reported here were based on 200 replications from simulation models with $n/2$ observations per population, $r = 10$, $\boldsymbol{\beta}^T = (\sqrt{10}, \ldots, \sqrt{10})$, $u = 1$, and variance $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. Two versions of T were used. In the first, T_{pop}, we set $\boldsymbol{\Delta}_{\mathrm{em}} = \mathrm{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_{\mathrm{em}})$ (see (5.7)) and $\boldsymbol{\Delta}_{\mathrm{fm}} = \boldsymbol{\Sigma}/\sigma_X^2$, with all parameters at the values used in the simulations. It follows immediately from (6.6) that

$$T_{\mathrm{pop}}^2 = 1 + (1 - r^{-1})\frac{(\sigma^2 - \sigma_0^2)^2}{\|\boldsymbol{\beta}\|^2\sigma_X^2\sigma^2}.$$

In the second version, T_n, we set $\boldsymbol{\Delta}_{\mathrm{em}}$ and $\boldsymbol{\Delta}_{\mathrm{fm}}$ to be the sample covariance matrices of the 200 replications of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$. If our asymptotic calculations are correct, then for a sufficiently large n we should have $T_{\mathrm{pop}} \approx T_n$.

The simulated data underlying Figure 7.1a were drawn using $\sigma^2 = 1$ and $\boldsymbol{\Omega}_0 = \sigma_0^2\mathbf{I}_9$. We used the true value $u = 1$ when forming the estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ based on (3.2). The upper curve, identified by the open circles, is a plot of T_{pop} for various values of σ_0. The other curves correspond to T_n for four samples sizes. As n increases T_n evidently approaches T_{pop} from below, with T_{80} being quite close to T_{pop}. The unlabeled curve that lies between T_{pop} and T_{80} was obtained with $n = 160$. The results in Figure 7.1a show that estimates from the envelope model can be much more efficient than the usual full-model estimates. They also support our previous conclusion that there is little difference between the methods when $\sigma \approx \sigma_0$.

Figure 7.1b was constructed as Figure 7.1a except that, for the T_n curves, u was estimated as follows for each of the 200 simulated data sets. The hypothesis $u = u_0$ can be tested by using the likelihood ratio statistic $\Lambda(u_0) = 2(\widehat{L}_{\mathrm{fm}} - \widehat{L}^{(u_0)})$, where $\widehat{L}_{\mathrm{fm}}$ denotes the maximum value of the log likelihood for the full model, and $\widehat{L}^{(u_0)}$ the maximum value of the log likelihood for (3.2). Following standard likelihood theory, under the null hypothesis $\Lambda(u_0)$ is distributed asymptotically as a chi-squared random variable with $p(r - u_0)$ degrees of freedom. We employed the statistic $\Lambda(u_0)$ in a sequential scheme to choose u. Using a common test level of 0.01 and starting with $u_0 = 0$, we chose the estimate $\widehat{u}$ of u as the first hypothesized value that was not rejected. The results in Figure 7.1b show as expected that estimating u increases the variability of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$, but substantial gains are still possible for modest sample sizes. The drop for $n = 20$ is due mainly to



(a) $u = 1$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_9$

(b) $\widehat{u}$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_9$

(c) $u = 1$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{A}^T \mathbf{A}/9$

(d) $\widehat{u}$, $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{A}^T \mathbf{A}/9$

Figure 7.1. Simulation results for comparing the means of two multivariate normal populations.

the tendency of the likelihood ratio test to reject too frequently for small samples. The bounding dimension u could also be selected using an information criterion like AIC or BIC. Our intent here is to demonstrate only that reasonable inference on u is possible, without recommending a particular method.

Figures 7.1c and 7.1d were constructed as were Figures 7.1a and 7.1b, except that $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{I}_9$ was replaced by $\mathbf{\Omega}_0 = \sigma_0^2 \mathbf{A}^T \mathbf{A}/9$, where $\mathbf{A} \in \mathbb{R}^{9 \times 9}$ was generated once as a matrix of standard normal variates. The range of the y-axis in Figures 7.1c and 7.1d is nearly twice that for Figures 7.1a and 7.1b, suggesting that correlation improves the performance of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ relative to $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$.

7.2. Data analysis

We applied the proposed methodology to a number of data sets from the literature and found an advantage in most of them, suggesting that envelope models may have wide applicability. We present two brief illustrations in this section, one with a real but modest gain for envelope models, and one with a dramatic gain.

In a sample of 172 New Zealand mussels, 61 were found to have pea crabs living within the shell and 111 were free of pea crabs. We compared the means of these two populations on $r = 6$ response variables: the logarithms of shell height, shell width, shell length, shell mass, muscle mass, and viscera mass. Bartlett's test statistic for equality of covariance matrices has the value 27.8 on 21 degrees of freedom, so the assumption of equal covariance matrices seems reasonable. The p-values for the likelihood ratio tests of $u = 1$ and $u = 2$ were 0.024 and 0.18, suggesting that either of these values might be appropriate. Letting $\widehat{T}$ denote the estimate of T by using the plug-in method, we found that $\widehat{T} = 4.7$ for $u = 1$ and $\widehat{T} = 2.9$ for $u = 2$. In either case, it seems that the estimate of the mean difference from model (3.2) is notably less variable than the full-model estimate. Even with $u = 2$ these results indicate that it would take a sample about $2.9^2 = 8.41$ times as large for the efficiency of $\widehat{\boldsymbol{\beta}}_{\mathrm{fm}}$ to equal that of $\widehat{\boldsymbol{\beta}}_{\mathrm{em}}$ with the present sample size. The $\widehat{T}$ summary reflects the ratio of standard errors over all linear combinations of the coefficients. The standard error ratios are more modest when considering only individual coefficients, the individual standard errors for the full-model estimates ranging between 1.18 and 1.05 times the respective standard errors for the envelope estimates. For the largest of these, the envelope estimates achieve a reduction equivalent to full-model estimates with roughly a 40 percent increase in sample size. We expect that this would be judged worthwhile in most analyses.

The second data set is the infrared reflectance example described in the Introduction. We chose this data set because the marginal response correlations are high, ranging between 0.9118 and 0.9991. This is the kind of situation in which the proposed methodology might give massive gains over a full-model analysis. The likelihood ratio test statistic for the hypothesis $u = 1$ has the value 1.09 on 5 degrees of freedom for a p-value of 0.95. With $u = 1$, $\widehat{T} = 219.2$. The standard deviation ratios for the individual mean differences were described in the Introduction.

To confirm the results for the infrared reflectance data we constructed a simulation model using all of the estimates from the original data as the population values. The population standard deviation ratio for this simulation scenario is $T_{\rm pop} = 219.2$, which is the same as the plug-in estimate from the original data. We then constructed estimates based on 24 low protein observations and 26 high protein observations from the simulation model, repeating the process 200 times. This gave $T_n = 221.6$ and average plug-in estimate $\widehat{T} = 240.4$, which seems to support the results of the original analysis. To see if the high response correlations might introduce a notable small sample bias in $\widehat{\boldsymbol{\beta}}_{\rm em}$ we computed $({\rm ave}(\widehat{\boldsymbol{\beta}}_{\rm em}) - \boldsymbol{\beta})/\boldsymbol{\beta}$ element-wise, where ${\rm ave}(\widehat{\boldsymbol{\beta}}_{\rm em})$ denotes the replication average of $\widehat{\boldsymbol{\beta}}_{\rm em}$. These six ratios ranged between -0.018 and 0.011. The same calculations using the 200 replications of $\widehat{\boldsymbol{\beta}}_{\rm fm}$ produced six ratios ranging between -0.122 and 0.175.

The fit of model (3.2) to the original reflectance data gave $\widehat{\boldsymbol{\Sigma}} = \widehat{\boldsymbol{\Sigma}}_1 + \widehat{\boldsymbol{\Sigma}}_2$, where $\widehat{\boldsymbol{\Sigma}}_1$ had rank 1 with non-zero eigenvalue 7.88, and $\widehat{\boldsymbol{\Sigma}}_2$ had rank 5 with eigenvalues $6,516.61$, 208.29, 20.08, 0.42, and 0.27. Evidently, the proposed method offers truly substantial gains in this example because the collinearity in $\boldsymbol{\Sigma}$ is quite large, and because $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is inferred to lie in an eigenspace of $\boldsymbol{\Sigma}$ with a relatively small eigenvalue.

8. Extensions and Relationships with Other Theory and Methods

In the previous sections we focused on one way in which the notion of enveloping can be employed; namely, creating a parsimonious, alternative parameterization for the multivariate linear model. However, envelopes can be used in other ways and in other contexts to allow more control over parameterizations, and to develop methodology affording substantial gains in efficiency. We expect enveloping to have considerable potential in multivariate analysis: whenever we are dealing with a random vector $\mathbf{U}$ and an associated covariance matrix $\boldsymbol{\Lambda}$, we can consider a parsimonious parameterization of the latter in reference to the former. Mathematically, the essence of enveloping is to find the smallest reducing subspace of $\boldsymbol{\Lambda}$ to which $\mathbf{U}$ belongs almost surely. In this section, we offer conjectures about a number of multivariate analysis contexts that share this form – the discussion is largely at the population level.

8.1. Reduced rank envelope models

Maximum likelihood estimation under model (1.1) does not require the coefficient matrix $\boldsymbol{\beta}$ to be of full rank $\min(r, p)$. Similarly, the envelope models introduced in the previous sections permit the rank of $\boldsymbol{\beta}$ to be less than $\min(r, p)$. In some regressions it may be useful to explicitly fit an envelope model with a specified rank d for $\boldsymbol{\beta}$. This, in effect, combines envelope models with models for multivariate reduced rank regression (Anderson (1951); Izenman (1975); Davies and Tso (1982); Bura and Cook (2003)).

Recall from (3.2) that the mean function for the envelope model is $\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{X}$, where $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$ is a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. If we restrict $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta}$ to have rank $d < \min(r, p)$, then $\boldsymbol{\eta} \in \mathbb{R}^{u \times p}$ must have rank d and thus can be factored as $\boldsymbol{\eta} = \boldsymbol{\gamma}\boldsymbol{\phi}$, where $\boldsymbol{\gamma} \in \mathbb{R}^{u \times d}$ is a semi-orthogonal matrix and $\boldsymbol{\phi} \in \mathbb{R}^{d \times p}$ is unconstrained. This gives a reduced rank version of envelope model (3.2):

$$\begin{aligned} \mathbf{Y} &= \boldsymbol{\alpha} + \boldsymbol{\Gamma}\boldsymbol{\gamma}\boldsymbol{\phi}\mathbf{X} + \boldsymbol{\varepsilon} \\ \boldsymbol{\Sigma} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T. \end{aligned} \tag{8.1}$$

As before, $\boldsymbol{\Gamma}$ is not identified but $\mathrm{span}(\boldsymbol{\Gamma}) = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is identified and estimable. Similarly, $\mathrm{span}(\boldsymbol{\gamma}) \in \mathbb{G}^{u \times d}$ and $\boldsymbol{\phi}$ are identified and estimable. Like the envelope version of model (1.1), this model has the potential for substantial gains in efficiency relative to the usual multivariate reduced rank model. Maximum likelihood and other methods of estimation for this model are currently under study.

It may be clear that we do not view reduced rank and envelope models as direct competitors, since combining them leads to (8.1) which is a more versatile model than either one alone, and allows for more control over dimensionality. Similarly, many other methods for reducing dimensionality, like factor analysis, variable selection, and coefficient penalization (Yuan et al. (2007)), could be extended for use with envelope models.

8.2. Discriminant analysis

Consider classifying a new observation $\mathbf{y}$ on a feature vector $\mathbf{Y} \in \mathbb{R}^r$ into one of two normal populations C_1 and C_2, with means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ and common covariance matrix $\boldsymbol{\Sigma}$. Assuming equal prior probabilities, the optimal population rule, which is the same as Fisher's linear discriminant (Seber (1984, p. 331)), is to classify $\mathbf{y}$ as arising from C_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}\mathbf{y} > \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Letting $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$, $u \le r$, denote a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathrm{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$, it follows from Corollary 2.1 that $\boldsymbol{\Sigma}^{-1}$ is of the form $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^T$. The optimal population rule expressed in terms of $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathrm{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2))$ is to classify into C_1 if

$$(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{y} > \frac{1}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2).$$

Estimates of u, $\boldsymbol{\Gamma}$ and $\boldsymbol{\Omega}$ can be found using the methods discussed in previous sections, specifying $\mathbf{Y}$ as the response vector. When $u \ll p$ or the eigenvalues of

$\mathbf{\Omega}$ are substantially larger than those of $\mathbf{\Omega}_0$, we expect misclassification rates for this rule to be significantly lower than those for the standard rule. In cases where $u = 1$, $\mathcal{E}_{\mathbf{\Sigma}}(\text{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)) = \text{span}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ and, assuming that $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{\Gamma} > 0$, the rule simplifies to $(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T\mathbf{y} > (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2)/2$. Extension to multiple populations with common covariance matrix seems straightforward conceptually.

Principal components have long been considered for dimension reduction prior to discriminant analysis. The first two methods discussed by Jolliffe (2002, Sec. 9) reduce $\mathbf{Y}$ by using the first few principal components from either the intra-population covariance matrix or the marginal covariance of $\mathbf{Y}$, computed without regard to population membership. Neither method is entirely satisfactory because there is no guarantee that the first few principal components will be the “best” for discrimination. The envelope approach proposed here has the potential to achieve what has long been attempted through principal component methodology.

8.3. Principal components

There are numerous ways to motivate the use of principal components for the reduction of a multivariate vector $\mathbf{Y} \in \mathbb{R}^r$ (Jolliffe (2002)). In this section we describe how an envelope construction might aid us in understanding a foundation for principal components based on latent variables (Tipping and Bishop (1999)).

Again consider model (1.1), only now with $\mathbf{X} \in \mathbb{R}^p$ as an unobserved vector of latent variables, standardized to have mean 0 and variance $\mathbf{I}_p$, and with $\boldsymbol{\beta}$ assumed to have rank $p < r$. The latent vector represents extrinsic variation in $\mathbf{Y}$, while the error $\boldsymbol{\varepsilon}$ represents intrinsic variation. The goal is to reduce the dimension of $\mathbf{Y}$ accounting for its extrinsic variation. Under this model it can be shown that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\mathbf{Y}$ (Cook (2007)), and thus $\mathbf{R} = \boldsymbol{\beta}^T\mathbf{\Sigma}^{-1}\mathbf{Y}$ is the reduction we would like to estimate. Any full rank linear transformation $\mathbf{A}$ of $\mathbf{R}$ results in an equivalent reduction; $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{R}$ if and only if $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{AR}$, so it is sufficient to estimate $\mathcal{S} = \text{span}(\mathbf{\Sigma}^{-1}\boldsymbol{\beta})$. Additionally $\mathcal{S}$ is minimal, if $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}^T\mathbf{Y}$ then $\mathcal{S} \subseteq \text{span}(\mathbf{B})$. Note that here we focus on estimating $\mathcal{S}$, though additional considerations may be necessary to translate knowledge about $\mathcal{S}$ into actions, depending on the application context.

Since $\mathbf{X}$ is not observed, only the marginal distribution of $\mathbf{Y}$ is available for the purpose of estimating $\mathcal{S}$. Following Tipping and Bishop (1999) we assume that $\mathbf{X}$ is normally distributed, and thus $\mathbf{Y}$ is normal with mean $\boldsymbol{\alpha}$ and variance $\mathbf{\Sigma}_{\mathbf{Y}} = \mathbf{\Sigma} + \boldsymbol{\beta}\boldsymbol{\beta}^T$. The maximum likelihood estimator of $\boldsymbol{\alpha}$ is just the sample mean of $\mathbf{Y}$, but $\mathbf{\Sigma}$ and $\boldsymbol{\beta}$ are confounded and cannot be separated without additional structure. Tipping and Bishop (1999) assumed isotropic errors, i.e., $\mathbf{\Sigma} = \sigma^2\mathbf{I}_r$, and it follows from their results that the maximum likelihood estimator of $\mathcal{S} =$

span($\boldsymbol{\beta}$) is the span of the first p eigenvectors of $\widehat{\boldsymbol{\Sigma}}_{\mathbf{Y}}$, the sample version of $\boldsymbol{\Sigma}_{\mathbf{Y}}$. Consequently, $\mathbf{R}$ is estimated by the first p principal components of the marginal variance of $\mathbf{Y}$ when the errors are isotropic.

The assumption of isotropic errors is limiting relative to the range of applications where it may be desirable to reduce multivariate observations. In the envelope parameterization of model (3.2), $\mathcal{S} = \mathrm{span}(\boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\eta})$ and $\mathbf{Y}$ is normally distributed with mean $\boldsymbol{\alpha}$ and $\boldsymbol{\Sigma}_{\mathbf{Y}} = \boldsymbol{\Gamma}(\boldsymbol{\Omega} + \boldsymbol{\eta}\boldsymbol{\eta}^T)\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T$. The coefficients $\boldsymbol{\eta}$ are not identified since they are confounded with $\boldsymbol{\Omega}$, so it is still not possible to estimate $\mathcal{S}$. However, $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\eta}^T\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^T\mathbf{Y}$ implies that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\boldsymbol{\Gamma}^T\mathbf{Y}$, so $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ provides an upper bound on the space of interest, $\mathcal{S} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. With isotropic errors, we have $\mathcal{S} = \mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. We conjecture that, with a sufficiently large intrinsic signal $\boldsymbol{\eta}$, $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ can be estimated from the marginal of $\mathbf{Y}$. The envelope model (3.2) with $\mathbf{X}$ as a latent vector would then allow estimation of the upper bound $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, which may be helpful in some applications, and could provide insights onto the usefulness of principal components under a general error structure.

8.4. Envelopes in the predictor space

Recall from the discussion in Section 3 that the envelope model expressed by (3.2) has the greatest potential for improvement in regressions with many responses (r) and relatively few predictors (p). The novel parametrization we propose has nothing to offer when $r \leq p$ and $d = \dim(\mathcal{B}) = r$. This is the case, for example, in univariate linear regression ($r = 1$). Nevertheless, it may still be possible to achieve efficiency gains by using an envelope construction *in the predictor space.*

Assuming that $\mathbf{X}$ is random, the population coefficient matrix $\boldsymbol{\beta}$ can be represented as $\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\,\mathrm{Cov}\,(\mathbf{X}, \mathbf{Y})$, where $\boldsymbol{\Sigma}_{\mathbf{X}} = \mathrm{Var}\,(\mathbf{X})$ is the marginal variance of $\mathbf{X}$. Let $\mathcal{C} = \mathrm{span}(\mathrm{Cov}\,(\mathbf{X}, \mathbf{Y})) \subseteq \mathbb{R}^p$ and let $\boldsymbol{\chi}$ be a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_X}(\mathcal{C})$, the $\boldsymbol{\Sigma}_{\mathbf{X}}$-envelope of $\mathcal{C}$. Then the coefficient matrix can be written as $\boldsymbol{\beta}^T = \mathbf{P}_{\boldsymbol{\chi}(\boldsymbol{\Sigma}_X)}\boldsymbol{\beta}^T$. This suggests that we estimate $\boldsymbol{\beta}^T$ by projecting the usual maximum likelihood estimate onto an estimate of $\mathcal{E}_{\boldsymbol{\Sigma}_X}(\mathcal{C})$, using the sample version of $\boldsymbol{\Sigma}_{\mathbf{X}}$ for the inner product. We would again expect notable efficiency gains if $\dim(\mathcal{E}_{\boldsymbol{\Sigma}_X}(\mathcal{C})) < p$ and we can find a good way to estimate $\mathcal{E}_{\boldsymbol{\Sigma}_X}(\mathcal{C})$. One method of estimating $\mathcal{E}_{\boldsymbol{\Sigma}_X}(\mathcal{C})$ that permits $n < p$ is described in Section 8.7.

8.5. Simultaneous envelopes

There is also the possibility of combining predictor space envelopes $\mathcal{E}_{\boldsymbol{\Sigma}_X}(\mathcal{C}) \subseteq \mathbb{R}^p$ with response space envelopes $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B}) \subseteq \mathbb{R}^r$ in a single multivariate regression. The predictor space envelopes of Section 8.4 rely on the identity $\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\,\mathrm{Cov}\,(\mathbf{X}, \mathbf{Y})$, which connects the coefficient matrix with population moment matrices. The corresponding expression for the envelope model (3.2) follows from (1.2); $\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\,\mathrm{Cov}\,(\mathbf{X},\mathbf{Y})\mathbf{P}_{\boldsymbol{\Sigma}_1}$, where $\mathbf{P}_{\boldsymbol{\Sigma}_1}$ is still the projection onto $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$. It now follows from Section 8.4 that

$$\boldsymbol{\beta}^T = \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}\,\mathrm{Cov}\,(\mathbf{X},\mathbf{Y})\mathbf{P}_{\boldsymbol{\Sigma}_1} = \boldsymbol{\beta}^T\mathbf{P}_{\boldsymbol{\Sigma}_1} = \mathbf{P}_{\boldsymbol{\mathcal{X}}(\boldsymbol{\Sigma}_X)}\boldsymbol{\beta}^T\mathbf{P}_{\boldsymbol{\Sigma}_1}.$$

This may serve as a conceptual starting point for the development of methods based on enveloping in both the predictor and response spaces.

8.6. Sufficient dimension reduction

There are various methods for reducing the dimension of a random predictor $\mathbf{X} \in \mathbb{R}^p$ in a regression with univariate response $Y \in \mathbb{R}^1$. Among them, *sufficient dimension reduction* methods estimate the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ (Cook (1994, 1998)), defined as the intersection of all subspaces $\mathcal{S}$ with the property that $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{P}_{\mathcal{S}}\mathbf{X}$. Since the conditional distributions of $Y|\mathbf{X}$ and $Y|\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ are identical, we can substitute $\mathbf{P}_{\mathcal{S}_{Y|\mathbf{X}}}\mathbf{X}$ for $\mathbf{X}$ without loss of information on the regression.

Cook (2007) proposed that estimation of $\mathcal{S}_{Y|\mathbf{X}}$ be based on modeling the conditional distribution of $\mathbf{X}|Y$. Suppose that

$$\mathbf{X} = \boldsymbol{\mu} + \boldsymbol{\beta}\mathbf{f}_y + \boldsymbol{\varepsilon}, \tag{8.2}$$

where $\boldsymbol{\mu} \in \mathbb{R}^p$, $\mathbf{f}_y \in \mathbb{R}^r$ is a known user-specified function of y, and $\boldsymbol{\varepsilon}$ is normally distributed with mean 0 and covariance matrix $\boldsymbol{\Sigma}_{\mathbf{X}|Y}$. This is a multivariate linear model like (1.1), with the predictor vector $\mathbf{X}$ taking on the role of the response, and $\mathbf{f}_y$ taking the role of the predictor. However, in pursuing our sufficient dimension reduction, we have no particular interest in the coefficient matrix $\boldsymbol{\beta} \in \mathbb{R}^{p\times r}$. Instead, interest lies in the central subspace $\mathcal{S}_{Y|\mathbf{X}} = \boldsymbol{\Sigma}_{\mathbf{X}|Y}^{-1}\mathcal{B}$ (Cook (2007)), where still $\mathcal{B} = \mathrm{span}(\boldsymbol{\beta})$. We can now use $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$ to parameterize (8.2), leading to an envelope model with the same form as (3.2), or perhaps a reduced rank envelope model like (8.1). Because of the importance of $\mathcal{S}_{Y|\mathbf{X}}$, we might instead consider parameterizing in terms of the $\boldsymbol{\Sigma}_{\mathbf{X}|Y}$-envelope of $\mathcal{S}_{Y|\mathbf{X}}$, $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}|Y}}(\mathcal{S}_{Y|\mathbf{X}})$. In view of Proposition 3.1, however, these and several other envelopes are equal and thus lead to the same parameterization.

These considerations allow us a better understanding of the PFC model proposed by Cook (2007, eq. 13) without reference to envelopes:

$$\begin{aligned}\mathbf{X} &= \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\eta}\mathbf{f}_y + \boldsymbol{\varepsilon}\\ \boldsymbol{\Sigma}_{\mathbf{X}|Y} &= \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T.\end{aligned}$$

If we take $\boldsymbol{\Gamma}$ to be a semi-orthogonal basis matrix for $\mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$, then this is the envelope version of (8.2). Since $\mathcal{S}_{Y|\mathbf{X}} \subseteq \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$, the model only allows estimation of an upper bound on $\mathcal{S}_{Y|\mathbf{X}}$. To estimate the central subspace itself it is necessary to use a reduced rank envelope model, except in special cases where $\mathcal{S}_{Y|\mathbf{X}} = \mathcal{E}_{\mathbf{\Sigma}_{\mathbf{X}|Y}}(\mathcal{B})$.

8.7. Seeded reductions when $n < p$

In addition to the model-based approaches proposed by Cook (2007), there are numerous moment-based methods for estimating the central subspace $\mathcal{S}_{Y|\mathbf{X}}$ without using models. Under various conditions, many of these approaches exploit population identities of the form $\mathcal{S}(\boldsymbol{\nu}) = \mathbf{\Sigma}_{\mathbf{X}}\mathcal{S}_{Y|\mathbf{X}}$, where $\mathbf{\Sigma}_{\mathbf{X}} = \mathrm{Var}\,(\mathbf{X})$ as defined previously, and $\boldsymbol{\nu}$ is a method-specific *seed* matrix (Cook, Li and Chiaromonte (2007)) that can be estimated from the sample moments of $(\mathbf{Y}, \mathbf{X})$ without inverting the sample version of $\mathbf{\Sigma}_{\mathbf{X}}$. For example, the least square seed (which corresponds to the multivariate linear model; see Section 8.4) is $\boldsymbol{\nu} = \mathrm{Cov}\,(\mathbf{X}, \mathbf{Y})$, and the seed for sliced inverse regression (Li (1991)) is $\boldsymbol{\nu} = \mathrm{Var}\,(\mathrm{E}\,(\mathbf{X}|\mathbf{Y}))$. Of course, when n is sufficiently large, $\mathcal{S}_{Y|\mathbf{X}}$ can simply be estimated from the spectral structure of the sample version of $\mathbf{\Sigma}_{\mathbf{X}}^{-1}\boldsymbol{\nu}$.

By using the $\mathbf{\Sigma}_{\mathbf{X}}$-envelope of $\mathcal{S}_{Y|\mathbf{X}}$, Cook, Li and Chiaromonte (2007) developed a method of estimating $\mathcal{S}_{Y|\mathbf{X}}$ that does not require $n > p$. Their method is based on estimating $\mathcal{E}_{\mathbf{\Sigma}_X}(\mathcal{S}_{Y|\mathbf{X}})$ as the span of the sample version of $\mathbf{R}_u = (\boldsymbol{\nu}, \mathbf{\Sigma}_{\mathbf{X}}\boldsymbol{\nu}, \ldots, \mathbf{\Sigma}_{\mathbf{X}}^{u-1}\boldsymbol{\nu})$, where u is also estimated. A basis for $\mathcal{S}_{Y|\mathbf{X}}$ is then estimated by projecting an estimate of $\boldsymbol{\nu}$ onto the space spanned by the estimate of $\mathbf{R}_u$, using the sample version of $\mathbf{\Sigma}_{\mathbf{X}}$ to define the inner product. The method is equivalent to partial least squares (Helland (1990)) for univariate linear regression when the seed is $\boldsymbol{\nu} = \mathrm{Cov}\,(\mathbf{X}, Y) \in \mathbb{R}^p$.

We conjecture that envelopes can be used in a variety of settings to develop estimation methods that allow $p > n$, and that generally have the potential to yield substantial gains in efficiency relative to standard methods, even when $p \ll n$. The particular method proposed by Cook, Li and Chiaromonte (2007) is a first step along these lines, but we expect that more efficient methods for estimating $\mathcal{E}_{\mathbf{\Sigma}_X}(\mathcal{S}_{Y|\mathbf{X}})$ are possible.

8.8. Functional data analysis

Although the envelope models at the core of this article concern multivariate linear regression in which $\mathbf{Y}$, and possibly $\mathbf{X}$, are finite-dimensional random vectors, there is no fundamental difficulty in extending our approach to the case where $\mathbf{Y}$ and $\mathbf{X}$ are *random functions*; with an appropriate generalization, the ideas we presented are applicable to functional data analysis. The purpose of this subsection is to demonstrate this possibility by sketching such a generalization under somewhat strong simplifying assumptions. A fuller and more careful generalization will be considered in a future study. This generalization is significant because parsimony is even more important for functional data analysis.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $([0,1], \mathcal{G}, \lambda)$ the measure space, where $\mathcal{G}$ is the class of Borel sets in $[0,1]$ and λ the Lebesgue measure. Next, let $L_2(\Omega, P)$ be the class of all random variables on Ω that are square integrable with respect to P, and $L_2([0,1], \lambda)$ the class of functions defined on $[0,1]$ that are square integrable with respect to λ. Suppose $\varepsilon : \Omega \times [0,1] \to \mathbb{R}$ and $X : \Omega \times [0,1] \to \mathbb{R}$ are mappings such that, for each $t \in [0,1]$, $\varepsilon(\cdot, t)$ and $X(\cdot, t)$ are members of $L_2(\Omega, P)$. Thus $t \mapsto \varepsilon(\cdot, t)$ (or $t \mapsto X(\cdot, t)$) is a random function from $[0,1]$ to $L_2(\Omega, P)$, instead of a random vector from $\{1, \ldots, p\}$ (or $\{1, \ldots, r\}$) to $L_2(\Omega, P)$, as in the multivariate regression model (1.1). For simplicity, we assume both ε and X to be zero-mean functions; that is

$$\int_\Omega \varepsilon(\omega, t) P(d\omega) = 0, \quad \int_\Omega X(\omega, t) P(d\omega) = 0$$

for each $t \in [0,1]$.

Let $\kappa : [0,1] \times [0,1] \to \mathbb{R}$ be a bivariate kernel function and define

$$U(\omega, t) = \int_0^1 X(\omega, s) \kappa(s, t) \lambda(ds).$$

Assume the kernel κ is such that, for each $t \in [0,1]$, $U(\cdot, t)$ belongs to $L_2(\Omega, P)$. Define the functional linear regression model as $Y = U + \varepsilon$. This is a functional version of (1.1), except for ignoring the intercept – which has no bearing on this generalization.

Now, let $\Sigma : [0,1] \times [0,1] \to \mathbb{R}$ and $\Lambda : [0,1] \times [0,1] \to \mathbb{R}$ be the bivariate functions

$$\Sigma(s, t) = \int_\Omega \varepsilon(\omega, s) \varepsilon(\omega, t) P(d\omega), \quad \Lambda(s, t) = \int_\Omega U(\omega, s) U(\omega, t) P(d\omega).$$

For each $f \in L_2([0,1], \lambda)$, let $T_\Sigma(f)$ and $T_\Lambda(f)$ be the functions

$$t \mapsto \int_0^1 f(s) \Sigma(s, t) \lambda(ds), \quad t \mapsto \int_0^1 f(s) \Lambda(s, t) \lambda(ds).$$

Then, under mild conditions on Σ and Λ, T_Σ and T_Λ are bounded linear operators from $L_2([0,1], \lambda)$ to $L_2([0,1], \lambda)$.

Indicating with $\mathcal{B}$ the closure of the linear subspace span$\{T_\Lambda(f) : f \in L_2([0,1], \lambda)\}$, the random function U belongs to $\mathcal{B}$ P-almost surely. We can then define the Σ-envelope of $\mathcal{B}$, say $\mathcal{E}_\Sigma(\mathcal{B})$, as the smallest reducing subspace of the linear operator T_Σ that contains the subspace $\mathcal{S}$. Furthermore, if we assume

Σ to be such that T_Σ is a compact operator, then its spectral decomposition is very similar to that of $\mathbf{\Sigma}$ in model (1.1), with the eigenvectors of $\mathbf{\Sigma}$ replaced by the eigenfunctions of T_Σ. Essentially all the results we developed for the $\mathbf{\Sigma}$-envelope in the previous sections can be extended to this functional linear regression setting.

Restricting the size of the Σ-envelope is not only a means of parsimoniously modeling the variance operator T_Σ, but can also be used to constrain the mean function U. For example, if we assume that the Σ-envelope is finite-dimensional, then we have in effect constrained the mean to be a linear combination of a finite number of functions in $L_2([0,1],\lambda)$. Such a constraint can be useful when the sample size n is relatively small compared to the number of observations on each subject.

9. Conclusions

Our results reveal a crucial property of the classical multivariate linear regression model (1.1), namely, that if the column space of the regression parameter $\boldsymbol{\beta}$ lies within a reducing subspace of the error covariance matrix $\mathbf{\Sigma}$, then far fewer parameters are needed to specify the likelihood. To express this parsimonious parameterization, we introduced the $\mathbf{\Sigma}$-envelope of $\boldsymbol{\beta}$, defined as the smallest reducing subspace of $\mathbf{\Sigma}$ that contains $\mathrm{span}(\boldsymbol{\beta})$. The reparameterized likelihood can be maximized explicitly with the $\mathbf{\Sigma}$-envelope fixed, and maximization with respect to the latter can be performed numerically using Grassmann-manifold optimization. As we demonstrated analytically and on real and simulated data examples, this approach can bring dramatic improvements in accuracy relative to the traditional multivariate linear regression estimator.

We also argued that the notion of enveloping extends well beyond the parsimonious parameterization of the classical multivariate linear model. Obviously, any multivariate model that can be posed as a special case of (1.1) (e.g., a MANOVA model; Johnson and Wichern (2007, Chap. 6)) can be modified through a $\mathbf{\Sigma}$-envelope parameterization. Moreover, parameterizations based on error covariance envelopes could be devised for models that stem from generalizations of (1.1) – e.g., multivariate generalized linear models with non-Gaussian responses depending on the predictor through a multivariate non-linear link function (Fahrmeir and Tutz (1994)), or the functional linear models discussed in Section 8.8. Finally, reaching past linear models, we showed (Section 8) that enveloping can serve as a means to reinterpret, connect, and improve efficiency for a broad range of multivariate statistical techniques.

Acknowledgement

We would like to thank the Editors, an associate editor, and three referees for their thorough and insightful reviews of several versions of the manuscript, which led to significant improvements in both content and presentation. We also thank Inge Helland and Zhihua Su for their comments on previous versions. Zhihua reproduced all of the numerical calculations using independent code. Research for this article was supported in part by NSF grants DMS-0704098 (R. Dennis Cook) and DMS-0704621 (Bing Li and Francesca Chiaromonte).

References

Aldrich, J. (2005). Fisher and regression. *Statist. Sci.* **20**, 401-417.

Anderson, T. (1951). Estimating linear restrictions on regression coefficients for multivariate normal distributions. *Ann. Math. Statist.* **22**, 327-351.

Bura, E. and Cook, R. D. (2003). Rank estimation in reduced-rank regression. *J. Multivariate Anal.* **87**, 159–176.

Christensen, R. (2001). *Advanced Linear Modeling.* Springer, New York.

Conway, J. (1990). *A Course in Functional Analysis.* Second Edition. Springer, New York.

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences*, 18-25. American Statistical Association, Alexandria, VA.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* Wiley, New York.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statist. Sci.* **22**, 1–26.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction without matrix inversion. *Biometrika* **94**, 569-584.

Davies, P. T. and Tso, M. K.-S. (1982). Procedures for reduced rank regression. *Appl. Statist.* **31**, 244-255.

Edelman, A., Tomás, A. A. and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303-353.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models.* Springer, New York.

Faraway, J. and Reed, M. P. (2007). Statistics for digital human motion modeling and ergonomics. *Technometrics* **49**, 277-290.

Helland, I. S. (1990). On the structure of partial least squares regression. *Scand. J. Statist.* **17**, 97-114.

Henderson, H. V. and Searle, S. R. (1979). Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Canad. J. Statist.* **7**, 65-81.

Hoffmann, K., Zyriax, B. C., Boeing, H. and Windler, E. (2004). A dietary pattern derived to explain biomarker variation in strongly associated with the risk of coronary heart disease. *Amer. J. Clin. Nutr.* **80**, 633-40.

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

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis.* Sixth Edition. Pearson Prentice Hall.

Jolliffe, I. T. (2002). *Principal Component Analysis.* Second Edition, Springer, New York.

Leek, J. T. and Storey, J. D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics* **3**, 1724-1735.

Li, K. -C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-327.

Li, K. -C., Aragon, Y., Shedden, K. and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* **98**, 99-109.

Liu, X., Srivastava, A. and Gallivan, K. (2004). Optimal linear representations of images for object recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 662-666.

Reinsel, G. C. and Velu, P. (1998). *Multivariate Reduced Rank Regression, Theory and Applications.* Lecture Notes in Statistics 136. Springer, New York.

Seber, G. A. F. (1984). *Multivariate Observations.* Wiley, New York.

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *J. Amer. Statist. Assoc.* **81**, 142-149.

Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal components. *J. Roy. Statist. Soc. Ser. B* **61**, 611-622.

Yuan, M., Ekici, A., Lu, Z. and Monteiro, R. (2007). Dimension reduction and coefficient estimation in multivariate linear regression. *J. Roy. Statist. Soc. Ser. B* **69**, 329-346.

Zyskind, G. (1967). On canonical forms, non-negative covariance matrices and best and simple least squares linear estimators in linear models. *Ann. Math. Statist.* **38**, 1092-1109.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: dennis@stat.umn.edu

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.

E-mail: bing@stat.psu.edu

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.

E-mail: chiaro@stat.psu.edu

(Received April 2008; accepted June 2009)

COMMENT

Jinzhu Jia, Yuval Benjamini, Chinghway Lim, Garvesh Raskutti and Bin Yu

UC Berkeley

We thank the authors for a very interesting paper and the editors for inviting us to discuss it. Cook, Li and Chiaromonte (2010) develop the envelope model that imposes relationships between the mean parameter matrix β and covariance matrix Σ of the error vector in a linear multi-response regression model. They use the maximum likelihood estimator (MLE) under the envelope model to estimate the mean parameters. As expected, this MLE is asymptotically less variable than the usual OLS if the envelope model holds and the dimension p of the predictor is fixed while sample size n goes to infinity. The question is to what extent this superiority of envelope-MLE remains when the envelope model might not hold, which is typically the case with data.

Reducing the variability of estimates of β is critical in many modern regression settings, even more so when both the dimension p of predictors and number r of responses are large compared to sample size n (Greenland (2000)). To deal with this problem, a common strategy is to use regularization. Regularization for multi-response linear models can be achieved by constraining the parameters of the model or by pooling information from different responses to produce better estimates. Both of these aspects of regularization are found in the envelope model.

The envelope model links the linear space spanned by the parameter vectors in individual models, $\beta \in \mathbb{R}^{p\times r}$, to the covariance matrix of responses errors ($\Sigma \in \mathbb{R}^{r\times r}$), where p is the dimension of the predictor and r is the number of responses for each sample. To be precise, the envelope model assumes that the space spanned by β lies in the linear space spanned by some u eigen vectors of Σ. This link is non-trivial, and the resulting model could be computationally hard to estimate. In this discussion, we call the estimate of β under the envelope model the *envelope-MLE.* Although the authors compare their method to OLS in Cook, Li and Chiaromonte (2010), they do not compare it to standard methods used to reduce variability in estimation. One such method is ridge regression (RR) - a regularization method that uses an ℓ_2 penalty on the estimated β. Another method, Curds and Whey (CW) introduced by Breiman and Friedman (1997), exploits the multiple responses (and β structure) to improve the estimation. That is, find $B \in \mathbb{R}^{r\times r}$, such that

$$B_{i,:} = \arg\min_{b} E\|Y_i - b^T\hat{Y}_{ols}\|_2^2, \quad i = 1,\ldots,r, \tag{1}$$

where $\hat{Y}_{ols}$ is the fitted OLS responses for a given observation with predictors $X \in \mathbb{R}^p$, and $B_{i,:}$ is the ith row of matrix B.

Because it is not usually known with data whether such a link between β and Σ exists, it is crucial to evaluate the performance of different methods in cases where the link does not necessarily hold or the sample size is not large even when the link holds. We compare the performance of envelope-MLE with the algorithms Ridge and CW, both for simulated data and real data. Our experience (admittedly limited) with the envelope model suggests that

(1) the envelope model is best suited to the regime $u < p < r < n$;

(2) the envelope-MLE, as currently implemented, is computationally more intensive than Ridge and CW.

1. Experiments

1.1. Simulated data

Two simulation scenarios are used. The first is based on the envelope model

$$\begin{aligned} Y &= \Gamma\eta X + \epsilon, \\ \Sigma &= \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T, \end{aligned} \tag{1.1}$$

where $Y \in \mathbb{R}^r$, (Γ, Γ_0) are the eigenvectors of $\Sigma := \mathrm{Cov}\,(\epsilon)$, $\Omega = \Gamma^T\Sigma\Gamma \in \mathbb{R}^{u\times u}$, and $\Omega_0 = \Gamma_0^T\Sigma\Gamma_0 \in \mathbb{R}^{(r-u)\times(r-u)}$. We take $\Sigma = \Gamma^T D_r\Gamma$, where Γ are the eigenvectors of a random matrix (elements in $N(0,1)$), and D_r is a diagonal matrix with the elements $1,\ldots,r$. We simulated $\eta \in R^{u\times p}$ from $\eta_{ij} \sim N(0,1)$ and generated $\beta = \Gamma\eta$ accordingly.

In the second simulation scenario, no structural link between β and Σ is assumed. We generated Σ similar as before, and β was a random matrix, $\beta = G\eta$, where the elements of $G \in \mathbb{R}^{r\times u}$ were $N(0,1)$. Note that u in this "random model" is the rank of β, but it is not related to the structure of Σ. For both models, X was generated from $\mathcal{N}(0, V)$, where $V_{ij} = 0.5^{|i-j|}, i, j = 1, 2, \ldots, p$.

The measure used here for comparison is the overall average mean-squared prediction error. Following Breiman and Friedman (1997), the mean-squared prediction error of response i, for a particular method m, is

$$e_i^2(m) = E_X\left(\beta_{i,:}X - \hat{\beta}(m)_{i,:}X\right)^2 = (\beta_{i,:} - \hat{\beta}_{i,:}(m))V(\beta_{i,:} - \hat{\beta}_{i,:}(m))^T,$$

where V is the covariance matrix of X. We compare the average prediction error of each method normalized by the OLS average prediction error.

We considered the following four set-ups: $p > r$ or $p < r$, $u = \min(p, r)$ or $u < \min(p, r)$. When $p > r$, $p = 50$, $r = 10$; when $p < r$, $p = 10$, $r = 50$; when $u < \min(p, r)$, $u = 3$. For each of the four set-ups, we did 50 repetitions to evaluate each estimation method, and for each repetition we took sample size $n = 500$.

1.2. Data

The data is taken from Kay et al. (2008), and consists of a training set of $n = 1,750$ samples and a validation set of $n = 120$ samples. Each sample consists of $p = 64$ predictor variables and $r = 143$ responses. The data is from an experiment measuring hemodynamic response to natural image stimuli in the visual cortex of the brain using functional Magnetic Resonance Imaging (fMRI). The predictor variables measure magnitude of a spatial grating (Gabor filter) at different positions in the image (an 8 by 8 grid). The responses are measures of the fMRI response at different locations in the visual cortex. The task is to predict the fMRI responses in these locations to a new natural image stimuli – or the encoding problem in computational neuroscience.

The training set was split into an estimation set $n = 1,500$, and a set for model selection and regularization parameter optimization ($n = 250$). The best models based on the model selection set were compared on a third validation set ($n = 120$) with a better signal-to-noise ratio. Prediction accuracy was estimated using mean-square-error of the prediction to the measured responses on the validation set. These results were then normalized by the OLS mean square error for the corresponding response.

2. Results

Our results show that in most scenarios the competing methods achieved predictions as good as the envelope-MLE, and were much faster.

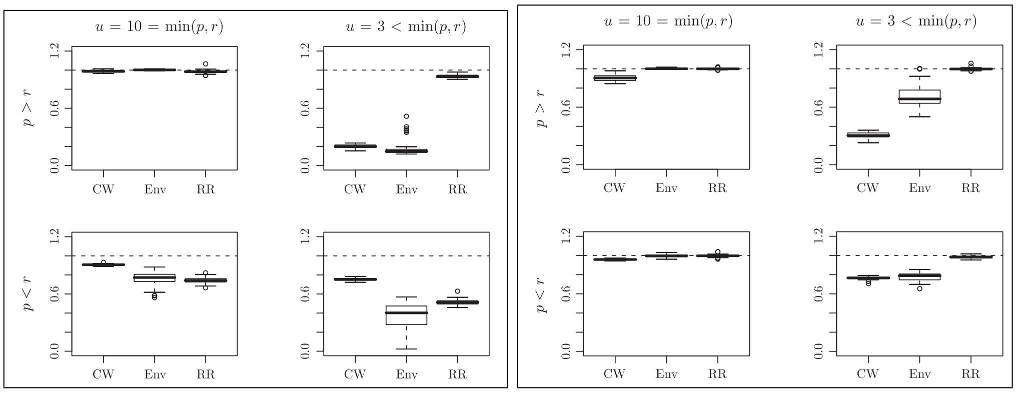
2.1. Prediction performance comparison

When the envelope model held, envelope-MLE was the most successful method when $u < p < r$ (right hand corner of (a) in Figure 2.1): it achieved lower prediction errors compared to the other methods, although it had a larger variability. In other cases, Ridge and/or CW achieved comparable results as envelope-MLE: CW when $u < r < p$, Ridge when $u = p < r$, or both when $u = r < p$.

However, this was not the case when data was not generated from an envelope model. While the envelope-MLE procedure performed better than unrestricted OLS, CW outperformed envelope-MLE (Figure 2.1 (b)), and Ridge was comparable to envelope-MLE in two cases and worse in the other two.

In the data example (Figure 2.2), all three methods gave comparable performance: the error of envelope-MLE was slightly worse than Ridge and slightly better than CW, even though envelope-MLE seems to have the largest variability. All methods performed better than OLS, ($SE \approx 0.003$, differences were significant; the results are lower than those reported in Kay et al. (2008) because of the restriction to 64 predictors).

As our results show (Figure 2.1), the envelope model is best suited to a classical regime where $u < p < r \ll n$. If $n < p$ or $n < r$, either Σ or β would be under-complete and would not be identifiable under the constraints of the model. We need $r > u$ for the regularization to be effective (otherwise the dimension of the envelope is already r and we get the OLS results).



(a) Envelope Models (b) Random Models

Figure 2.1. The average (over responses) prediction error of each method normalized by OLS average prediction error. The distribution in each boxplot corresponds to 50 repetitions of the simulation. Left panel (a): Envelope Models. Right panel (b): Random Models. Note that u in "Random Models" (b) is not the dimension of the envelope. CW denotes the Curds and Whey, Env the Envelope model, and RR the Ridge regression with tuning parameter tuned by 5-fold cross validation. The dashed line denotes the benchmark that one method performs the same as OLS (and thus the ratio is 1).

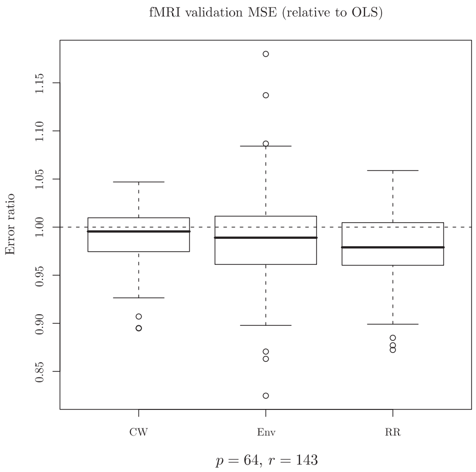


Figure 2.2. Prediction error for individual responses (normalized by OLS error) for image-fMRI data. Boxplots show distribution of $r = 143$ responses. All methods are better than OLS (Median of error ratios < 1). Ridge performs best, followed by envelope-MLE. Each boxplot corresponds to a single point in the simulations of Figure 2.1.

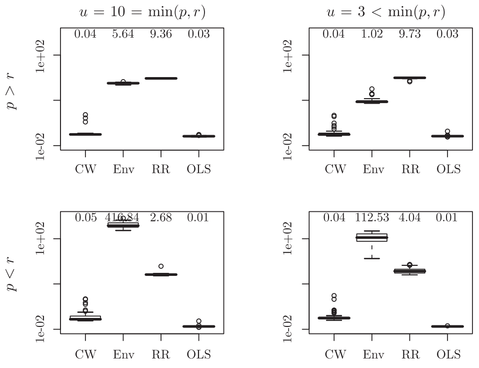


Figure 2.3. Run Time (in Seconds) for each iteration. The times are shown in log scale. CW denotes the Curds and Whey, Env the Envelope model, RR the Ridge regression with tuning parameter tuned by 5-fold cross validation, OLS the ordinary least square method.

2.2. Runtime comparison

The envelope-MLE method was always more computationally intensive than all other methods used in the regime $p < r < n$ (see Figure 2.3). Since this is the regime where the envelope model is most useful, the result highlights the need to improve efficiency of the algorithm. In fact, for the fMRI data, using the original implementation of envelope-MLE, the algorithm could not run in reasonable time. Instead we used a parameter tuning set to tune u (the dimension of $\mathcal{E}_{\Sigma}(\mathcal{B})$) to improve computation speed. It took the Envelope model estimation more than 300 seconds to run in the optimal setting $u = 20$, and up to 2,500 seconds for an envelope dimension of $u = 100$, while the other methods are very efficient (less than 0.1 seconds for each).

The model selection procedure for the envelope model was fairly stable. The error in the training set reduces as u increases (when $u = 143$ the model is equivalent to OLS which minimizes the training set error). However, in both the parameter tuning set and validation set the prediction errors were minimized by the same value, $u = 20$ (see Figure 2.5).

3. Conclusion

To summarize, the envelope model provides a novel way of regularization for multi-response linear regression problems. Our simulation and data results suggest that the envelope model works best in the classical domain when $u < p < r < n$ and the envelope model holds. More experience is needed for us to better understand the envelope model relative to other regularization methods such as Ridge regression and CW, especially when $\min(r, p) \gg n$. This is feasible only when faster codes become available for fitting the envelope model.

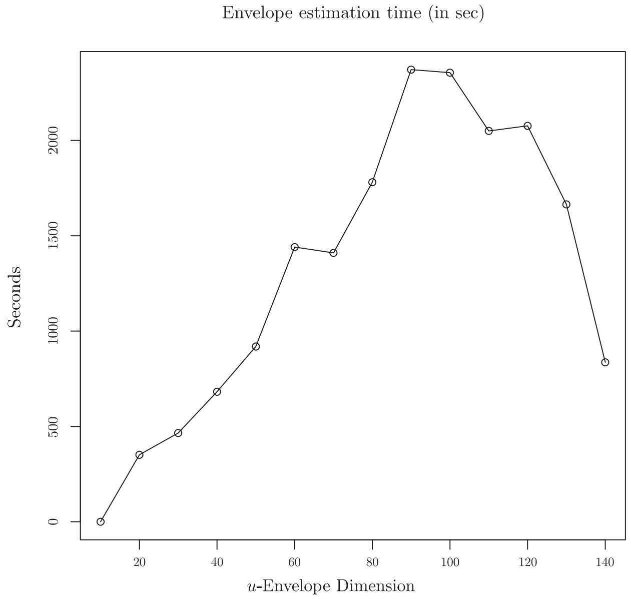


Figure 2.4. Run time for different u on the fMRI data. The time required for the envelope-MLE estimation (300-2,400 seconds) restricted testing larger parameter and responses sets.

Acknowledgement

This research is partially supported by NSF grants DMS-0605165, DMS-0907632, and SES-0835531. Yuval Benjamini is partially supported by an NSF VIGRE Graduate Fellowship. In addition, Bin Yu is partially supported by a grant from MSRA.

References

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

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica.* **20**, 927-960.

Greenland, S. (2000). Principles of multilevel modelling. *International Journal of Epidemiology* **29**, 158-167.

Kay, K., Naselaris, T., Prenger, R. and Gallant, J. (2008). Identifying natural images from human brain activity. *Nature* **452**, 352 - 355.

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.

E-mail: jzjia@stat.berkeley.edu

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.

E-mail: yuvalb@stat.berkeley.edu

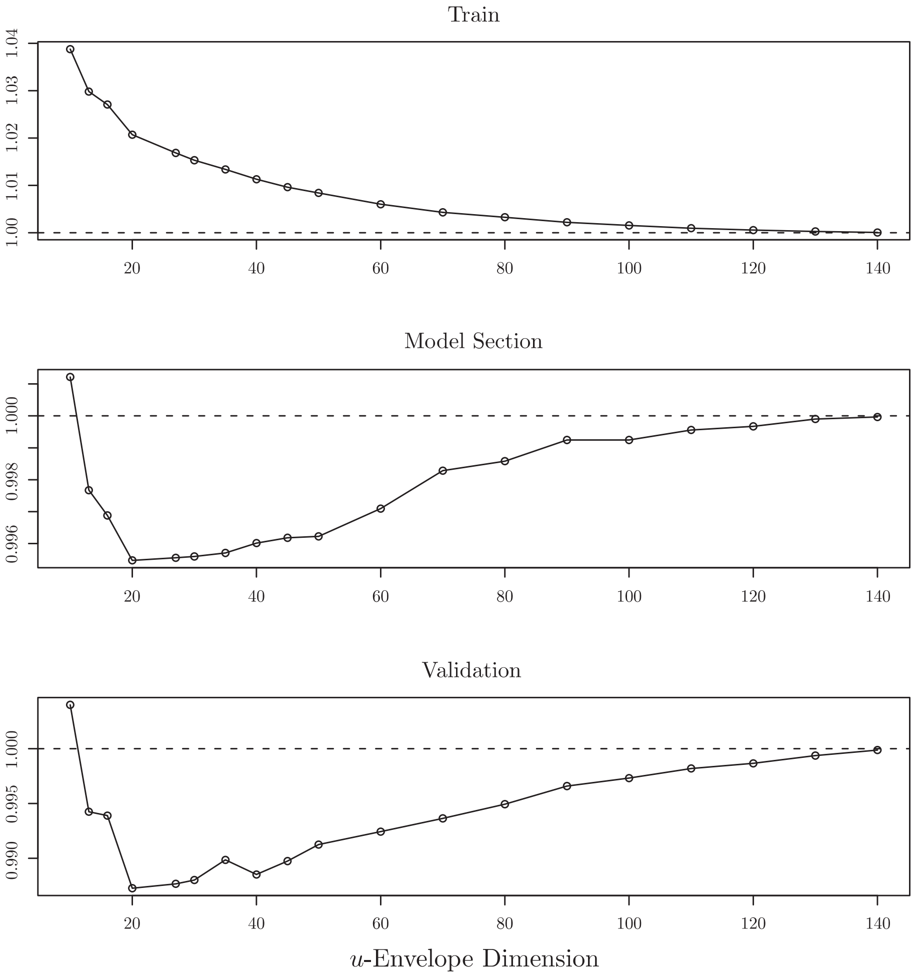


Figure 2.5. Prediction errors of envelope-MLE (normalized by OLS errors) on fMRI data for training, parameter selection, and validation sets when u varies. For $u = r$ the envelope errors are similar to OLS errors. Both parameter tuning set and validation set prediction errors were minimized at $u = 20$.

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.
E-mail: lim@stat.berkeley.edu

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.
E-mail: garveshr@stat.berkeley.edu

Department of Statistics, UC Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A.
E-mail: binyu@stat.berkeley.edu

(Received January 2010; accepted February 2010)

COMMENT

Liqiang Ni

University of Central Florida

I congratulate the authors for their innovative approach to multivariate linear regression. Like the authors, I was puzzled that, with or without known Σ, the MLE of β stayed the same. Should not the knowledge of Σ help the estimation of the mean function? The proposed envelope models provide an answer. I believe this simple but powerful idea will have a deep impact on a variety of topics where establishing connections between seemingly unrelated parts of the enquiry can be beneficial.

My comments focus on the minimization of (4.9) and the comparison between its estimator and an alternative. First let me introduce a lemma and a proposition, which facilitate the discussion.

Lemma 1. *Suppose that A and B are positive definite matrices in $\mathbb{R}^{p\times p}$. Let $X \in \mathbb{R}^{p\times d}$, $\mathrm{rank}(X) = d$. If X minimizes $\det(X^TAX)/\det(X^TBX)$, then $\mathrm{span}(X) = \mathrm{span}(U)$, where U is the d leading eigenvector of $A^{-1}B$.*

Proof. Suppose $A^{1/2}X$ has a QR decomposition VW, where $V \in \mathbb{R}^{p\times d}$ and $V^TV = I_d$.

$$\frac{\det(X^TAX)}{\det(X^TBX)} = \frac{\det(W^TW)}{\det(W^TV^TA^{-1/2}BA^{-1/2}VW)} = \frac{1}{\det(V^TA^{-1/2}BA^{-1/2}V)},$$

which reaches the minimum when V consists of the d leading eigenvectors of $A^{-1/2}BA^{-1/2}$ or, equivalently, $\mathrm{span}(X) = \mathrm{span}(A^{-1/2}V)$ is spanned by the d leading eigenvectors of $A^{-1}B$.

Proposition 1. *Suppose that A and B are positive definite matrices in $\mathbb{R}^{p\times p}$. Suppose that A has the spectral decomposition $A = U\Lambda U^T$, where $U = (U_1, U_2)$, the diagonal matrix $\Lambda = \mathrm{diag}(\Lambda_1, \Lambda_2)$, $U_1 \in \mathbb{R}^{p\times d}$, $\Lambda_1 \in \mathbb{R}^{d\times d}$. Meanwhile $B = U\Phi U^T$, where the block diagonal matrix $\Phi = \mathrm{diag}(\Phi_1, \Phi_2)$, where $\Phi_1^{-1/2}\Lambda_1\Phi_1^{-1/2} < \Phi_2^{-1/2}\Lambda_2\Phi_2^{-1/2}$. Let $P \subset \mathbb{R}^d$ be any projection with $\dim(P) = d < p$, and let Q be its complement. Then $P = P_{U_1}$ minimizes $\det(PAP + QBQ)$.*

Proof. Suppose there is a projection $P_1 = P_{U\Gamma_1}$, where $\Gamma_1 \in \mathbb{R}^{p\times d}$ is a semi-orthonormal matrix. Let $\Gamma_2 \in \mathbb{R}^{p\times(p-d)}$ be the orthogonal complement of Γ_1. It is easy to see that

$$\begin{aligned}\det(\Phi) &= \det[(\Gamma_2,\Gamma_1)^T\Phi(\Gamma_2,\Gamma_1)]\\ &= \det(\Gamma_2^T\Phi\Gamma_2)\det[\Gamma_1^T\Phi\Gamma_1 - \Gamma_1^T\Phi\Gamma_2(\Gamma_2^T\Phi\Gamma_2)^{-1}\Gamma_2^T\Phi\Gamma_1]\\ &\le \det(\Gamma_2^T\Phi\Gamma_2)\det(\Gamma_1^T\Phi\Gamma_1),\end{aligned}$$

where the equality holds if and only if $\Gamma_1^T \Phi \Gamma_2 = 0$. Therefore,

$$\begin{aligned}\det(P_1 A P_1 + Q_1 B Q_1) &= \det(\Gamma_1^T \Lambda \Gamma_1) * \det(\Gamma_2^T \Phi \Gamma_2) \\ &\geq \det(\Phi) \det(\Gamma_1^T \Lambda \Gamma_1) \det(\Gamma_1^T \Phi \Gamma_1)^{-1}. \qquad (1)\end{aligned}$$

Based on Lemma 1, the RHS of (1) reaches its minimum by letting $\Gamma_1 = (I_d, 0)$, the d leading eigenvectors of $\Lambda^{-1}\Phi$. It is easy to verify that $\Gamma_1^T \Phi \Gamma_2 = 0$, hence the conclusion.

We have $\hat{\Sigma}_{res} \to \Sigma$ and $\hat{\Sigma}_Y \to \Sigma_Y = \Sigma + \beta \Sigma_X \beta^T$ as $n \to \infty$. Therefore, (4.9) converges to

$$\det(P_{\Sigma_1} \Sigma P_{\Sigma_1} + Q_{\Sigma_1} \Sigma_Y Q_{\Sigma_1}). \qquad (2)$$

If we assume the envelope structure with $\beta = \Gamma\eta$, we have $\Sigma = P_\Gamma \Sigma P_\Gamma + Q_\Gamma \Sigma Q_\Gamma$ and $\Sigma_Y = P_\Gamma \Sigma_Y P_\Gamma + Q_\Gamma \Sigma Q_\Gamma$. In other words, there are two blocks of spectra where Σ_Y is the same as Σ in one block, and is larger than Σ in the other in which they do not have to share the same eigenvectors. This is a special case of Proposition 1 with $\Lambda_1 < \Phi_1$ and $\Lambda_2 = \Phi_2$. The subspace $P_{\Sigma_1} = P_\Gamma$ minimizes (2), which is spanned by the leading eigenvector of $\Sigma^{-1}\Sigma_Y$. This leads to a Henderson's method-3-type estimator (Henderson (1953)) based on $\hat{\Sigma}_{res}^{-1}\hat{\Sigma}_Y$, denoted as $\hat{\Gamma}_{H3}$. This argument can be considered as an extension of Proposition 5 in Cook (2007). On reading this paper, I was fascinated by Christensen's comments (Christensen (2007)) which adeptly connected Cook's main theme with multivariate linear regression. I also agree with Christensen that the MLE is probably sharper than $\hat{\Gamma}_{H3}$.

If we do not assume the block structure above, we face the task of minimizing $\det(PAP{+}QBQ)$ as in (4.9). While special algorithms on manifolds are available, here I propose a relatively simple iterative scheme. Let $U_1 \in \mathbb{R}^{p\times d}$ and $U_2 \in \mathbb{R}^{p\times(p-d)}$ be orthonormal complements. It is easy to see that

$$\begin{aligned}\det(B) &= \det(U_2^T B U_2) \det(U_1^T B U_1^T - U_1^T B U_2 (U_2^T B U_2)^{-1} U_2^T B U_1) \\ &= \det(U_2^T B U_2) \det(U_1^T B^{1/2} Q_{B^{1/2}U_2} B^{1/2} U_1) \\ &= \det(U_2^T B U_2) \det(U_1^T B^{1/2} P_{B^{-1/2}U_1} B^{1/2} U_1) \\ &= \det(U_2^T B U_2) \det(U_1^T B^{-1} U_1).\end{aligned}$$

Therefore, we only need to minimize $\det(U_1^T A U_1) \det(U_1^T B^{-1} U_1)$. If $d = 1$, this minimization can be easily handled by build-in functions in statistical softwares such as *nlm* in *R*. For $d > 1$, we may minimize the function with respect to one target column with the remaining $d-1$ columns fixed. Then, we rotate columns. This always converges since the minimum values are monotonically decreasing.

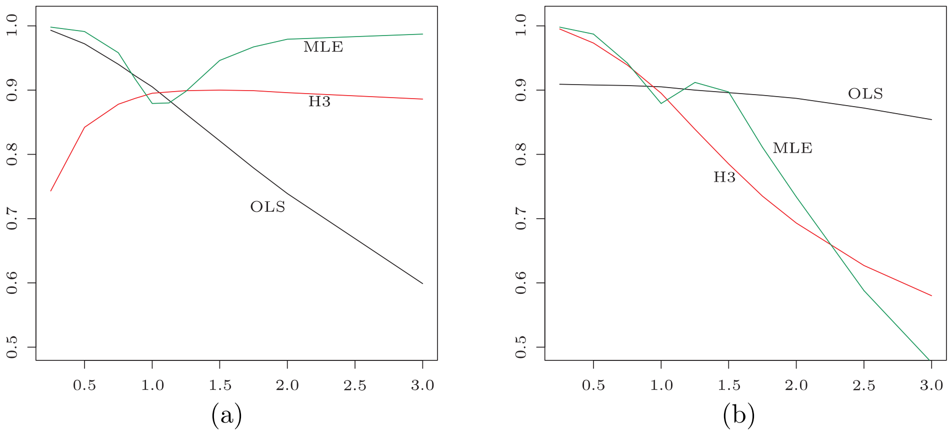


Figure 1.

To avoid traps in local minima, we can start with multiple randomly generated initial values.

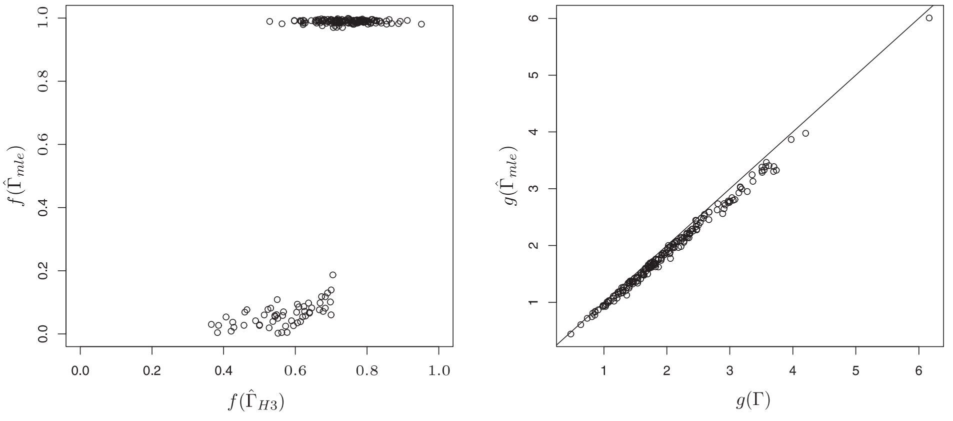
I ran a simulation using the two normal means setting in Section 7.1. Consider

$$Y = \alpha + \Gamma X + \sigma\Gamma\epsilon + \sigma_0\Gamma_0\epsilon_0,$$

where $Y \in \mathbb{R}^{10}$, $\alpha = 0$, $\Gamma = (1/\sqrt{10}, ..., 1/\sqrt{10})$, $X \in \mathbb{R}$ equals either 0 or 1 with equal proportions, Γ_0 is the orthonormal complement, ϵ and ϵ_0 are independent $N(0, I_{10})$ variates. To evaluate the performance of $\hat{\Gamma}_{ols}$, $\hat{\Gamma}_{h3}$, and $\hat{\Gamma}_{mle}$, we use the criterion $f(v) = |v^T\Gamma|/(\|v\|\|\Gamma\|)$. We also define $g(v) = \det(P_v\hat{\Sigma}_{res}P_v + Q_v\hat{\Sigma}_Y Q_v)$, where both matrices have been re-scaled for numerical stability such that the fifth eigenvalue of $\hat{\Sigma}_{res}$ is 1.

First fix $\sigma = 1$ and vary σ_0. Figure 1(a) shows the average of f of 200 replications with sample size $n = 160$. No surprise that the OLS estimator degenerates with an increasing level of the noise since it does not incorporate any structural information. There is an interesting comparison between $\hat{\Gamma}_{H3}$ and $\hat{\Gamma}_{mle}$. The performance of $\hat{\Gamma}_{H3}$ deteriorates with increasing σ_0. At $\sigma_0 = 1$, its performance is similar to those of Γ_{ols} and Γ_{lme} as expected. In contrast, Γ_{lme} seems to have a dip at $\sigma_0 = 1$, and is uniformly superior to the other two.

Next fix $n = 160$, $\sigma_0 = 1$ and vary σ. Figure 1(b) presents the results. With a dilution of the proportion of the signal (in the block of Γ) which can contribute to X, Γ_{H3} degenerates monotonically as expected, since $\Sigma_{res}^{-1}\Sigma_Y \approx (\sigma_x^2/\sigma^2)\Gamma\Gamma^T + I_{10}$. It seems interesting that the Γ_{mle} improves initially beyond $\sigma = 1$ then follows the same pattern as Γ_{H3}. This saddle pattern in the neighborhood of $\sigma = 1$ resembles the one near $\sigma_0 = 1$ in Figure 1(a), and it even becomes more conspicuous with increasing sample sizes where the right peak shifts further right (not shown here).



(a)

(b)

Figure 2.

There is also a huge difference between these two methods even when both are lackluster. For a given large σ, all the Γ_{H3} estimates are closely clustered, the Γ_{mle} estimates go to extremes with either $f \approx 1$ or $f \approx 0$. For example, let $n = 160$ and $\sigma = 2$, Figure 2(a) shows the 200 pairs of $(f(\hat{\Gamma}_{H3}), f(\hat{\Gamma}_{mle}))$. One might wonder if the Γ_{mle}'s were trapped in local minima. Figure 2(b) shows the 200 pairs of $(g(\Gamma), g(\hat{\Gamma}_{mle}))$, which dispels this doubt. If we fix σ but increase n, the proportion of Γ_{mle} with $f \approx 1$ increases. This exploratory simulation reassures one about the superiority of the MLE.

References

Christensen, R. (2007). Comments on Fisher lecture. *Statist. Sci.* **22**, 27-31.

Cook, R. D. (2007). Fisher lecture: Dimension reduction in regression (with discussion). *Statist. Sci.* **22**, 1-43.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226-252.

Department of Statistics, University of Central Florida, Orlando, FL 32816-2370, USA.

E-mail: lni@mail.ucf.edu

(Received November 2009; accepted February 2010)

COMMENT

Xuming He and Jianhui Zhou

University of Illinois at Urbana-Champaign and University of Virginia

The paper by Professors Cook, Li, and Chiaromente is stimulating, pioneering what they call envelope models to achieve dimension reduction by taking advantage of a possible parametric link between the regression coefficients and the conditional covariance matrix of the response vector. They develop a sophisticated computational algorithm for data-adaptive pursuits of the envelope models, and demonstrate visible efficiency gains through their efforts. In this discussion, we present some of our own perspectives and explorations that aim to take us back to some more transparent methods of building parsimonious models.

1. A Model of Idealism ?

Parsimonious modeling is at the core of statistical analysis. The pursuit of parsimonious models through dimension reduction has become more important as we are increasingly challenged by the need to analyze high dimensional data. Cook, Li, and Chiaromente are to be congratulated for their skillful use of envelope models for parsimonious and efficient multivariate regression analysis. They aim to capture a parametric link between the regression coefficient and the (conditional) covariance matrix of the response vector, and apply the maximum likelihood estimator to a problem of reduced (and hopefully much reduced) dimension for higher efficiency.

The theoretical foundation for the envelope models for multivariate regression is very appealing. When a small number of the eigenvectors of Σ, the conditional covariance matrix of the response vector $y \in R^r$, span a linear space that contains the space spanned by β, the unknown regression coefficient matrix, we should perform MLE on the reduced model by removing the uninformative sub-space of y. The envelope is the reduced model with the smallest possible dimension, denoted by u. The smaller u is relative to r, the more efficiency gain one can expect from their proposed estimator.

Their theory comes with some strings attached. Because β and Σ are unknown in practice, finding the envelope with $u < r$ has to be partly faith and partly luck. If we are presented problems of multivariate linear regression with fixed β, but Σ is drawn randomly from a reasonable class of distributions, finding such an envelope would fail almost surely. Just take the simple case of $dim(X) = 1$ with $\beta \in R^r$; to find an envelope with $u = 1$, β has to be an eigenvector of Σ, a "probability zero" event.

Suppose that we are lucky to be in a scenario where β is an eigenvector of Σ, and

$$\Sigma = \sigma^2 \Gamma\Gamma' + \sigma_0^2 \Gamma_0 \Gamma_0',$$

where $\Gamma = \beta/||\beta||$, and $\Gamma_0\Gamma_0' = I - \Gamma\Gamma'$ in the notation of the discussed paper. If the first component of y is scaled up by a factor of 2, the response vector satisfies a similar multivariate response with $\beta^* = A\beta$ and $\Sigma^* = A\Sigma A$, with

$A = diag(2, 1, \ldots, 1)$. Except for some lucky choices of β and Σ, this new β^* is no longer an eigenvector of Σ^*, indicating that the envelope model with $u = 1$ can no longer be found.

The dependence of the envelope model on specific choices of scale for the components of y could be taken as evidence that the exact mathematical framework for the envelope model does not capture the essence of the underlying regression model. We naturally ask why the concept and the methodology are still useful.

2. A Form of Shrinkage?

Like any other parsimonious models, the envelope model does not have to be exactly correct for it to do its job, and it is often worth reducing variability of the estimator by tolerating some bias in the model. We emphasize this point, because Section 7 of the discussed paper could be misleading when it states that one need only compare variability with no appreciable bias detected. However, in realistic problems, we do not expect the envelope model to hold exactly, and bias should be part of the performance metric. When r is large, β is likely to lie in a small neighbored of u eigenvectors of Σ, which often makes bias a worthy sacrifice.

We bring up the bias-variance trade-off to lead us to the concept of shrinkage (James and Stein (1961)). More efficient estimation is often achieved without reliance on any formal dimension reduction model. To see how shrinkage works in the simulation setting of Section 7.1 of the discussed paper, we conducted a simulation study using the penalized likelihood on the full model

$$-\log(likelihood) + \lambda||\beta - \bar{\beta}||_1,$$

where λ is a tuning parameter, and $\bar{\beta}$ is the average of all components of β. The penalty used here provides shrinkage of the regression parameters to a common value.

Using the model in Section 7.1 of Cook, Li, and Chiaromente with $\beta = (\sqrt{10}, \sqrt{10}, \ldots, \sqrt{10}) \in R^{10}$, $\Sigma = \beta\beta'/||\beta||^2 + \sigma_0^2\Gamma_0\Gamma_0'$, and $\Gamma_0\Gamma_0' = I - \beta\beta'/||\beta||^2$, we obtained the ratio of the mean squared error of the MLE under the full model versus the shrinkage estimator, under several values of σ_0^2. Figure 2.1 (a) plots these ratios for $\lambda = 1$. Together with Figure 7.1 of the discussed paper, it indicates that the shrinkage estimator improves on the efficiency of not only the MLE under the full model, but also the proposed estimate based on the envelope model of $u = 1$.

It is probably not surprising that shrinking toward a common value would do well when all the components of β are indeed equal. The dramatic improvement in the efficiency for the cases with large values of σ_0^2 is due to the nature of Σ

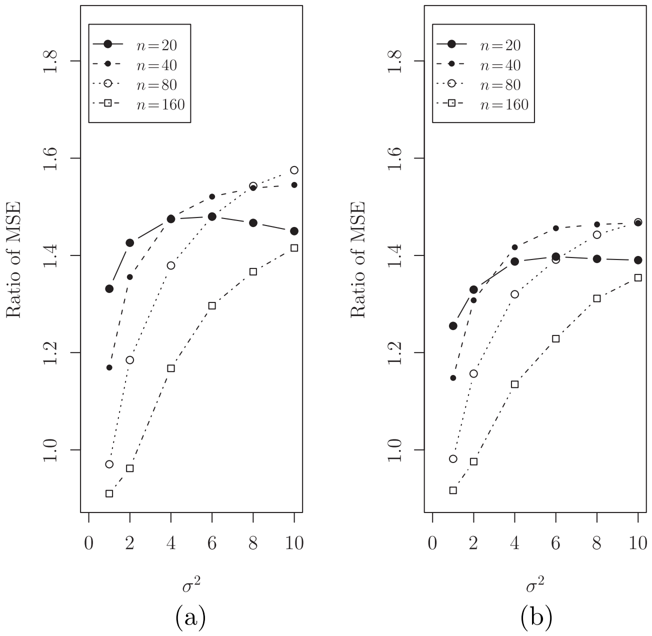


Figure 2.1. Ratio of MSE (MLE under the full model versus the shrinkage estimates with $\lambda = 1.0$). Panel (a) corresponds to the model with $\beta = (\sqrt{10}, \sqrt{10}, \ldots, \sqrt{10})$. Panel (b) corresponds to the model with $\beta = (2\sqrt{10}, \sqrt{10}, \ldots, \sqrt{10})$.

in this particular model. When $\sigma_0^2 = 8$, for example, the errors are negatively correlated with one another and Σ is nearly singular, so that the average over the ten components of y has a very small variance.

If we consider a less favorable setting with $\beta = (2\sqrt{10}, \sqrt{10}, \ldots, \sqrt{10})$ and no other changes in the model, then the results of the shrinkage estimate are shown in Figure 2.1 (b).

Like the estimator of Cook, Li and Chiaromente, the simple shrinkage estimator achieves greater efficiency gains for smaller n and for higher values of σ_0. The issue of λ selection certainly needs to be discussed if the shrinkage method is to be used in more general problems, but the connection, and the differences, between shrinkage and envelope models is worth exploring.

3. The Road Not Taken?

In the spirit of dimension reduction, we explored our favorite method of canonical correlation (CANCOR) as in Zhou and He (2008). In this case, we can simply find the significant canonical variates of Y. If $r > p$, we use $u \leq p$ significant canonical variates in lieu of Y itself.

Suppose that R is an $r \times r$ rotation matrix, and

$$Y_i^* = R^T Y_i = (y_{i,1}^*, y_{i,2}^*, \ldots, y_{i,r}^*)^T,$$

where the first u components of Y^* are the significant canonical variates. let $\beta^* = R^T \beta$.

We consider a reduced model

$$Y_{iu}^* = \begin{pmatrix} y_{i,1}^* \\ \vdots \\ y_{i,u}^* \end{pmatrix} = \beta_u^* X_i + \epsilon_{iu}^*, \tag{3.1}$$

where ϵ_{iu}^* is the first uth component of $\epsilon_i^* = R^T \epsilon_i$. The covariance matrix of ϵ_i^* given X_i is $R^T \Sigma R$. We obtain the least squares estimate $\hat{\beta}_u^*$ of β_u^* or, in general, the MLE under the reduced model, and then use $(\hat{\beta}_u^*, 0)$ as the estimate of β^*, that is, the last $p - u$ rows of β^* are set to zero. Finally, we obtain the estimate of β based on $\beta = R\beta^*$.

This is a simple estimator from the computational viewpoint, because we only need to use the standard software on least squares and canonical correlation. The determination of u can also be done based on standard multivariate tests on canonical correlations (e.g., Anderson (1984)). For convenience, we call this the "C-estimator". Does the C-estimator improve on the efficiency of the MLE under the full model?

If $u = p$, the last $r - p$ canonical variates of Y are uncorrelated with X, so the C-estimator is equivalent to the Gaussian MLE under the full model. Gains of efficiency are possible when $u < p$.

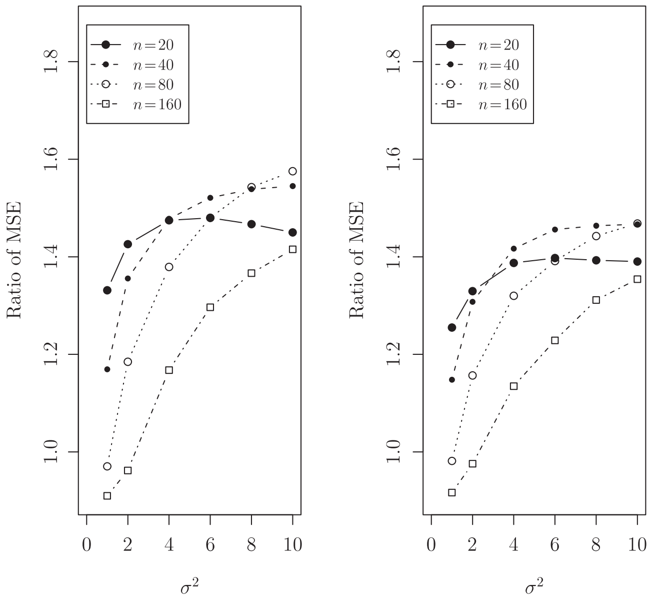
To see how much efficiency gain can be obtained with the C-estimator, we conducted a simulation study with samples generated from

$$Y_i = \beta X_i + \epsilon_i, \tag{3.2}$$

where $Y_i \in R^{10}$, $\beta = (\beta_1, \ldots, \beta_p)$ in an $10 \times p$ matrix, $X_i \in R^p$ are generated from the standard multivariate normal, and $\epsilon_i \in R^{10}$ is multivariate normal with mean 0 and covariance matrix $diag(\sigma^2, \ldots, \sigma^2)$.

In Case I, we chose $p = 2$, $\beta_1 = (1, 1, \ldots, 1)$ and $\beta_2 = (1.5, 1, \ldots, 1)$. Figure 3.2 gives the MSE ratio between the C-estimator and the MLE under the full model at various sample sizes and various values of σ^2. The average value of the estimated u in the simulation based on 500 Monte carlo samples was closer to 1 for smaller values of n and larger values of σ^2; see Table 3.1.

In Case II, we chose $p = 3$, $\beta_1 = (1, 1, \ldots, 1)$, $\beta_2 = (2, 2, 2, 2, 2, 1, 1, 1, 1, 1)$, and $\beta_3 = \beta_1 + \beta_2$. Figure 3.3 shows the ratios of MSE, where the efficiency gain was less impressive than in Case I, but still non-negligible.



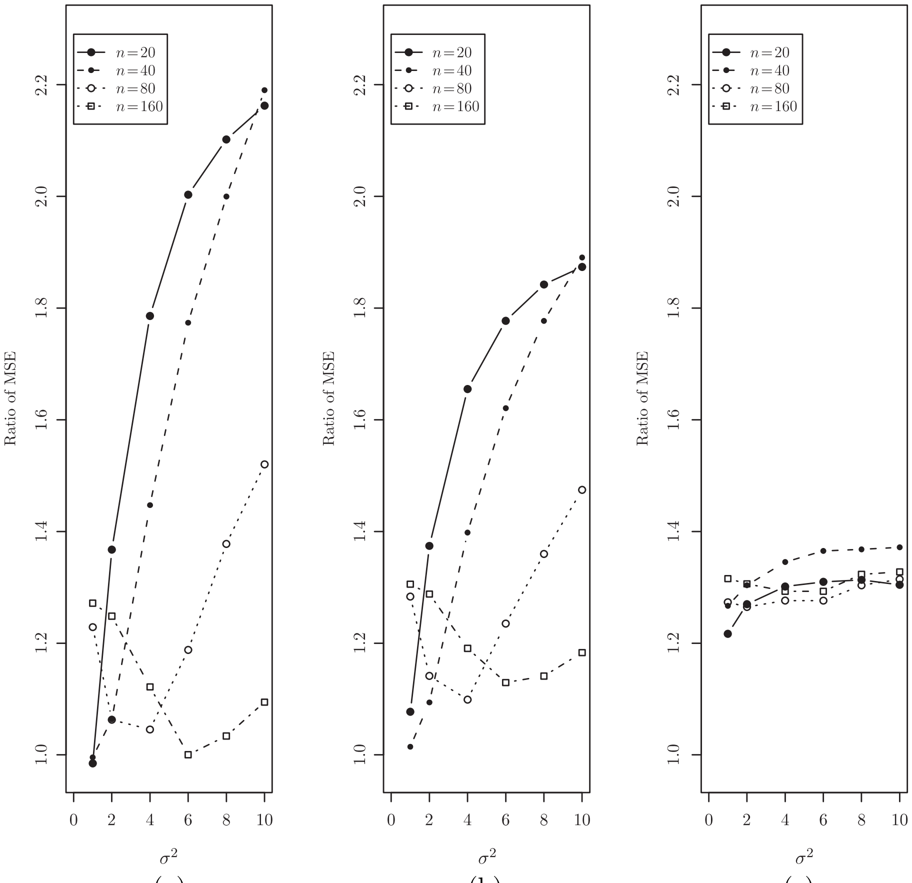
(a) (b)

Figure 3.2. Ratio of MSE (MLE under the full model versus the C-estimator). The model parameters are $\beta_1 = (1, 1, \ldots, 1)$, $\beta_2 = (1.5, 1, \ldots, 1)$. Panel (a) corresponds to the ratio for β_1 and Panel (b) corresponds to the ratio for β_2.

Table 3.1. Mean values of the estimated u by CANCOR in Case I and Case II, at different values of n and σ^2.

	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 4$	$\sigma^2 = 6$	$\sigma^2 = 8$	$\sigma^2 = 10$
Case I						
$n =$ 20	1.12	1.09	1.07	1.07	1.07	1.06
$n =$ 40	1.16	1.10	1.08	1.07	1.07	1.07
$n =$ 80	1.42	1.21	1.14	1.11	1.09	1.08
$n =$ 160	1.80	1.46	1.23	1.16	1.13	1.11
Case II						
$n =$ 20	1.35	1.23	1.15	1.13	1.11	1.11
$n =$ 40	1.82	1.47	1.22	1.16	1.14	1.12
$n =$ 80	2.05	1.94	1.57	1.41	1.31	1.26
$n =$ 160	2.05	2.04	1.97	1.75	1.60	1.47

We do not have a systematic comparison between the C-estimator and the estimator of Cook, Li and Chiaromente, but we note that the C-estimator does not take advantage of the the relationship between β and Σ. Rather, it is based on the marginal correlation matrices of X and Y. The general message we have learned from the work of Cook, Li and Chiaromente is invariant of the road we



(a) (b) (c)

Figure 3.3. Ratio of MSE (MLE under the full model versus the C-estimator). The model parameters are $\beta_1 = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$, $\beta_2 = (2, 2, 2, 2, 2, 1, 1, 1, 1, 1)$, and $\beta_3 = \beta_1 + \beta_3$. Panels (a), (b), and (c) correspond to β_1, β_2, and β_3, respectively.

may take when faced with high dimensional data: statistical inference is often better done with a reduced model. It is up to our imagination to choose a reduced model, and we are glad to have had the opportunity to learn about a rather sophisticated approach taken by Cook, Li and Chiaromente, and to explore a few simpler alternatives.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis.* John Wiley, New York.

Cook, R. D., Li, B. and Chiaromonte, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica.* **20**, 927-960.

James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proc. Fourth Berkeley Symp. Math. Statist. Prob.* **1**, 361-379.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *Ann. Statist.* **36**, 1649-1668.

Department of Statistics, University of Illinois, 725 S. Wright, Champaign, IL 61820, USA.

E-mail: x-he@uiuc.edu

Department of Statistics, 131 Kerchof Hall, P.O. Box 400135, University of Virginia, Charlottesville, VA 22904-4135, USA.

E-mail: jz9p@virginia.edu

(Received November 2009; accepted February 2010)

COMMENT

Inge S. Helland

University of Oslo

For decades the field of model reduction in regression and in related areas has been dominated by ad hoc proposals and competing methodology. Now, with the present paper and with the related ideas on sufficient model reduction put forward in Cook (1994, 1998) and later papers, it seems that we are entering a new era. We may be approaching a fundamental and general theory on how to improve models in a unique and - perhaps in some sense - optimal way.

Most of the paper by Cook, Li and Chiaromonte is concerned with the multivariate linear regression model (1.1), but as indicated in the Sections 3.2 and 8, this can be extended. The methods discussed are relevant for any model which involves a covariance matrix Σ and regression space $\mathcal{S}$. Many regression models may be seen in this way. One case which has not been explored much so far is that of the Generalized Linear Models with many covariates, but even for ordinary regression models there is much to learn.

In this discussion I concentrate on the population model of Partial Least Squares Regression, which turns out to be closely connected to the envelope model.

Proposition 2.2 in the paper states that the subspace spanned by any set of eigenvectors of $M = \Sigma$ is a reducing subspace of M. In fact, there is an if and only if here. From Proposition 2.2 (1 & 2) it follows that if $\mathcal{R}$ reduces M, then $\mathcal{R}$ is spanned by egenvectors of M. The question is which eigenvectors span the minimal reduced space $\mathcal{E}_M(\mathcal{S})$. The simple answer is given by Proposition 2.3; $\mathcal{E}_M(\mathcal{S}) = \oplus_{i=1}^q P_i\mathcal{S}$, with projections onto eigenspaces.

The fact that $\text{span}(\beta) = \mathcal{S} \subseteq \mathcal{E}_M(\mathcal{S})$ does not indicate that the envelope model is a wider model than the original regression model. It only implies that the new regression vector is of the form $R\eta$ for some η, where R is a matrix spanning $\mathcal{E}_M(\mathcal{S})$, which in the nontrivial case implies a model reduction. In addition, the formula of Proposition 2.3 implies an orthogonality property of the envelope model, so there are arguments indicating improved prediction properties.

Consider now the random x regression model with centered variables

$$y = \beta^T x + e,$$

where y is a scalar and x is a p-vector. Let $M = \Sigma_{xx}$, and note that $\sigma = \sigma_{xy}$ is one-dimensional.

For this model the population version of the chemometricians' Partial Least Squares Regression (PLS) was discussed in Helland (1990), and several characteristica of the regression vector were considered. One has the formula

$$\beta_{m,PLS} = \sum_{k=1}^{m} (\nu_k)^{-1} \eta_k \eta_k^T \sigma,$$

where ν_k is the eigenvalue corresponding to the M-eigenvector η_k, and where m is the unique number of terms in the population PLS-algorithm where the algorithm stops in a natural way. The number of terms in this formula is minimal in the following sense: for single eigenvalues we only require $\eta_k^T\sigma \neq 0$. For multiple eigenvalues we rotate so that a unique η_k has a non-zero component along σ, this is always possible, and this solution is found automatically by the population PLS-algorithm. Note that $\beta_{m,PLS}$, in a very precise sense, lies in a minimal space spanned by eigenvectors of M.

In Helland and Cook (2010) it is proved first that if the Σ_{xx}-envelope has dimension m, then this space is spanned by the m eigenvectors of a population PLS model. Next, the two models give the same predictions. Thus for one-dimensional random x regression, envelope models and population PLS models are identical.

It is useful to recall the general results on the same regression model of Næs and Helland (1993), defining two concepts of relevance in a regression model. First, $z = R^T x$ is called weakly relevant for predicting y if the best linear predictor based on z is the best linear predictor based on x. This turns out to be equivalent to $\text{span}(\beta) \subseteq \text{span}(R)$. Next, z is called strongly relevant for predicting y if in addition we can write $x = Rz + Uv$, with $R^T U = 0$ in such a way that

v is uncorrelated with both z and y. This turns out to be equivalent to weak relevance together with the property that span(R) is spanned by eigenvectors of M. The link to span(R) reducing M is obvious.

There is also a link to PLS (Næs and Helland (1993); Helland and Cook (2010)): assume that $z = R^T x$, with R having r columns, is strongly relevant for predicting y. Then there is a $m \leq r$ such that the PLS model of dimension m holds. The dimension is m if and only if there is a minimal set of eigenvectors with non-zero intersection with σ. Let R_1 be the matrix of these eigenvectors. Then span$(R_1) \subseteq$ span(R). On the other hand, consider a PLS model with m components. Let R_1 be the matrix composed of the m components. Then $z = R_1^T x$ is strongly relevant for predicting y.

From the formula above for $\beta_{m,PLS}$, it is clear that this last R_1 is minimal. Thus $\mathcal{R} = \text{span}(R_1)$ is equal to $\mathcal{E}_M(\mathcal{S})$ with $\mathcal{S} = \text{span}(R)$. In conclusion: taking the random regression model as a point of departure, the the population PLS model is identical to an envelope model, being a minimal reduction of any model constructed according to the strong relevance concept, and to any model being reduced by the span(R) of a strongly relevant model.

Much of the discussion above can be generalized to the multivariate random regression model $y = B^T x + e$, where y and e are vectors, and where chemometricians have proposed non-equivalent sample PLS algorithms. In this case the $M = \Sigma_{xx}$-envelope model is of course unique. The relation to PLS is discussed in Helland and Cook (2010), but may require further investigations. For this model we also have an Σ_{yy}-envelope model. The question is when it is appropriate to do both reductions. This can be done in a meaningful way by first projecting out immaterial dimensions in y to give a new dependent variable Py, and then projecting out immaterial dimension in x for the regression of Py upon x. It is conjecture that estimation from this procedure will beat the multivariate PLS-algorithm with respect to prediction properties.

Many other properties of the PLS-model are formulated in the literature. It is a hope that these can also be used to increase our understanding of the envelope model.

It is known that sample PLS cannot be optimal for prediction in any way (Helland (2001)). A question that has been open until now has been to find some criterion under which population PLS is an optimal model for prediction purposes. Any result in this direction which can be found for the envelope model will automatically also be valid for population PLS.

One very important aspect of Cook, Li and Chiaromonte's paper is that they have developed a feasible maximum likelihood procedure for the envelope model. This is likely to give better regression predictors than any sample estimates proposed for equivalent models, and it makes the idea of an envelope model a very practical notion.

References

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of phyical systems. In *Proceedings of the Section on Physical and Engineering Sciences*, 18-25. American Statistical Association, Alexandria VA.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* Wiley, New York.

Helland, I. S. (1990). Partial least squares regression and statistical models. *Scand. J. Statist.* **17**, 97-114.

Helland, I. S. (2001). Some theoretical aspects of partial least squares regression. *Chemom. Intell. Labor. Systems* **58**, 97-107.

Helland, I. S. and Cook, R. D. (2010). Partial least squares regression from envelope models. Submitted.

Næs, T. and Helland, I.S. (1993). Relevant components in regression. *Scand. J. Statist.* **20**, 239-250.

Department of Mathematics, University of Oslo, Box 1053 Blindern, NO-0316 Oslo, Norway.

E-mail: ingeh@math.uio.no

(Received November 2009; accepted February 2010)

COMMENT

Hung Hung and Su-Yun Huang

National Taiwan University and Academia Sinica

Abstract: Cook, Li and Chiaromonte have introduced the interesting notion of an envelope model, which provides an efficient approach for high-dimensional data analysis. The envelope model looks for the “minimal sufficient response” for the regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Our discussion focuses mainly on a two-staged estimation approach, PLS-MLE, which is workable for $n \ll r + p$. The partial-least-squares approach is used to extract an intermediate response subspace. This subspace is assumed to be big enough to contain the envelope subspace, but also small enough to accommodate a stable MLE.

1. A Two-Staged Alternative Approach for Estimation

1.1. Two-staged MLE

The authors' MLE relies on minimizing the objective function $D = \det(\boldsymbol{H})$ over $\mathrm{span}(\boldsymbol{\Gamma})$, where $\boldsymbol{H} = \boldsymbol{H}(\boldsymbol{\Gamma}) = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T\hat{\boldsymbol{\Sigma}}_{\mathrm{res}}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + (\boldsymbol{I} - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)\hat{\boldsymbol{\Sigma}}_Y(\boldsymbol{I} - \boldsymbol{\Gamma}\boldsymbol{\Gamma}^T)$.

This minimization problem, using Grassmannian optimization, involves the first and second derivatives of D with respect to the envelope parameter $\boldsymbol{\Gamma}$. The first derivative is

$$\frac{\partial D}{\partial \boldsymbol{\Gamma}} = 2D\,\boldsymbol{\Gamma}^T\left(\boldsymbol{M}\boldsymbol{H}^{-1} + \boldsymbol{H}^{-1}\boldsymbol{M}^T - \boldsymbol{H}^{-1}\hat{\boldsymbol{\Sigma}}_Y - \hat{\boldsymbol{\Sigma}}_Y\boldsymbol{H}^{-1}\right), \qquad (1.1)$$

where $\boldsymbol{M} = (\hat{\boldsymbol{\Sigma}}_{\text{res}} + \hat{\boldsymbol{\Sigma}}_Y)\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T$. At each iterative update, (1.1) involves inverting the matrix $\boldsymbol{H}$ at its current envelope parameter value. It then requires $n \gg r + p$, otherwise inverting $\boldsymbol{H}$ can be numerically unstable. Also notice that the computational complexity of inverting an $r \times r$ matrix is $O(r^3)$, which can be time consuming if r is large. We provide a two-staged estimation method, that not only stabilizes but speeds up the numerical computation for large r. Assume there is a known matrix $\boldsymbol{R}_{r\times q}$ (its selection will be discussed later) such that $\text{span}(\boldsymbol{\Gamma}) \subseteq \text{span}(\boldsymbol{R})$ and $u \leq q \ll n$. That is, $\boldsymbol{\Gamma} = \boldsymbol{R}\boldsymbol{\xi}$ for some matrix $\boldsymbol{\xi}_{q\times u}$. Without loss of generality, we assume $\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}_q$, which implies $\boldsymbol{\xi}^T\boldsymbol{\xi} = \boldsymbol{I}_u$. Consider the $\boldsymbol{R}$-induced model

$$\boldsymbol{R}^T\boldsymbol{Y} = \boldsymbol{R}^T\boldsymbol{\alpha} + \boldsymbol{\beta}^*\boldsymbol{X} + \boldsymbol{\varepsilon}^*, \text{ where } \boldsymbol{\beta}^* = \boldsymbol{R}^T\boldsymbol{\beta} = \boldsymbol{\xi}\boldsymbol{\eta} \text{ and } \boldsymbol{\varepsilon}^* \sim \boldsymbol{N}_q(\boldsymbol{0}, \boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R}). \qquad (1.2)$$

We could adopt the authors' MLE to estimate $\boldsymbol{\beta}^*$. The validity of this approach relies on the envelope model structure of model (1.2), which is stated in the following proposition.

Proposition 1. *The subspace* $\text{span}(\boldsymbol{\xi})$ *is the* $\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R}$*-envelope of* $\text{span}(\boldsymbol{\beta}^*)$*.*

Proof. Since $\boldsymbol{\beta}^* = \boldsymbol{\xi}\boldsymbol{\eta}$, it is obvious that $\boldsymbol{\beta}^* \in \text{span}(\boldsymbol{\xi})$. Observe that $(\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R})\boldsymbol{\xi} = \boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \boldsymbol{\xi}\boldsymbol{\Omega} \in \text{span}(\boldsymbol{\xi})$. Together with the symmetry of $\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R}$, we have that $\text{span}(\boldsymbol{\xi})$ is a reducing subspace of $\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R}$. The desired property can be established by showing the minimality of $\text{span}(\boldsymbol{\xi})$ among all $\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R}$-reducing subspaces that contain $\boldsymbol{\beta}^*$. If $\text{span}(\boldsymbol{\xi})$ is not minimal, there must exist a $\boldsymbol{\xi}^*$ such that (a) $\text{span}(\boldsymbol{\beta}^*) \subseteq \text{span}(\boldsymbol{\xi}^*) \subsetneq \text{span}(\boldsymbol{\xi})$, and (b) $\text{span}(\boldsymbol{\xi}^*)$ reduces $\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R}$. From (a), we have $\boldsymbol{\beta}^* = \boldsymbol{\xi}^*\boldsymbol{\eta}^*$ and $\boldsymbol{\xi}^* = \boldsymbol{\xi}\boldsymbol{c}$ for some $\boldsymbol{\eta}^*$ and $\boldsymbol{c}$ and $\text{rank}(\boldsymbol{c}) < u$, since $\text{span}(\boldsymbol{\xi}^*)$ is a proper subset of $\text{span}(\boldsymbol{\xi})$. From (b), we have $\boldsymbol{R}^T\boldsymbol{\Sigma}\boldsymbol{R} = \boldsymbol{\xi}^*\boldsymbol{A}\boldsymbol{\xi}^{*T} + \boldsymbol{\xi}_0^*\boldsymbol{A}_0\boldsymbol{\xi}_0^{*T}$ for some $\boldsymbol{A}$ and $\boldsymbol{A}_0$, where $\boldsymbol{\xi}_0^*$ is the orthogonal complement of $\boldsymbol{\xi}^*$. Take $\boldsymbol{\Gamma}^* = \boldsymbol{R}\boldsymbol{\xi}^*$. It is easy to see that $\text{span}(\boldsymbol{\Gamma}^*) \subsetneq \text{span}(\boldsymbol{\Gamma})$ since $\text{span}(\boldsymbol{\xi}^*)$ is a proper subset of $\text{span}(\boldsymbol{\xi})$. Also, $\boldsymbol{\beta} = \boldsymbol{\Gamma}\boldsymbol{\eta} = R\boldsymbol{\beta}^* = \boldsymbol{\Gamma}^*\boldsymbol{\eta}^* \in \text{span}(\boldsymbol{\Gamma}^*)$. Finally we show that $\boldsymbol{\Gamma}^*$ reduces $\boldsymbol{\Sigma}$. Let $\boldsymbol{R}_0$ be the orthogonal complement of $\boldsymbol{R}$. Observe that

$$\begin{aligned}
\boldsymbol{\Sigma}\boldsymbol{\Gamma}^* &= (\boldsymbol{R}\boldsymbol{R}^T + \boldsymbol{R}_0\boldsymbol{R}_0^T)\boldsymbol{\Sigma}\boldsymbol{R}\boldsymbol{\xi}^* \\
&= \boldsymbol{R}(\boldsymbol{\xi}^*\boldsymbol{A}\boldsymbol{\xi}^{*T} + \boldsymbol{\xi}_0^*\boldsymbol{A}_0\boldsymbol{\xi}_0^{*T})\boldsymbol{\xi}^* + \boldsymbol{R}_0\boldsymbol{R}_0^T(\boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^T)\boldsymbol{\Gamma}\boldsymbol{c} \\
&= \boldsymbol{R}\boldsymbol{\xi}^*\boldsymbol{A} + \boldsymbol{R}_0\boldsymbol{R}_0^T(\boldsymbol{R}\boldsymbol{\xi})\boldsymbol{\Omega}\boldsymbol{c} = \boldsymbol{\Gamma}^*\boldsymbol{A} \in \text{span}(\boldsymbol{\Gamma}^*).
\end{aligned}$$

By the symmetry of $\boldsymbol{\Sigma}$, $\boldsymbol{\Gamma}^*$ reduces $\boldsymbol{\Sigma}$. Then, span($\boldsymbol{\Gamma}^*$) is the $\boldsymbol{\Sigma}$-envelope for $\mathcal{B}$, which contradicts the assumption that span($\boldsymbol{\Gamma}$) is the $\boldsymbol{\Sigma}$-envelope. Therefore, span($\boldsymbol{\xi}$) must be the minimal reducing subspace that contains $\boldsymbol{\beta}^*$.

The estimate of $\boldsymbol{\beta}$ can be easily recovered from $\boldsymbol{R}\hat{\boldsymbol{\beta}}^*_{\text{em}}$ through the fact that $\boldsymbol{\beta} = \boldsymbol{R}\boldsymbol{\beta}^*$. Alternatively, since $\boldsymbol{\Gamma} = \boldsymbol{R}\boldsymbol{\xi}$, $\boldsymbol{\Gamma}$ can be estimated by $\boldsymbol{R}\hat{\boldsymbol{\xi}}$ with $\hat{\boldsymbol{\xi}}$ being the MLE of $\boldsymbol{\xi}$ under model (1.2). Paralleling the authors' (4.6), $\boldsymbol{\beta}$ can also be estimated by $\boldsymbol{P}_{\boldsymbol{R}\hat{\boldsymbol{\xi}}}\hat{\boldsymbol{\beta}}_{\text{fm}}$. It is easy to show that these two perspectives provide the same estimate, i.e., $\boldsymbol{R}\hat{\boldsymbol{\beta}}^*_{\text{em}} = \boldsymbol{P}_{\boldsymbol{R}\hat{\boldsymbol{\xi}}}\hat{\boldsymbol{\beta}}_{\text{fm}}$, which further justifies the use of an $\boldsymbol{R}$-induced model, assuming that we have the knowledge of an $\boldsymbol{R}$ to encapsulate $\boldsymbol{\Gamma}$. Notice that $\boldsymbol{R}$ is not necessarily a reducing subspace of $\boldsymbol{\Sigma}$. One only requires the column span of $\boldsymbol{R}$ to be large enough to contain the envelope subspace span($\boldsymbol{\Gamma}$), but small enough so that $q + p \ll n$, to accommodate a stable MLE.

1.2. PLS-aided selection of $\boldsymbol{R}$ and $\boldsymbol{q}$

Suppose q is fixed. We choose $\boldsymbol{R}$ by applying the PLS approach. PLS seeks a lower dimensional transformation of $\boldsymbol{Y}$ (denoted by $\boldsymbol{W}_q^T\boldsymbol{Y}$ in the following algorithm) that contains most of the association information between $\boldsymbol{Y}$ and $\boldsymbol{X}$. This motivates us to choose $\boldsymbol{R} = \boldsymbol{W}_q$, or $\boldsymbol{R} = \boldsymbol{W}_q(\boldsymbol{W}_q^T\boldsymbol{W}_q)^{-1/2}$, if $\boldsymbol{W}_q$ is not already orthonormal. There are many PLS algorithms, we used SIMPLS algorithm (de Jong (1993)), which is briefly summarized below. Let $\boldsymbol{S}_0 = \boldsymbol{U}^T\boldsymbol{F}$, with $\boldsymbol{U}$ and $\boldsymbol{F}$ being the centered data matrices of $\boldsymbol{Y}$ and $\boldsymbol{X}$, respectively. At the i^{th} update, $i = 1, \cdots, q$,

- w_i is taken to be the left singular vector of $\boldsymbol{S}_{i-1}$ with the largest singular value, denoted by λ_i. Let $t_i = \boldsymbol{U}w_i$ and $p_i = \boldsymbol{U}^Tt_i/(t_i^Tt_i)$.
- $\boldsymbol{W}_i = [w_1, \cdots, w_i]$, $\boldsymbol{P}_i = [p_1, \cdots, p_i]$, and $\boldsymbol{S}_i = \left(\boldsymbol{I} - \boldsymbol{P}_i(\boldsymbol{P}_i^T\boldsymbol{P}_i)^{-1}\boldsymbol{P}_i^T\right)\boldsymbol{S}_{i-1}$.

We denote the two-staged MLE using PLS to select $\boldsymbol{R}$ as PLS-MLE. As to the selection of q, we use a simple criterion. Observe that in the i^{th} iteration, the left singular vector of $\boldsymbol{S}_{i-1}$ with the largest singular value is included in $\boldsymbol{W}_i$. We expect the singular value λ_{i+1} in next iteration to be relatively small if most of the association information between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is already explained by $\boldsymbol{W}_i^T\boldsymbol{Y}$. We thus proceed with the algorithm until $d_{q+1} \triangleq \lambda_{q+1}/\sum_{i=1}^{q+1}\lambda_i$ is smaller than a pre-specified threshold value δ.

Example 1. (Comparing two normal means) We take the same simulation setting as in the authors' Figure 1a (with the modification $\boldsymbol{\beta} = 1.4736\sqrt{2\ln r}\,\boldsymbol{1}_r$) for $n = 40$, 80, and $r = 10$, 80. When $r = 80$, MLE is infeasible and only PLS-MLE is reported. The threshold value is set to $\delta = 0.008$ for data-driven q selection. The results depicted in Figure 1 are based on 200 replicate runs. It is found

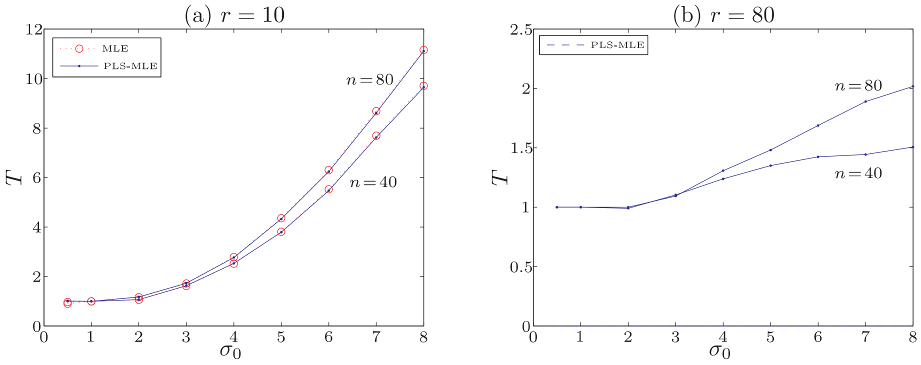


Figure 1. Relative efficiency at different values σ_0.

that PLS-MLE and MLE have indistinguishable behavior at $r = 10$ for all cases. When $r = 80$, PLS-MLE can still gain efficiency for $\sigma_0 \geq 2$, even if the gain is not as great as for the case $r = 10$.

2. An Application to Linear SVMs

Assume the envelope model in the authors' (3.2) with the explanatory variable vector $\boldsymbol{X} \in \mathbb{R}^k$, where $\boldsymbol{X}$ is an indicator coding for class labels. To indicate the c^{th} class, $\boldsymbol{X}$ is set to $(0, \ldots, 1, 0, \ldots)^T$ with one in the c^{th} place and zeros elsewhere. The PLS-MLE approach is used to estimate $\boldsymbol{\Gamma}$ by $\boldsymbol{R}\hat{\boldsymbol{\xi}}$, as described at the end of Section 1.1. By plugging in its estimate, $\boldsymbol{\Gamma}$ is regarded as known in the following discussion for simplicity, but do keep in mind that $\boldsymbol{\Gamma}$ is estimated by the PLS-MLE. The SVM classification is then trained in the envelope subspace spanned by $\boldsymbol{\Gamma}$. A new test instance can be first projected to the envelope subspace span($\boldsymbol{\Gamma}$), and then classified according to the trained SVM model in the envelope subspace.

Example 2. [Medline data] The Medline data set is a document-term matrix. Each row, which represents a document, consists of term frequencies for 22,095 distinct terms. This data set has 2,500 documents uniformly over 5 classes, and each document belongs to one class. The set was equally divided into 1,250 training documents and 1,250 test documents. The document-term data set is sparse and has many zero term frequencies. We remove those columns with zero variance in the training set and this results in 15,109 terms. We then adopt the envelope model at authors' (3.2) with $r = 15,109$ and $p = 5$. Three different SVM variants were considered:

- Standard SVM, i.e., without enveloping;

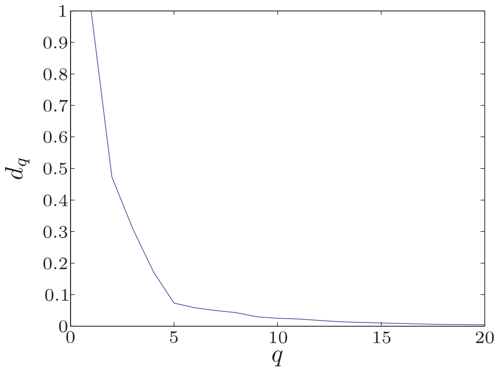


Figure 2. The value of d_q for Medline data.

- SVM on the PLS-extracted subspace;
- SVM on the PLS-MLE envelope subspace with $u = 4$ (the number of classes minus one).

There are many SVM algorithms, we used the smooth SVM of Lee and Mangasarian (2001). As this is a multi-class problem, a series of small binary SVM classifiers was trained (known as one-against-one) and a test instance was classified by combining these binary SVMs by majority vote (see Schölkopf and Smola (2002)). Figure 2 plots the value of d_q versus q that can guide us in selecting q. The d_q value descends quickly and stays below 0.01 for $q > 15$. An empirical value $\delta = 0.01$ then suggests a q around 15. The analysis results for $6 \leq q \leq 20$ are provided in Figures 3$-$4. Notice that, for $q \leq 15$, the SVM in the PLS subspace has a bit higher accuracy than the standard SVM, and a further reduction by MLE leads to a further small accuracy gain (it attains the maximum value 0.9024 at $q = 7$). However, when the PLS-size gets larger than or equal to 16, the MLE starts to behave unstably, and the quality depends on the random initial. This reveals the usefulness of a PLS-aided MLE with a moderate q. We also found that the classification accuracies for the SVM based merely on PLS were roughly the same for all q. Without further enveloping by MLE, however, we need more PLS components ($q = 13$) to achieve accuracy 0.9024. The accuracy gain of the reduced model (by PLS or by PLS-MLE) over the full model was about 1%-1.4%, which is probably not much (see Figure 3). However, the reduced model has a lot of reduction in memory space and computing time complexity (see Figure 4).

3. Concluding Remarks

Enveloping is a general idea of parsimonious modeling for efficiency gain. The $\mathbf{\Sigma}$-envelope for multivariate linear models relies on the notion of $\mathbf{\Sigma}$-invariance,

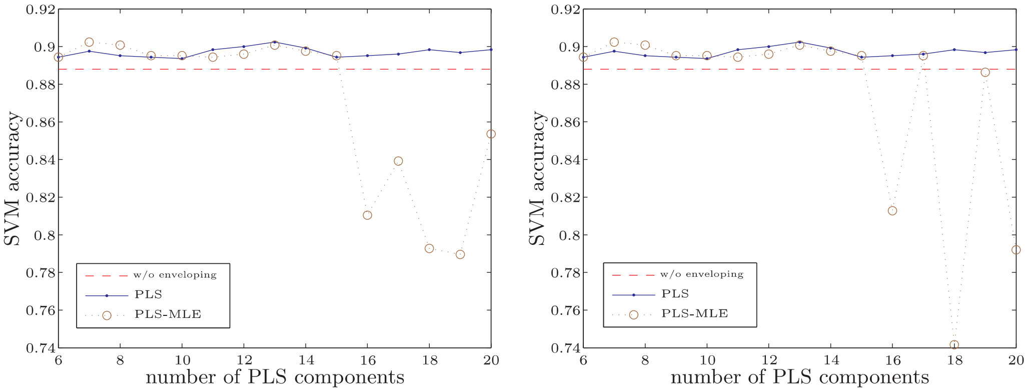


Figure 3. Accuracy comparison for Medline data, results of 2 replicate runs.

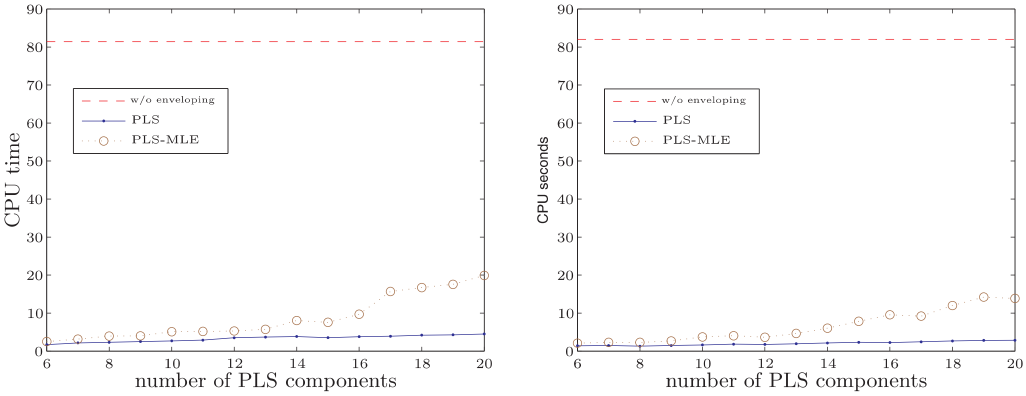


Figure 4. CPU time comparison for Medline data, results of 2 replicate runs.

defined by $\mathbf{\Sigma}\mathcal{S} \subseteq \mathcal{S}$ for symmetric $\mathbf{\Sigma}$ acting on a subspace $\mathcal{S}$. Different learning tasks and learning algorithms may have their own route to make inference. Therefore, we may consider different notions of invariance natural to specific learning methods/algorithms in future study, as well as the efficiency gain in memory space and computing time complexity, seen in the Medline example, rather than accuracy improvement.

References

de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems* **18**, 251-263.

Lee, Y.-J. and Mangasarian, O. L. (2001). SSVM: a smooth support vector machine for classification. *Computational Optimization and Applications* **20**, 5-21.

Schölkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* MIT Press.

Graduate Institute of Epidemiology, College of Public Health, National Taiwan University, 17, XuZhou Road, Taipei City 100, Taiwan.

E-mail: hhung@ntu.edu.tw

Institute of Statistical Science, Academia Sinica Taipei 11529, Taiwan.

E-mail: syhuang@stat.sinica.edu.tw

(Received January 2010; accepted February 2010)

COMMENT

Xuerong Meggie Wen

Missouri University of Science and Technology

In the context of the classical multivariate linear regression model, Cook, Li and Chiaromonte proposed a parsimonious parameterization using the novel concept of an "*envelope*", which yields an asymptotically less variable MLE comparing to the traditional multivariate linear regression estimator. One question that arises naturally is how this method can be generalized to deal with nonlinear or (and) nonnormal multivariate regression models. In this discussion, we explore the applicability of the *envelope* method when the mean function is nonlinear. Specifically, we assume the multivariate nonlinear regression model

$$Y^j = f_j(\alpha_j + \boldsymbol{\beta}_j \mathbf{X}) + \epsilon_j, \tag{1}$$

where $\mathbf{Y} = (Y^1, \ldots, Y^r)$ is the random response vector, $\mathbf{X}$ is a p-dimensional random vector of predictors, $\epsilon = (\epsilon_1, \ldots, \epsilon_r)$ is independent with $\mathbf{X}$ and is normally distributed with mean 0 and unknown covariance matrix $\mathbf{\Sigma}$, $f_j(.)$, $j = 1, \ldots, r$, are arbitrary unknown link functions. Both $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)$ and $\boldsymbol{\beta} \in \mathbb{R}^{r \times p}$ are unknown, while $\boldsymbol{\beta}_j$, $j = 1, \ldots, r$, is the jth row of $\boldsymbol{\beta}$. This is the same model that Cook and Setodji (2003) considered in the context of sufficient dimension reduction. Our goal is to estimate span$(\boldsymbol{\beta})$, referred to as the *multivariate central mean subspace* in sufficient dimension reduction literature (Cook and Setodji 2003).

Assuming that $\mathbf{X}$ satisfies the so-called linearity condition (Li and Duan 1989), $\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}\mathbf{X})$ is a linear function of $\boldsymbol{\beta}\mathbf{X}$, commonly used in sufficient dimension reduction methods and one which holds for elliptically contoured predictors

(Eaton (1986)), hence holds when $\mathbf{X}$ is multivariate normal, the following Lemma suggests that the *envelope* method that Cook et. al. proposed still holds under link violation. The justification of our Lemma is similar to that of Proposition 8.1 of Cook (1998) which stems from Theorem 2.1 of Li and Duan (1989).

Lemma 1. *If* (1) *holds and* $\mathrm{E}(\mathbf{X}|\boldsymbol{\beta}\mathbf{X})$ *is a linear function of* $\boldsymbol{\beta}\mathbf{X}$*, then* $\hat{\boldsymbol{\beta}}$*, the maximum likelihood estimator obtained via envelope method under the misspecified linear link function, is a Fisher consistent estimator of* $\boldsymbol{\beta}$ *up to a multiplicative scalar.*

We may also compare the envelope estimator with those obtained via sufficient dimension reduction methods assuming (1). We conjecture that it would perform better than those sufficient dimension reduction methods which are currently available for multivariate responses due to the property of MLE. As to the restriction of the Gaussian distribution for the multivariate linear model considered in Cook et al., we may be able to extend the envelope method to the natural exponential family. Further research along this line is underway.

References

Cook, R. D. (1998). *Regression Graphics.* Wiley, New York.

Cook, R. D. and Setodji, C. M. (2003). A model-free test for reduced rank in multivariate regression. *J. Amer. Statist. Assoc.* **98**, 340-351.

Eaton, M. L. (1986). A characterization of spherical distributions. *J. Multivariate Anal.* **20**, 272-276.

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009-1052.

Department of Mathematics and Statistics, Missouri University of Science and Technology, MO 65409, U.S.A.

E-mail: wenx@mst.edu

(Received February 2010; accepted February 2010)

COMMENT

Zhou Yu and Lixing Zhu

East China Normal University and Hong Kong Baptist University

We congratulate the authors for their path breaking work that provides us new insight into multivariate linear regression. By introducing the envelope models, the authors open up an avenue toward efficient redundancy reduction. Our discussion here will focus on three issues: (1) intuition of efficiency gain; (2) second order bias; (3) the envelope model with sparse structure.

1. Intuition of Efficiency Gain

The authors provide impressively strong supporting asymptotics that evidences the significant benefits of the MLE from the envelope models. When $\mathbf{\Sigma}_1$ is assumed to be given, we can understand the theoretical conclusion in a more direct and simple way. With $\mathbf{\Sigma}_1$, $\hat{\boldsymbol{\beta}}_{\text{em}} = \mathbf{P}_{\mathbf{\Sigma}_1}\hat{\boldsymbol{\beta}}_{\text{fm}}$ leads to $\text{var}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\text{em}})] = \mathbf{\Sigma}_X^{-1} \otimes (\mathbf{P}_{\mathbf{\Sigma}_1}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Sigma}_1})$. Noting that $\mathbf{P}_{\mathbf{\Sigma}_1} = \mathbf{\Sigma}_1^{\dagger}\mathbf{\Sigma}_1 = \mathbf{\Gamma}\mathbf{\Omega}^{1/2}(\mathbf{\Omega}^{1/2}\mathbf{\Gamma}^T\mathbf{\Gamma}\mathbf{\Omega}^{1/2})^{-1}\mathbf{\Omega}^{1/2}\mathbf{\Gamma}^T = \mathbf{\Gamma}\mathbf{\Gamma}^T = \text{P}_{\mathbf{\Gamma}}$, and hence $\mathbf{P}_{\mathbf{\Sigma}_1}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Sigma}_1} = \text{P}_{\mathbf{\Gamma}}\mathbf{\Sigma}\text{P}_{\mathbf{\Gamma}} = \mathbf{\Sigma} - \text{Q}_{\mathbf{\Gamma}}\mathbf{\Sigma}\text{Q}_{\mathbf{\Gamma}}$, it is then obvious that $\text{var}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\text{em}})] \leq \mathbf{\Sigma}_X^{-1} \otimes \mathbf{\Sigma} = \text{var}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\text{fm}})]$. When $\mathbf{\Sigma}_1$ needs to be estimated, it is not straightforward to arrive at the conclusion $\text{var}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\text{em}})] \leq \text{var}[\sqrt{n}\text{vec}(\hat{\boldsymbol{\beta}}_{\text{fm}})]$ because the estimator $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}$ for $\mathbf{P}_{\mathbf{\Sigma}_1}$ plugged in $\hat{\boldsymbol{\beta}}_{\text{em}}$ would contribute to the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\text{em}}$. However, with theoretical support as in **Theorem 5.1**, the effect of variance inflation brought by $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}$should be no more than the oracle efficiency gain $\mathbf{\Sigma}_X^{-1} \otimes (\boldsymbol{Q}_{\mathbf{\Gamma}}\mathbf{\Sigma}\boldsymbol{Q}_{\mathbf{\Gamma}})$.

2. Second Order Bias

Although $\hat{\boldsymbol{\beta}}_{\text{em}}$ is attractive with smaller variance asymptotically than $\hat{\boldsymbol{\beta}}_{\text{fm}}$, $\hat{\boldsymbol{\beta}}_{\text{em}}$ is not an unbiased estimator as $\hat{\boldsymbol{\beta}}_{\text{fm}}$ is when the predictors vector $\mathbf{X}$ is non-random. $\hat{\boldsymbol{\beta}}_{\text{em}}$ and $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}$ should admit the following asymptotic expansions respectively: $\hat{\boldsymbol{\beta}}_{\text{fm}} = \boldsymbol{\beta} + \boldsymbol{\beta}_{\text{fm}}^*$ and $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1} = +\mathbf{P}_{\mathbf{\Sigma}_1}^* + \mathbf{P}_{\mathbf{\Sigma}_1}^{**} + O_p(n^{-3/2})$, where $\boldsymbol{\beta}_{\text{fm}}^* = O_p(n^{-1/2})$, $\mathbf{P}_{\mathbf{\Sigma}_1}^* = O_p(n^{-1/2})$, and $\mathbf{P}_{\mathbf{\Sigma}_1}^{**} = O_p(n^{-1/2})$. Moreover, $E(\boldsymbol{\beta}_{\text{fm}}^*) = 0$ and $E(\mathbf{P}_{\mathbf{\Sigma}_1}^*) = 0$. Then $E(\hat{\boldsymbol{\beta}}_{\text{em}}) = E(\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}\hat{\boldsymbol{\beta}}_{\text{fm}}) = \mathbf{P}_{\mathbf{\Sigma}_1}\boldsymbol{\beta} + [E(\mathbf{P}_{\mathbf{\Sigma}_1}^{**}\boldsymbol{\beta}) + E(\mathbf{P}_{\mathbf{\Sigma}_1}^*\boldsymbol{\beta}_{\text{fm}}^*)] + O(n^{-3/2})$. The $O(n^{-1})$ bias $\boldsymbol{\delta} = E(\mathbf{P}_{\mathbf{\Sigma}_1}^{**}\boldsymbol{\beta}) + E(\mathbf{P}_{\mathbf{\Sigma}_1}^*\boldsymbol{\beta}_{\text{fm}}^*)$ may be non-negligible. We conducted a small simulation with the same set-up as in Section 7.1 to compare the biases of $\hat{\boldsymbol{\beta}}_{\text{fm}}$ and $\hat{\boldsymbol{\beta}}_{\text{em}}$. Let $\hat{\boldsymbol{\beta}}_{\text{em}}^i$'s be the sample estimators obtained based on 200 replications, $i = 1, \cdots, 200$. Then one can estimate the square bias of $\hat{\boldsymbol{\beta}}_{\text{em}}$ simply as $\text{BIAS}^2(\hat{\boldsymbol{\beta}}_{\text{em}}) = \|(1/200)\sum_{i=1}^{200}\hat{\boldsymbol{\beta}}_{\text{em}}^i - \boldsymbol{\beta}\|^2$, where $\|.\|$ is the Euclidean matrix norm, see Li, Zha and Chiaromonte (2005). From the simulation results reported in Table 1, it is clear that $\hat{\boldsymbol{\beta}}_{\text{fm}}$ is less biased than $\hat{\boldsymbol{\beta}}_{\text{em}}$. In general, the difference between the biases of $\hat{\boldsymbol{\beta}}_{\text{fm}}$ and $\hat{\boldsymbol{\beta}}_{\text{em}}$ gets smaller as the sample size n increases, which justifies the large sample property.

Two standard approaches are proven to be successful in removing the leading bias. Let $\hat{\boldsymbol{\delta}}$ be a consistent sample estimator of $\boldsymbol{\delta}$. A straightforward bias corrected estimator for the envelope model is given by $\hat{\boldsymbol{\beta}}_{\text{BC-em}} = \hat{\boldsymbol{\beta}}_{\text{em}} - \hat{\boldsymbol{\delta}}$, which is unbiased to order $O(n^{-1})$. Moreover, the mean squared error of $\hat{\boldsymbol{\beta}}_{\text{BC-em}}$ is equal to that of $\hat{\boldsymbol{\beta}}_{\text{em}}$ to order $O(n^{-2})$. Another way to remove the leading bias term from $\hat{\boldsymbol{\beta}}_{\text{em}}$ is through a suitable modification of the score equation, see Firth

Table 1. Comparison of biases ($\times 10^{-3}$). $\boldsymbol{\Omega}_0 = \sigma_0^2 \boldsymbol{I}_9$ and $\sigma = 1$.

		$n = 20$	$n = 30$	$n = 50$	$n = 100$
$\sigma_0 = 0.5$	$\mathrm{BIAS}^2(\hat{\boldsymbol{\beta}}_{\mathrm{em}})$	1.6002	0.9890	0.4866	0.2106
	$\mathrm{BIAS}^2(\hat{\boldsymbol{\beta}}_{\mathrm{fm}})$	1.5562	0.9140	0.4739	0.1986
$\sigma_0 = 1$	$\mathrm{BIAS}^2(\hat{\boldsymbol{\beta}}_{\mathrm{em}})$	4.6689	3.7902	0.9210	0.8217
	$\mathrm{BIAS}^2(\hat{\boldsymbol{\beta}}_{\mathrm{fm}})$	4.4905	3.3266	0.8920	0.8055
$\sigma_0 = 2$	$\mathrm{BIAS}^2(\hat{\boldsymbol{\beta}}_{\mathrm{em}})$	10.661	7.0053	3.1738	1.1381
	$\mathrm{BIAS}^2(\hat{\boldsymbol{\beta}}_{\mathrm{fm}})$	8.0645	6.8613	3.1003	1.0332

(1993). It is well known that the bias corrected MLE is second order efficient with respect to mean squared error criterion, see Efron (1975).

3. Envelope Model with Sparse Structure

Sparse penalized approaches for variable selection have generated much interest, including least absolute shrinkage and selection operator (LASSO; Tibshirani (1996)), smoothing clipped absolute deviation (SCAD; Fan and Li (2001)) estimator, Dantzig selector (Candés and Tao (2007)). We believe that these promising variable selection methods can be applied to the envelope model, particularly when $\boldsymbol{\beta}$ is sparse in the sense that many of its elements are zero. We take the Dantzig selector as an example to illustrate how its idea can be transplanted into sparse envelope model. The Dantzig selector is designed to strike a balance between nearly solving the score function of full linear model and minimizing the ℓ_1 norm of regression coefficients. Let $\mathbf{U}_{\boldsymbol{i}}$ and $\boldsymbol{\beta}_i^T$ be the i-th column of $\mathbf{U}$ and $\boldsymbol{\beta}^T$, respectively, for $i = 1, \cdots, r$. The Dantzig selector solves $\boldsymbol{\beta}_i^T$ by

$$\min \|\boldsymbol{\beta}_i^T\|_{\ell_1}, \quad \text{subject to} \|\mathbf{F}^{\boldsymbol{T}}\mathbf{F}\boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{T}} - \mathbf{F}^{\boldsymbol{T}}\mathbf{U}_{\boldsymbol{i}}\|_{\ell_\infty} \leq \lambda_i^{\mathrm{fm}}, \quad i = 1, \cdots, r,$$

where λ_i^{fm} is a tuning parameter that gives certain relaxation of the score equation. Note that when $\boldsymbol{\Sigma}_1$ is known, the score equation to solve $\boldsymbol{\beta}_{\mathrm{em}}$ is $\mathbf{F}^{\boldsymbol{T}}\mathbf{F}\boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{T}} - \mathbf{F}^{\boldsymbol{T}}\mathbf{U}_{\boldsymbol{i}}\mathbf{P}_{\boldsymbol{\Sigma}_{\mathbf{1}}} = 0$. Incorporating the Dantzig selector, the algorithm for estimating parameters in sparse envelope models is modified a little as follows.

- A. Estimate $\boldsymbol{\beta}$ by the Dantzig selector $\hat{\boldsymbol{\beta}}_{\mathrm{fm}}^{\mathrm{DS}}$. Obtain the residual covariance matrix as $\boldsymbol{\Sigma}_{\mathrm{res}} = n^{-1}\sum_{i=1}^{n}(\mathbf{Y}_i - \hat{\boldsymbol{\beta}}_{\mathrm{fm}}^{\mathrm{DS}}\mathbf{X}_i)(\mathbf{Y}_i - \hat{\boldsymbol{\beta}}_{\mathrm{fm}}^{\mathrm{DS}}\mathbf{X}_i)^T$.

- B. Do the same Step b of the original algorithm described in 4.3.

- C. Estimate $\boldsymbol{\beta}$ by the Dantzig selector with $\mathbf{X}$ as the predictor and $\widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_1}\mathbf{Y}$ as the predictor :

$$\min \|\boldsymbol{\beta}_i^T\|_{\ell_1}, \quad \text{subject to} \|\mathbf{F}^{\boldsymbol{T}}\mathbf{F}\boldsymbol{\beta}_{\boldsymbol{i}}^{\boldsymbol{T}} - \mathbf{F}^{\boldsymbol{T}}\mathbf{U}_{\boldsymbol{i}}\widehat{\mathbf{P}}_{\boldsymbol{\Sigma}_{\mathbf{1}}}\|_{\ell_\infty} \leq \lambda_i^{\mathrm{em}}, \quad i = 1, \cdots, r.$$

The solution for i-th direction is denoted by $(\hat{\boldsymbol{\beta}}_{\mathrm{em}}^{\mathrm{DS}})_i^T$.

Table 2. Simulation comparisons of $\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}}$ and $\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}}$.

	Model	$i=1$	$i=2$	$i=3$	$i=4$
MSE	$(\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}})_i^T$	0.4749	0.4545	0.4510	0.4611
	$(\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}})_i^T$	0.8868	0.8473	0.8722	0.8666
TPR	$(\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}})_i^T$	1.0000	1.0000	1.0000	1.0000
	$(\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}})_i^T$	1.0000	1.0000	1.0000	1.0000
FPR	$(\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}})_i^T$	0.0732	0.0747	0.0732	0.0724
	$(\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}})_i^T$	0.0812	0.0815	0.0808	0.0827

- D. Do the same as Step d of the original algorithm described in 4.3.

The tuning parameters λ_i^{fm} and λ_i^{em} can be selected by minimizing the BIC criterion proposed in Dicker and Lin (2009). Under the assumptions of the envelope model, we believe that $\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}}$ is more efficient than $\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}}$. A simulation study is presented to verify our conjecture. The results reported here were based on 200 replications from simulation models with $n = 100$, $p = 110$, $r = 4$, and $u = 2$. The first ten elements of $\boldsymbol{\beta}_1^T$ were 1, and of $\boldsymbol{\beta}_2^T$, -1. The eleventh to twentieth elements of $\boldsymbol{\beta}_3^T$ were -1, and of $\boldsymbol{\beta}_4^T$, 1. All the other elements of $\boldsymbol{\beta}$ were 0. We took

$$\boldsymbol{\Gamma}^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}.$$

Let $\sigma^2 = 1$ and $\sigma_0^2 = 0.5$. The variance of the error term was constructed as $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{\Gamma}\boldsymbol{\Gamma}^T + \sigma_0^2\boldsymbol{\Gamma}_0\boldsymbol{\Gamma}_0^T$. We employed mean squared error, the true positive rate (TPR), and the false positive rate (FPR) as the performance measures to compare $\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}}$ and $\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}}$ for four directions. From Table 2, we can see that the MSE of $\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}}$ was much less than that of $\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}}$. Moreover, while achieving the same efficiency in selecting active predictors as $\hat{\boldsymbol{\beta}}_{\text{fm}}^{\text{DS}}$, $\hat{\boldsymbol{\beta}}_{\text{em}}^{\text{DS}}$ was more efficient in screening out inactive predictors.

The aforementioned approach to incorporate the Dantzig selector is useful when $\boldsymbol{\beta}$ is sparse. For the envelope model, it would be more meaningful to consider the case when $\boldsymbol{\Gamma}$ is sparse or, more precisely, when $\mathbf{P}_{\boldsymbol{\Sigma}_1}$ is sparse. In the previous simulation study,

$$\mathbf{P}_{\boldsymbol{\Sigma}_1} = \begin{pmatrix} 0.5 & -0.5 & 0 & 0 \\ -0.5 & 0.5 & 0 & 0 \\ 0 & 0 & 0.5 & -0.5 \\ 0 & 0 & -0.5 & 0.5 \end{pmatrix}.$$

However, a typical sample estimator for $\mathbf{P}_{\mathbf{\Sigma}_1}$ was:

$$\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1} = \begin{pmatrix} 0.4603 & -0.4982 & 0.0142 & -0.0032 \\ -0.4982 & 0.5399 & 0.0023 & -0.0139 \\ 0.0142 & 0.0023 & 0.5049 & -0.4998 \\ -0.0032 & -0.0139 & -0.4998 & 0.4949 \end{pmatrix}.$$

If the small coefficients, such as 0.0142 and -0.0032, can be shrunk to zero, we believe the efficiency of MLE from the envelope model can be improved. A possible LASSO type approach to obtain a regularized version of $\mathbf{P}_{\mathbf{\Sigma}_1}$ is by minimizing the following objective function over the Grassmann manifold $\mathbb{G}^{r\times u}$:

$$\log\det(\mathbf{\Gamma}^T\hat{\mathbf{\Sigma}}_{\mathrm{res}}\mathbf{\Gamma}) + \log\det(\mathbf{\Gamma}_0\hat{\mathbf{\Sigma}}_{\mathbf{Y}}\mathbf{\Gamma}_0^T) + \lambda\sum_{i=1}^{r}\sum_{j=1}^{r}|(\mathbf{\Gamma}\mathbf{\Gamma}^T)_{ij}|,$$

where λ is a tuning parameter. However, it is challenging to combine available algorithms for LASSO and the Stiefel Grassmann optimization algorithm to solve such a problem. How to develop a feasible method both theoretically and practically to regularize $\mathbf{P}_{\mathbf{\Sigma}_1}$ deserves further study. In general sparse cases, an interesting extension is simultaneous regularization of $\hat{\boldsymbol{\beta}}$ and $\widehat{\mathbf{P}}_{\mathbf{\Sigma}_1}$.

Lastly, we thank the authors again for their clear and imaginative work that will stimulate plenty of future research, such as in quasi-likelihood and sufficient dimension reduction, especially with multivariate responses.

Acknowledgement

The research was supported by a RGC grant from the Research Grants Council of Hong Kong, Hong Kong, China, and a NSF grant from National Natural Science Foundation of China (No. 10701035).

References

Candés, E. J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35**, 2313-2351.

Dicker, L. and Lin, X. (2009). A large sample analysis of the Dantzig selector and extensions. Unpublished manuscript.

Efron, B. (1975). Defining the curvature of a statistical problem (with application second-order efficiency) (with discussion). *Ann. Statist.* **3**, 1189-1242.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96**, 1348-1360.

Firth, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27-38.

Li, B., Zha, H. and Chiaromonte, C. (2005). Contour regression: a general approach to dimension reduction. *Ann. Statist.* **33**, 1580-1616.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

School of Finance and Statistics, East China Normal University, Shanghai 200062, China.
E-mail: yz19830224@gmail.com

Department of Mathematics, The Hong Kong Baptist University, Kowloon Tong, Hong Kong.
E-mail: lzhu@hkbu.edu.hk

(Received January 2010; accepted February 2010)

COMMENT

Yuexiao Dong and Li-Ping Zhu

Temple University and East China Normal University

The authors are to be congratulated on this groundbreaking work. Envelope models are introduced to control over-parameterization in the classical multivariate normal linear model. Substantial improvement by envelope models is shown in terms of estimating efficiency over a wide range of applications. It is hard to overstate the importance of multivariate linear regression as the cornerstone of the multivariate analysis, and we have no doubt about the significant impact this paper will have.

To get some insights into envelope models, we assume that $X \in \mathbb{R}^1$, $\mathbf{Y} \in \mathbb{R}^r$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^r$ is normally distributed with mean $\mathbf{0}$ and unknown covariance $\boldsymbol{\Sigma} \geq 0$. The asymptotic variance of the usual maximum likelihood estimate of $\boldsymbol{\beta}$ in the multivariate normal linear model

$$\mathbf{Y} = \boldsymbol{\alpha} + \boldsymbol{\beta} X + \boldsymbol{\varepsilon} \tag{1}$$

is $n^{-1}\,\mathrm{diag}(\boldsymbol{\Sigma})/\mathrm{Var}\,(X)$ for a sample of size n. However, this paper demonstrates a surprising result that the estimation accuracy of $\boldsymbol{\beta}$ can be further improved if we replace the original response $\mathbf{Y}$ by a proper combination $\boldsymbol{\Gamma}^\top \mathbf{Y}$. The key is to find an orthogonal matrix $(\boldsymbol{\Gamma}, \boldsymbol{\Gamma}_0) \in \mathbb{R}^{r\times r}$ such that (i) $\mathrm{span}(\boldsymbol{\beta}) \subseteq \mathrm{span}(\boldsymbol{\Gamma})$, and (ii) $\boldsymbol{\Gamma}^\top \mathbf{Y} \perp\!\!\!\perp \boldsymbol{\Gamma}_0^\top \mathbf{Y} | X$, where the notation "$\perp\!\!\!\perp$" denotes independence. Model (1) can now be recast as

$$\mathbf{P}_{\boldsymbol{\Gamma}} \mathbf{Y} = \mathbf{P}_{\boldsymbol{\Gamma}} \boldsymbol{\alpha} + \boldsymbol{\beta} X + \mathbf{P}_{\boldsymbol{\Gamma}} \boldsymbol{\varepsilon}, \text{ and } \mathbf{P}_{\boldsymbol{\Gamma_0}} \mathbf{Y} = \mathbf{P}_{\boldsymbol{\Gamma}} \boldsymbol{\alpha} + \mathbf{P}_{\boldsymbol{\Gamma_0}} \boldsymbol{\varepsilon}. \tag{2}$$

The joint likelihood function of $(\mathbf{\Gamma}, \mathbf{\Gamma}_0)^\top \mathbf{Y}|X$ is the product of the likelihood function of $\mathbf{\Gamma}^\top \mathbf{Y}|X$ and that of $\mathbf{\Gamma}_0^\top \mathbf{Y}|X$, but only $\mathbf{\Gamma}^\top \mathbf{Y}|X$ helps infer about $\boldsymbol{\beta}$. With a fixed $\mathbf{\Gamma}$, the asymptotic variance of the maximum likelihood estimate of $\boldsymbol{\beta}$ based on (2) becomes $n^{-1}\,\mathrm{diag}(\mathbf{P}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Gamma}})/\mathrm{Var}\,(X)$. Since $\mathbf{\Sigma} = \mathbf{P}_{\mathbf{\Gamma}}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Gamma}} + \mathbf{P}_{\mathbf{\Gamma_0}}\mathbf{\Sigma}\mathbf{P}_{\mathbf{\Gamma_0}}$, the estimation accuracy of $\boldsymbol{\beta}$ is consequently improved by the transformed model. The link between $\boldsymbol{\beta}$ and $\mathbf{\Sigma}$ is the envelope $\boldsymbol{\mathcal{E}}_{\mathbf{\Sigma}}(\boldsymbol{\beta})$, or the intersection of all $\mathbf{\Gamma}$'s that satisfies (i) and (ii) simultaneously.

In particular, if X is categorical with two levels, the envelope models can be used for classification. Specifically, we let $\boldsymbol{\mu}_0 \equiv \mathbb{E}(\mathbf{Y}|X=0)$ and $\boldsymbol{\mu}_1 \equiv \mathbb{E}(\mathbf{Y}|X=1)$, which relates to model (1) by setting $\boldsymbol{\alpha} = \boldsymbol{\mu}_0$ and $\boldsymbol{\beta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$. The classical linear discriminant analysis classifies $\boldsymbol{y}$ to the reference population $\{X = 0\}$ if

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \mathbf{\Sigma}^{-1} \boldsymbol{y} > \frac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1). \tag{3}$$

Let $\mathbf{\Gamma} \in \mathbb{R}^{r\times u}$, $u \le r$, denote a semi-orthogonal basis matrix for $\boldsymbol{\mathcal{E}}_{\mathbf{\Sigma}}(\boldsymbol{\beta})$. Plug $\mathbf{\Sigma}^{-1} = \mathbf{\Gamma}\mathbf{\Omega}^{-1}\mathbf{\Gamma}^\top + \mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^\top$ into (3), where $\mathbf{\Omega} = \mathbf{\Gamma}^\top\mathbf{\Sigma}\mathbf{\Gamma}$ and $\mathbf{\Omega}_0 = \mathbf{\Gamma}_0^\top\mathbf{\Sigma}\mathbf{\Gamma}_0$. We have a new classification rule based on the envelope model

$$(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \mathbf{\Gamma}\mathbf{\Omega}^{-1}\mathbf{\Gamma}^\top \boldsymbol{y} > \frac{1}{2}(\boldsymbol{\mu}_0 - \boldsymbol{\mu}_1)^\top \mathbf{\Gamma}\mathbf{\Omega}^{-1}\mathbf{\Gamma}^\top(\boldsymbol{\mu}_0 + \boldsymbol{\mu}_1). \tag{4}$$

The authors conjectured that classification based on (4) is more accurate than that based on (3), especially when $u \ll r$ or the eigenvalues of $\mathbf{\Omega}$ are substantially larger than those of $\mathbf{\Omega}_0$. In the sequel we examine this issue through synthetic examples.

We generated X from a series of Bernoulli trials with probability of success 0.5. Set $r = 10$, $\boldsymbol{\alpha} = \boldsymbol{\mu}_0$, and $\boldsymbol{\beta} = \boldsymbol{\mu}_1 - \boldsymbol{\mu}_0$, where $\boldsymbol{\mu}_0 = (0,\ldots,0)^\top$, $\boldsymbol{\mu}_1 = (5,\ldots,5)^\top$ in model (1). The error $\boldsymbol{\varepsilon}$ was generated as normal with mean zero and covariance matrix $\mathbf{\Sigma} = \sigma^2\mathbf{\Gamma}\mathbf{\Gamma}^\top + \sigma_0^2\mathbf{\Gamma}_0\mathbf{\Omega}_0^{-1}\mathbf{\Gamma}_0^\top$, where $\sigma^2 = 5^2$ and $\mathbf{\Gamma} = (1,\ldots,1)/\sqrt{10}$. We chose $\sigma_0^2 = 1^2, 5^2$, and 9^2 to examine the effect of different noise levels. We considered three scenarios for $\mathbf{\Omega}_0 = (\omega_{0,ij})_{9\times 9}$: (i) $\omega_{0,ij}$ takes 1 if $i = j$ and 0 otherwise; (ii) $\omega_{0,ij} = 1/9\sum_{k=1}^{9} Z_{ki}Z_{kj}$ where Z_{ki}'s are i.i.d standard normal; and (iii) $\omega_{0,ij} = 0.9^{|i-j|}$. To compare the classification rules (3) and (4), we calculated the misclassification rates by using test data of size 100. The mean misclassification rates are reported in Table 1. The standard deviations of the mean rates were all within 0.005 and are omitted here.

We can see that part of the authors' conjecture was confirmed: classification based on the envelope model was better than the full model when the eigenvalues of $\mathbf{\Omega}$ were substantially larger than those of $\mathbf{\Omega}_0$. This improvement was significant when the sample size was small, and became less obvious as the sample size increased. Note that the magnitude of the eigenvalues of $\mathbf{\Omega}_0$ was compounded

Table 1. Comparison of misclassification rates for the full model rule (3) and envelope model rule (4) based on 500 repetitions. All numbers are reported in %.

$\boldsymbol{\Omega}_0$		Scenario (i)		Scenario (ii)		Scenario (iii)	
Model		Full	Envelope	Full	Envelope	Full	Envelope
$\sigma_0 = 1$	$n = 50$	8.758	5.884	8.582	5.842	8.610	5.878
	$n = 100$	7.244	5.914	6.896	5.630	7.256	5.996
	$n = 500$	5.914	5.594	5.700	5.580	5.952	5.712
$\sigma_0 = 5$	$n = 50$	8.652	6.466	8.588	7.282	8.678	8.834
	$n = 100$	7.192	6.816	7.058	6.578	6.966	7.256
	$n = 500$	6.056	5.830	5.896	5.822	5.866	5.970
$\sigma_0 = 9$	$n = 50$	8.670	12.272	8.676	15.236	8.692	19.638
	$n = 100$	7.134	9.224	7.190	11.742	7.000	14.432
	$n = 500$	5.854	6.378	6.054	7.108	6.066	8.968

by two factors: the noise level σ_0 and the correlation structure in $\boldsymbol{\Omega_0}$. The eigenvalues of $\boldsymbol{\Omega}_0$ tended to increase if either factor grew. We make the following observations from these two aspects.

- The noise level σ_0 in $\boldsymbol{\Sigma}$ affected the envelope model much more than the full model. As σ_0 increased, the misclassification rates based on the full model stayed the same while the performance of the envelope model deteriorated significantly. When $\sigma_0 = 1$ (which is small relative to $\sigma = 5$), the misclassification rates based on the envelope model were uniformly smaller than those based on the full model. When σ_0 was moderate, both models performed comparatively. However, when σ_0 increased to 9, the envelope model performed much worse than the full model.

- The envelope model was slightly more sensitive to correlation structure of $\boldsymbol{\Omega}_0$ in $\boldsymbol{\Sigma}$ than the full model. In scenarios (i) and (ii), the envelope model performed better if σ_0 was small or moderate. However, in the extreme scenario (iii), where high correlation was present, the full model outperformed the envelope model even with moderate noise level $\sigma_0 = 5$. This indicates that the envelope model possibly breaks down in terms of misclassification rate due to strong correlation of $\boldsymbol{\varepsilon}$. The authors have shown that stronger correlation in $\boldsymbol{\varepsilon}$ makes the superiority of the envelope model more significant in terms of estimation efficiency in the parameter estimating setting, while we see the contrary was true in terms of misclassification rate in the discriminant analysis setting when $\sigma_0 = 9$.

The envelope model was better than the full model in terms of asymptotic variance of the estimate of $\boldsymbol{\beta}$ in model (1). However, our observations indicate that the superior performance of the envelope model for parameter estimation may not necessarily lead to better performance in the classification setting in terms of misclassification rate. Some data driven methods, such as generalized cross validation, might be used determine the better classification rule for any particular problem.

Statistics Department, The Fox School of Business and Management, Temple University, Alter Hall 326, 1801 Liacouras Walk, Philadelphia, PA 19122-6083, USA.
E-mail: ydong@temple.edu

School of Finance and Statistics, East China Normal University, Shanghai, 200241, P.R.China.
E-mail: lpzhu@stat.ecnu.edu.cn

(Received January 2010; accepted February 2010)

COMMENT

Heng-Hui Lue

Tunghai University

It is our great pleasure to congratulate the authors for making impressive contributions to multivariate response regression problems. Over the past few years, there has been a considerable amount of work in the dimension reduction area (see Li, et al. (2003); Setodji and Cook (2004); Yoo and Cook (2007); Li, Wen, and Zhu (2008) and references therein). The authors bring up a delicate but potentially powerful idea: developing a parsimonious approach to maximally reducing the number of parameters based on the minimal reducing subspace. The authors also derive the asymptotic results that offer the possibility of properly conducting statistical inferences on testing hypotheses in multivariate response regressions. The simultaneous estimation of both the parameters and directional components of the model leads to some significant benefits.

We have two comments on this work. The first concerns moderate violations of the multivariate linear regression assumptions, e.g. heterogeneity and monotonic nonlinearity. To evaluate the performance of estimation, we generate an i.i.d. example from the heterogeneity model

$$\begin{aligned} Y_1 &= x_1 + x_2 + c\,\epsilon_1 \exp\{2(1-x_3)\}, \\ Y_2 &= x_1 + x_2 - c\,\epsilon_2 \exp\{2x_3\}, \end{aligned} \tag{1}$$

Table 1. Means and standard deviations of $R^2(\hat{b}_1)$ and $R^2(\hat{b}_2)$ for model (1) by (a) mrSIR; (b) EM in 1,000 replications.

	(a)						(b)					
$n = 200$	$R^2(\hat{b}_1)$			$R^2(\hat{b}_2)$			$R^2(\hat{b}_1)$			$R^2(\hat{b}_2)$		
c	0.10	0.05	0.025	0.10	0.05	0.025	0.10	0.05	0.025	0.10	0.05	0.025
mean	0.990	0.993	0.992	0.958	0.898	0.677	0.886	0.962	0.988	0.526	0.525	0.501
s.d.	0.007	0.006	0.054	0.033	0.105	0.261	0.168	0.079	0.037	0.283	0.285	0.298
$n = 500$	$R^2(\hat{b}_1)$			$R^2(\hat{b}_2)$			$R^2(\hat{b}_1)$			$R^2(\hat{b}_2)$		
c	0.10	0.05	0.025	0.10	0.05	0.025	0.10	0.05	0.025	0.10	0.05	0.025
mean	0.996	0.997	0.998	0.986	0.971	0.915	0.927	0.974	0.993	0.529	0.518	0.524
s.d.	0.003	0.002	0.002	0.010	0.021	0.078	0.125	0.056	0.016	0.292	0.304	0.307

where c is a tuning constant, and six coordinates of $\mathbf{X}$, the rest of Y_i's, and ϵ_i's are independent standard normals. Set $p = 6$ and $r = 5$. Let $\beta_1 = (1,1,0,0,0,0)'$ and $\beta_2 = (0,0,1,0,0,0)'$ be the true directions. The data contain both linear and exponential features. For simplicity, we abbreviate the envelope model as EM and Lue's as mrSIR (Lue (2009)). The performance is then compared with mrSIR for illustration. We use an affine invariant criterion (Li (1991)), $R^2(b) = \max_{\beta \in \mathcal{B}} (b'\beta)^2/(b'b \cdot \beta'\beta)$, where $\mathcal{B}$ is the true dimension-reduction space, and we emphasize the effectiveness of estimated sufficient dimension reduction directions. The 1,000 datasets were then simulated from (1) with sample sizes $n = 200$ and 500. Table 1 summarizes the performance of the true direction estimation for two methods. Lue's estimates were close to the true directions except for the case of $n = 200$ and $c = 0.025$, with means ranging from 0.898 to 0.998, whereas the envelope model produced some bias in estimation. The best view for model (1) with $c = 0.05$ in a single run is shown in Figure 1, which reveals a fan-shaped pattern. The envelope model seems sensitive to the heteroscedasticity. Figure 2 shows the changes of means of $R^2(\hat{b}_j)$, for $j = 1, 2$, with $n = 500$, as p or r increases. As expected, the trend was stable as r increased; however, it decreased as p grew. Here we used the LDR-package for the envelope model (Cook, Forzani, and Tomassi (2009)), which is available at `http://sites.google.com/site/lilianaforzani/ldr-package`.

A second concern is with the asymptotic variance. We extend the asymptotic variance for an estimated direction, obtained by Chen and Li (1998), from a univariate response to multivariate responses. The result is described as follows. Given the most predictable variates $\check{\mathbf{Y}} = (\check{Y}_1, \cdots, \check{Y}_L)'$ and slices $H = h_1 \times \cdots \times h_L$, for $1 \le L \le r$, let $\delta_h(\check{\mathbf{Y}}) = 1$ if $\check{\mathbf{Y}}$ falls into the hth slice, $1 \le h \le H$, and 0 otherwise. Let $\mu_h = E(\mathbf{X}|\delta_h(\check{\mathbf{Y}}) = 1)$ and $\Sigma_\eta = \sum_{h=1}^{H} p_h(\mu_h - E\mathbf{X})(\mu_h -$

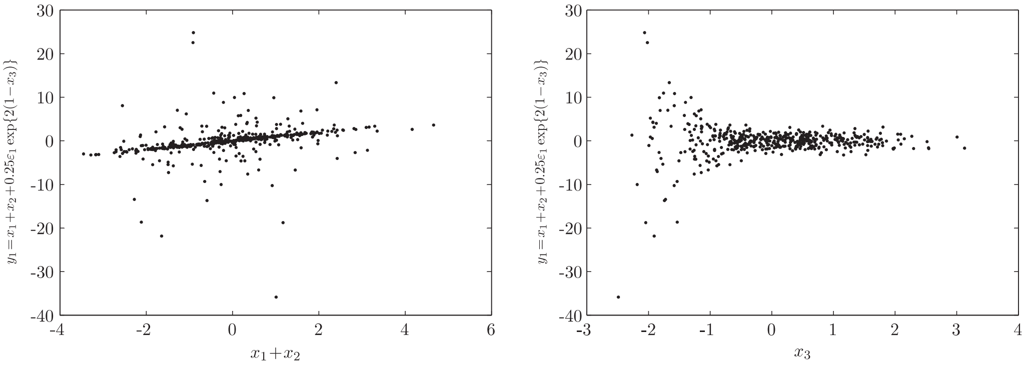


Figure 1. Best view for model 1 $c = 0.05$, $r = 5$, $p = 6$, and $n = 500$.

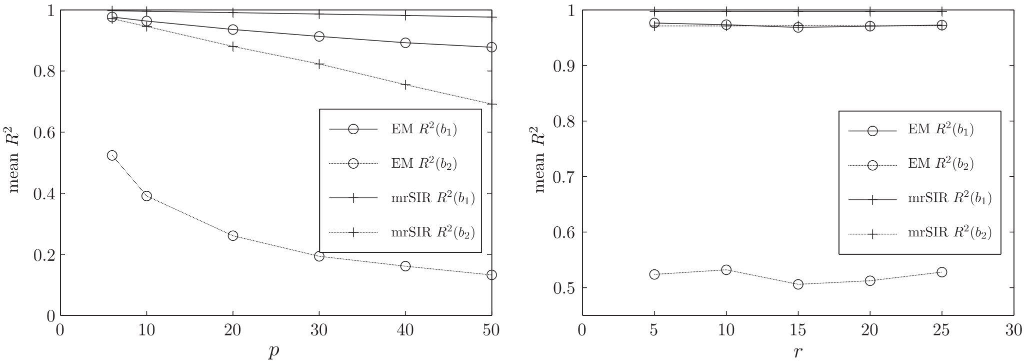


Figure 2. Means of R^2. Left panel with p increasing; right panel with r increasing.

$E\mathbf{X})'$, where $p_h = E\delta_h(\check{\mathbf{Y}})$, and let $\Delta v = \hat{v}_i - v_i$ and $\tilde{v}_i = \hat{\Sigma}_{\mathbf{X}}^{-1}\hat{\Sigma}_\eta v_i$, where v_i is the eigenvector for the eigenvalue decomposition $\Sigma_\eta v_i = \lambda_i \Sigma_{\mathbf{X}} v_i$. If e is any vector which is orthogonal to the effective dimension reduction space, then we have $e'\Delta v = \lambda_i^{-1} e'\tilde{v}_i + O_p(n^{-1})$. To approximate the term $e'\tilde{v}_i$, it can be shown that $e'\tilde{v}_i = e'\hat{\Sigma}_{\mathbf{X}}^{-1} n^{-1}\sum_{j=1}^n (\mathbf{X}_j - \bar{\mathbf{X}}) T_i(\check{Y}_j) + O_p(n^{-1})$, where $T_i(\check{Y}_j) = E(v_i'\mathbf{X}|\check{Y}_j) = v_i' \sum_h \delta_h(\check{Y}_j)\mu_h$ is the transformation of $\check{Y}_j$. For the residual $r_i = T_i(\check{\mathbf{Y}}) - ET_i(\check{\mathbf{Y}}) - b_i'(\mathbf{X} - E\mathbf{X})$, where b_i is the slope of multiple linear regression for $T_i(\check{\mathbf{Y}})$, it is straightforward to obtain $e'\hat{v}_i = \lambda_i^{-1} n^{-1} \sum_{j=1}^n r_{ij}\, e'\Sigma_{\mathbf{X}}^{-1}(\mathbf{X}_j - E\mathbf{X}) + O_p(n^{-1})$. When the error is homogeneous in the sense that $\mathrm{cov}(r_i^2, [e'\Sigma_{\mathbf{X}}^{-1}(\mathbf{X} - E\mathbf{X})]^2) = 0$, we have

$$\mathrm{var}(e'\hat{v}_i) = \frac{1-\lambda_i}{\lambda_i}\, n^{-1} e'\Sigma_{\mathbf{X}}^{-1} e. \tag{2}$$

Thus, we may use the diagonal elements of the covariance matrix in (2) as the asymptotic variance of $\hat{v}_i$ for inference.

Acknowledgement

We are grateful to Dr. Chun-Lung Su for helpful comments.

References

Chen, C. H. and Li, K. C. (1998). Can SIR be as popular as multiple linear regression? *Statist. Sinica* **8**, 289-316.

Cook, R. D., Forzani, L. and Tomassi, D. (2009). LDR a package for likelihood-based suNfficient dimension reduction. *J. Statist. Software*, in press.

Li, K. C. (1991). Sliced inverse regression for dimension reduction (with discussion). *J. Amer. Statist. Assoc.* **86**, 316-342.

Li, K. C., Aragon, Y., Shedden, K. and Agnan, C. T. (2003). Dimension reduction for multivariate response data. *J. Amer. Statist. Assoc.* **98**, 99-109.

Li, B., Wen, S. and Zhu, L. (2008). On a projective resampling method for dimension reduction with multivariate responses. *J. Amer. Statist. Assoc.* **103**, 1177-1186.

Lue, H. H. (2009). Sliced inverse regression for multivariate response regression. *J. Statist. Plann. Inference* **139**, 2656-2664.

Setodji, C. M. and Cook, R. D. (2004). K-means inverse regression. *Technometrics* **46**, 421-429.

Yoo, J. K. and Cook, R. D. (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression. *Biometrika* **94**, 231-242.

Department of Statistics, Tunghai University, Taichung, Taiwan 40704.

E-mail: hhlue@thu.edu.tw

(Received January 2010; accepted February 2010)

REJOINDER

R. Dennis Cook[1], Bing Li[2], Francesca Chiaromonte[2] and Zhihua Su[1]

[1] *University of Minnesota and* [2] *Pennsylvania State University*

We are grateful to the discussants for their encouraging reactions. We found their comments to be stimulating, many pointing to fresh directions that suggest envelopes may indeed have a place in the future of multivariate analysis. Since it was not possible to respond usefully to all of the discussants' comments, we focused our reply on common themes.

1. Advantages of Envelopes

We begin by considering circumstances under which envelopes might offer an advantage over the standard likelihood analysis when model (3.2) holds. A necessary condition to obtain an advantage is that $u < \min(r, p)$. When this is the case envelopes perform asymptotically better than the standard analysis simply because of parsimony; that is, because the envelope model comprises fewer parameters. When $\boldsymbol{\beta}$ has full column rank, this necessary condition reduces to $u < p < r$. Jia et al. reported simulation results for scenarios where $u < p < r$, and where $u < r < p$ and $\boldsymbol{\beta}$ has rank u.

However $u < \min(r, p)$, even when $u \ll \min(r, p)$, in and of itself does not guarantee substantial gains. Let $\|\mathbf{A}\|$ denote the spectral norm of the matrix $\mathbf{A}$. We have observed that the gains produced by envelopes are insubstantial when $\|\boldsymbol{\Sigma}_1\| \approx \|\boldsymbol{\Sigma}_2\|$, can be solid when $\|\boldsymbol{\Sigma}_1\| \gg \|\boldsymbol{\Sigma}_2\|$, and are typically massive when $\|\boldsymbol{\Sigma}_1\| \ll \|\boldsymbol{\Sigma}_2\|$. The latter observation is supported by the relative efficiency given in (6.6), and corroborated by both the simulations of Section 7.1 and the analysis of the wheat protein data in Section 7.2. In the first simulation scenario of Jia et al., neither $\|\boldsymbol{\Sigma}_1\|$ nor $\|\boldsymbol{\Sigma}_2\|$ dominate; consequently, we conjecture that the results they report in Figure 2.1 are due primarily to parsimony. We were encouraged by their simulation results overall, and anticipate that a stronger relative performance for envelopes can be demonstrated when one matrix clearly dominates, and in particular when $\|\boldsymbol{\Sigma}_1\| \ll \|\boldsymbol{\Sigma}_2\|$. In contrast, Ni controlled the relative sizes of $\|\boldsymbol{\Sigma}_2\|$ (his σ_0^2) and $\|\boldsymbol{\Sigma}_1\|$ (his σ^2), and when $\sigma_0 \gg \sigma$, observed good relative performance for the envelope estimator in Figure 1(a). Indeed, the peculiar dip in Figure 1(a) can be explained in light of our discussion around (6.6): OLS and envelope estimators have equal asymptotic efficiency when $\sigma_0 = \sigma$, but otherwise the latter is characterized by smaller variation. Ni's Figure 1(b) will be discussed in Section 3.

During the past few months we have analyzed many data set from the literature, mostly with small to moderate values of r. In some cases envelopes demonstrated no worthwhile gains over the standard analysis, in other cases they provided modest but desirable gains, and in yet others they indeed provided massive gains. For example, we considered a small data set from Johnson and Wichern (2008) comprising 42 air-pollution measurements recorded at noon in Los Angeles on different days. We took wind speed and solar radiation as predictors ($p = 2$), and measurements for CO, NO, NO2, O3, and HC as responses ($r = 5$). With $u = 1$, which is supported by the likelihood ratio test, the ratios of the standard errors between the full model and the envelope model range from 1.80 to 176.98. In this example $\|\widehat{\boldsymbol{\Sigma}}_1\| = 0.21$ and $\|\widehat{\boldsymbol{\Sigma}}_2\| = 31.06$, again supporting the notion that envelopes can give massive gains when $\|\boldsymbol{\Sigma}_1\| \ll \|\boldsymbol{\Sigma}_2\|$.

Our experience in comparing envelopes with other methods through simulations and data analysis led us to the empirical conclusion that, depending on u and the relationship between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, envelopes can perform about the same, better, or much better than other methods in prediction and estimation. Indeed, this conclusion seems to be supported by the simulation results reported by the discussants.

Using an invariance argument, He and Zhou reasoned that the gains produced by envelopes arise from general benefits of shrinkage, rather than from any intrinsic benefits of the envelope model *per se*. We agree that the notion of shrinkage plays a role in enveloping, as does the related notion of data regularization (see also discussion by Jia et al.). In fact, at the end of Section 1.1, we explicitly discussed enveloping as means of regularization, which pursues a measure of "eigen sparsity" through which the estimates are shrunk. In addition, we find value in our original motivation (Section 1.1) for the envelope model as a means of characterizing regressions in which the distribution of some linear combinations $\mathbf{\Gamma}_0^T\mathbf{Y}$ of the response does not change with the predictors. Shrinkage and regularization can then be seen as ways to downweight or eliminate the linear combinations that change relatively little.

We appreciate He and Zhou's expression of Box's memorable statement "All models are wrong, but some are useful." They rightly point out that if $\mathbf{\Sigma}$ is chosen randomly from an absolutely continuous distribution, then $\boldsymbol{\beta}$ falling into any lower-dimensional envelope is a zero-probability event. However, once more echoing Box, one could go a step further and say that essentially all useful models are zero-probability events. Take for example the sparse linear regression model $Y = \boldsymbol{\beta}^T X + \boldsymbol{\varepsilon}$ where $\boldsymbol{\beta}_j = 0$ for a subset $j \in J$ of the indexes $\{1, \ldots, p\}$. Since any proper subspace of $\mathbb{R}^p$ has Lebesgue measure 0, this model is also a zero-probability event in terms of any probability distribution for $\boldsymbol{\beta}$ dominated by this measure. Another example is the non-parametric variable selection model $Y \perp\!\!\!\perp X | X_j, j \in J$, which again is a zero-probability event in the same sense and for the same reason. Just which specific zero-probability event we should pay attention to is a piece of transcendental intuition with which, we hope, nature can strike a chord. We argue that assuming $\boldsymbol{\beta}$ to fall into a lower dimensional envelope is at least as reasonable as assuming some components of $\boldsymbol{\beta}$ to be 0; in fact, the latter is a special case of the former if the envelope is taken to be the span of $\{e_j : j \in J\}$, where e_j indicates the vector whose jth component is 1 and all other components are 0. In this sense, enveloping, shrinkage and regularization are manifestations of the same basic philosophy.

2. Partial Least Squares

In his discussion, Helland pointed out a very interesting and important connection with Partial Least Squares (PLS). The model described by Helland is in fact an envelope model, where the envelope is the one generated by the predictor covariance matrix, $\mathbf{\Sigma}_X = \text{var}(X)$. We discussed this model briefly in Section 8.4 and related it to our previous work in Cook, Li, and Chiaromonte (2007). However, the model at the core of our current article uses the envelope generated by the error covariance matrix or, equivalently, by the conditional covariance matrix of $\mathbf{Y}$ given X, $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathbf{Y}|X} = \text{var}(\mathbf{Y}|X)$.

The connection between partial least squares and envelopes can be summed up as follows. Let $\mathbf{v}_1, \ldots, \mathbf{v}_r$ be the r eigenvectors of $\mathbf{\Sigma} = \mathbf{\Sigma}_{\mathbf{Y}|X}$ and let $\mathbf{w}_1, \ldots, \mathbf{w}_p$ be the p eigenvectors of $\mathbf{\Sigma}_X$. We are interested in the conditional mean $\mathcal{E}(\mathbf{Y}|X)$, which is a focal point of regression. Intuitively, not all of the $\mathbf{v}$'s or $\mathbf{w}$'s are relevant to the regression. If all of the $\mathbf{v}$'s, but only some of the $\mathbf{w}$'s, are relevant, then the model is related to partial least squares in the way described by Helland. If all of the $\mathbf{w}$'s, but only some of the $\mathbf{v}$'s, are relevant, then the model is the main envelope model described in our article, and it is not directly related to partial least squares. As Helland pointed out, in the first case there is an explicit solution that essentially coincides with the projection of the least squares estimate onto the envelope. However, in the second case there is *no* explicit solution, and numerical maximization over a Grassmann manifold, or some other iterative algorithm, is necessary. Importantly, there is also a third scenario: the one in which not all of the $\mathbf{v}$'s and not all of the $\mathbf{w}$'s are relevant to the regression. This would induce further model reduction, and is a promising field to explore – relatedly, we discussed the possibility of simultaneous envelopes in Section 8.5.

Hung and Huang's Proposition 1 is a nice addition to the tools for studying envelopes. They used it as a foundation for combining PLS and envelopes in a novel algorithm for prediction, as illustrated in their classification example with the Support Vector Machine (SVM). The idea of applying PLS as an initial reduction stage, followed by enveloping is indeed intriguing, particularly for regressions where $n \ll p + r$. Here we would like to make two points. First, Chung and Keleş (2010) recently proved that the PLS estimator of the coefficient vector in the univariate linear regression of $Y \in \mathbb{R}^1$ on $X \in \mathbb{R}^p$ is consistent when $p/n \to 0$, but inconsistent otherwise. As a consequence, we hesitate to use PLS when $n \ll p + r$, although this requires a more detailed study in view of the results shown in Hung and Huang's Figure 1(b).

Second, in our experience the relative performance of PLS and envelopes again depends on the relationship between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$. To illustrate, we set $r = 1$ and $p = 7$ and simulated (Y, X) as multivariate normal data, with the

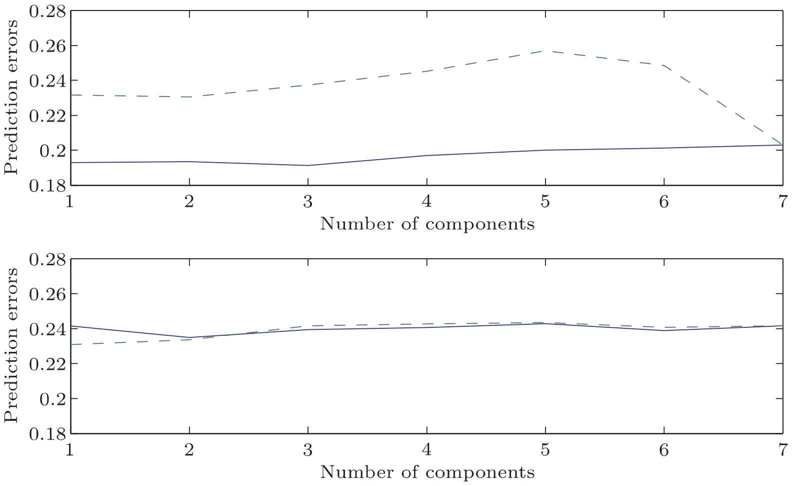


Figure 1. Simulation results for prediction error of envelope and PLS estimators. The solid and dashed line represent prediction errors obtained with enveloping and PLS, respectively.

ultimate goal of predicting the univariate response Y from u linear combinations of the 7 predictors in X. We reduced the dimension of X by using an envelope for the inverse regression of X on Y, essentially treating X as the response, and then predicted using the linear regression of Y on $\widehat{\mathbf{\Gamma}}^T X$, the u linear combinations of X arising from the estimated envelope. The performance of envelopes relative to PLS in this setting is controlled by u and by the relationship between $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$, which now refer to the inverse regression of X on Y. In our simulations we used $n = 60$, a true $u = 2$, and a $\mathbf{\Sigma}$ constructed to have eigenvalues about 0.05, 1.6, 3, 28, 80, 84 and 584. Results are shown in Figure 1, where the horizontal axis is the dimension of the envelope employed on the data, as well as the number of components used in PLS, and the vertical axis is the squared prediction error determined by 5-fold cross validation. In the top panel of Figure 1, $\mathbf{\Sigma}_1$ captures the eigenvalues 84 and 584, and enveloping does significantly better than PLS. In the bottom panel, $\mathbf{\Sigma}_1$ captures the eigenvalues 80 and 84, and the performance of the two approaches is essentially the same.

3. Computing

Many of the discussants raised the issue of computing, highlighting the fact that Grassmann optimization can be quite slow when r is large. This arises in part because the algebraic dimension of the Grassmann manifold is $u(r-u)$; if $u = 50$ and $r = 100$, our optimization is taking place essentially in $\mathbb{R}^{2500}$. Our current code is useful for r up to 100 with modest values of u, but is still annoyingly slow in larger problems – although we are working on faster versions. Local optima can also be an issue, mostly when the signal is weak.

Two packages, LDR and GrassmannOptim, are available for optimization over Grassmann manifolds. LDR is written in Matlab and is available at

`http://sites.google.com/sites/lilianaforzani/ldr-package.`

This package implements many methods for sufficient dimension reduction, and includes routines for envelope models which require analytic first derivatives and use numerical second derivatives. As a consequence, starting at a root-n consistent estimator results in a final estimator that is asymptotically equivalent to the MLE, even if local optima are present (see, for example, Small, Wang, and Yang (2000)). Nevertheless, like most programs for non-linear optimization, there is no guarantee that it will always reach the global maximum. This might not be worrisome in the analysis of a single data set where it is possible to study the objective function. However, local optima can bias simulation results targeted at the MLE and be quite annoying. GrassmannOptim is written in R and is available at

`http://CRAN.R-projects.org/package=GrassmannOptim.`

While there are as yet no special routines for envelope models based on GrassmannOptim, one advantage of this package is that it contains an option for simulated annealing that can avoid local optima at the expense of computing time. Computing time can be quite substantial with both packages if r is large.

We were initially a bit perplexed by the results shown in Ni's Figure 1(b). His simulation model should satisfy the relative efficiency properties described in (6.6), so enveloping should be asymptotically superior to ordinary least squares (OLS) when $\sigma_0 \neq \sigma_1$ – pointing to a disagreement between (6.6) and Ni's Figure 1(b). To see if Ni's routine might have gotten trapped in local optima, we reran his simulation scenario with our code using the true $\mathbf{\Gamma}$ as the starting value. The results, shown in Figure 2, agree qualitatively with the relative efficiency in (6.6). It seems then that algorithms using random starting points might indeed be prone to reaching local maxima. To investigate this further, we ran our code for Ni's simulation with $\sigma = 3$ and $\sigma_0 = 1$, the right-most point in Ni's Figure 1(b) – results are shown in Figure 3. The means of Ni's $f(v)$, 0.996 for envelopes and 0.840 for OLS, correspond reasonably to those at the right-most point in Figure 2, and the quality of the relative variations of the two estimators is predicted by (6.6). We expect that our implementation of Grassmann optimization worked well here because it includes an initial search over potential starting values for $\mathbf{\Gamma}$, including the eigenvectors of $\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}$. For instance, when $\sigma = 3$ and $\sigma_0 = 1$ in

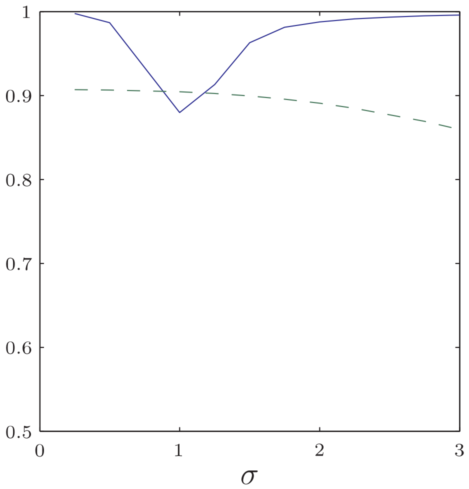
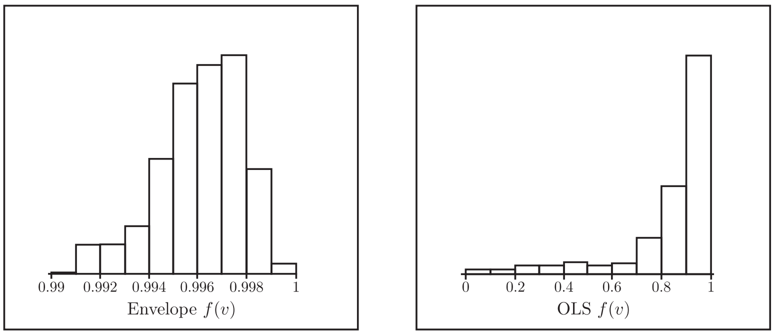


Figure 2. Rerun of Ni's Figure 1(b) using $\mathbf{\Gamma}$ as the starting value. The vertical axis is $h(v)$ and the horizontal axis is σ. The solid and dashed line correspond to enveloping and OLS, respectively.



(a) Envelopes (b) OLS

Figure 3. Histograms of $f(v)$ from 100 runs with $\sigma = 3$ and $\sigma_0 = 1$.

Ni's simulation model, $\mathbf{\Sigma_Y} = 9.5\mathbf{\Gamma}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Gamma}_0^T$, and the first eigenvector of $\widehat{\mathbf{\Sigma}}_\mathbf{Y}$ provides a root-n consistent starting value.

Hung and Huang's proposal of using PLS for preliminary reduction followed by enveloping is appealing from a computational point of view. With this approach, they were able to analyze a classification problem with $r = 15,109$, $p = 5$, and $u = 4$, which we found to be impressive because it is not possible to handle problems of this size with our current implementation of Grassmann optimization. We expect that the numerical instability they noticed in their Figure 3 for larger component numbers was caused by convergence to local optima, and does not reflect an intrinsic property of envelopes. Issues related to local optima may also be the cause of the extra variation in envelope predictions observed by Jia et al.

He and Zou's use of canonical correlations for dimension reduction was motivated in part by computational simplicity and availability of standard software. We would like to point out that indeed canonical correlations could be used for preliminary reduction followed by enveloping, in much the same way that Hung and Huang combined PLS and envelopes. Nevertheless, one possible criticism of the C-estimator is that it fails to take into account the asymmetric nature of regression; this is related to our rejoinder to Helland in Section 2. Since in regression we are often interested in estimating $\mathcal{E}(\mathbf{Y}|X)$, dimension reduction for $\mathbf{Y}$ is different in nature from dimension reduction for X. Using the cross-covariance matrix of X and $\mathbf{Y}$ to generate the envelope upon which to project $\boldsymbol{\beta}$ treats X and $\mathbf{Y}$ symmetrically. The C-estimator may be logically more appropriate in contexts where X and $\mathbf{Y}$ do play symmetric roles.

Finally, Ni proposed an interesting algorithm based on one-at-a-time minimization over basis vectors. This algorithm also merits further study, but its performance may depend heavily on having good starting values to avoid local optima.

4. Extensions and Combinations

The discussants mentioned several thought-provoking ways in which enveloping might be extended or combined with other methods. Wen's result on Fisher consistency of the envelope MLE under a misspecified link function is intriguing because it suggests that there may well be a useful link-free version of enveloping methodology. Along similar lines, Helland hinted that it may be possible to adapt enveloping for application with generalized linear models.

He and Zhou expressed the view that "More efficient estimation is often achieved without reliance on any formal dimension reduction method," and went on to illustrate their point by using a penalized full model log likelihood to demonstrate that shrinkage can result in gains comparable to, or exceeding those of enveloping. The differences between shrinkage and enveloping are certainly worth exploring, but we emphasize that this is not an either-or situation; there is nothing in principle that would prevent us from penalizing an envelope log likelihood, thereby combining the benefits of both approaches.

Indeed, using a penalized envelope log likelihood is one of the proposals by Yu and Zhu. We think that this is a promising direction to pursue. Consider, for example, the penalty function $\rho(\mathbf{\Gamma}) = \lambda \sum_{i=1}^{r} (\mathbf{\Gamma}\mathbf{\Gamma}^T)_{i,i}^{1/2}$, which is like the penalty suggested by Yu and Zhu, except that only the diagonal terms are used. For any orthogonal matrix $\mathbf{O} \in \mathbb{R}^{u \times u}$, $\rho(\mathbf{\Gamma}) = \rho(\mathbf{\Gamma}\mathbf{O})$, and consequently ρ depends only on span$(\mathbf{\Gamma})$. In effect, ρ penalizes the rows of $\mathbf{\Gamma}$, and this is exactly what is needed to tell which responses are independent of changes in X. Chen, Zou, and Cook (2010) used ρ in combination with standard methods like SIR and SAVE to produce sparse estimates of the central subspace. They showed that their penalized subspace estimator CISE possesses the oracle property, and that it dominates various other methods, including methods that penalize individual elements of $\mathbf{\Gamma}$. This is a specific instance of the synergies that can be created between dimension reduction and penalization. X. Chen (personal communication) has also conducted a small simulation study to explore the potential advantages of using ρ in combination with enveloping based on minimizing

$$\log\det(\mathbf{\Gamma}^T\widehat{\mathbf{\Sigma}}_{\mathrm{res}}\mathbf{\Gamma}) + \log\det(\mathbf{\Gamma}_0^T\widehat{\mathbf{\Sigma}}_{\mathbf{Y}}\mathbf{\Gamma}_0) + \rho(\mathbf{\Gamma})$$

over the Grassmann manifold $\mathbb{G}^{r\times u}$. Here too, the results support the notion of a synergy between dimension reduction and penalization.

We framed our development in the context of the multivariate normal linear model, but the underlying idea and formal definition of an envelope are based only on moments and do not require normality. Consequently, we are free to pursue envelope estimation in ways that rely less on an underlying distribution. Thinking along these lines, we pose the following approach: For each fixed $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$, let $\mathbf{v}_1(\boldsymbol{\beta}), \ldots, \mathbf{v}_r(\boldsymbol{\beta})$ be the eigenvectors of

$$\widehat{\mathbf{\Sigma}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{Y}_i - \boldsymbol{\beta}X_i)(\mathbf{Y}_i - \boldsymbol{\beta}X_i)^T.$$

We want $\boldsymbol{\beta}$ to be such that

1. $(\boldsymbol{\beta}X_1, \ldots, \boldsymbol{\beta}X_n)$ is as close to $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$ as possible, and
2. it is orthogonal to as many as eigenvectors $\mathbf{v}_\ell(\boldsymbol{\beta})$ of $\widehat{\mathbf{\Sigma}}(\boldsymbol{\beta})$ as possible, so that the remaining eigenvectors effectively form an envelope.

It therefore seems reasonable to minimize the objective function:

$$\sum_{i=1}^{n}\|\mathbf{Y}_i - \boldsymbol{\beta}X_i\|^2 + \sum_{\ell=1}^{r}\lambda_\ell\sqrt{\mathbf{v}_\ell^T(\boldsymbol{\beta})\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}\mathbf{v}_\ell(\boldsymbol{\beta})}.$$

Intuitively, minimizing this function would bring $\boldsymbol{\beta}X$ close to $\mathbf{Y}$, and at the same time force $\boldsymbol{\beta}$ to be orthogonal to a subset of the eigenvectors, depending on the tuning parameters λ_ℓ.

One can also consider sparsity of X and eigen-sparsity of $\mathbf{Y}$ together – for example by minimizing the function

$$\sum_{i=1}^{n}\|\mathbf{Y}_i - \boldsymbol{\beta}X_i\|^2 + \sum_{\ell=1}^{r}\lambda_\ell\sqrt{\mathbf{v}_\ell^T(\boldsymbol{\beta})\boldsymbol{\beta}(\boldsymbol{\beta}^T\boldsymbol{\beta})^{-1}\boldsymbol{\beta}\mathbf{v}_\ell(\boldsymbol{\beta})} + \sum_{k=1}^{r}\sum_{\ell=1}^{p}\tau_{k\ell}|\beta_{k\ell}|.$$

5. Discrimination

Hung and Huang and Dong and Zhu considered the value of enveloping in the context of discrimination, although from different perspectives. Hung and Huang demonstrated good performance of PLS and enveloping relative to SVM.

Dong and Zhu addressed directly a conjecture we made in Section 8.2; namely, that when $\|\mathbf{\Omega}_0\| \ll \|\mathbf{\Omega}\|$ (equivalently, $\|\mathbf{\Sigma}_2\| \ll \|\mathbf{\Sigma}_1\|$), misclassification rates based on enveloping are substantially less than those based on Fisher's linear discriminant. The required relation here, $\|\mathbf{\Sigma}_2\| \ll \|\mathbf{\Sigma}_1\|$, is the reverse of the most desirable relation $\|\mathbf{\Sigma}_1\| \ll \|\mathbf{\Sigma}_2\|$ considered in Section 1 of this rejoinder. Intuitively, the difference arises because, from a predictive point of view, the roles of $\mathbf{Y}$ and X are reversed: In Section 1 we were concerned with predicting $\mathbf{Y}$ from X, while discriminant analysis deals with predicting X from $\mathbf{Y}$. We were pleased to see that Dong and Zhu confirmed our conjecture with up to 30% gains in prediction error for enveloping. However, similar to the circumstances surrounding Ni's Figure 1(b), we did not anticipate that enveloping would be inferior to Fisher's linear discriminant when $\|\mathbf{\Sigma}_1\| \ll \|\mathbf{\Sigma}_2\|$; that is, when $\sigma_0 \gg \sigma$. Our intuition suggests that the results shown in Dong and Zhu's Table 1 for $\sigma_0 = 9$ are again due to issues with starting values and local optima. To further explore this, we reran two instances from their Table 1, both with $n = 50$ in scenario (i), but using 2000 replications. In the first, we set $\sigma_0 = 1$ and obtained misclassification rates (Full, Envelope) = $(8.707, 5.877)$, which agree well with their result (Full, Envelope) = $(8.758, 5.884)$. In the second instance, we set $\sigma_0 = 9$ and obtained (Full, Envelope) = $(8.862, 9.089)$, which shows a much closer agreement between the methods than does their result (Full, Envelope) = $(8.670, 12.272)$. We do not know why our results differ, but suspect the reason rests again operationally with starting values. In any case, we would like to emphasize that this, too, is not an either-or situation. Fisher's linear discriminant arises as a special case of enveloping when $u = r$. Consequently, in practical problems we might use cross-validation to choose u, perhaps arriving at Fisher's discriminant when $\|\mathbf{\Sigma}_1\| \ll \|\mathbf{\Sigma}_2\|$, but typically using proper envelope classification ($u < r$) with improved performance when $\|\mathbf{\Sigma}_2\| \ll \|\mathbf{\Sigma}_1\|$. Finally, revisiting a theme we introduced in Section 4, penalization might be combined with envelope discrimination to improve classification rates even further.

6. Other Issues

6.1. Second order bias

Yu and Zhu raised the possibility that there might be a worrisome second order bias in the envelope estimator of $\boldsymbol{\beta}$, and pointed to a couple of ways in which this bias could be mitigated. However, judging form their numerical results in Table 1, the gain by bias correction seems to be modest in this instance. The squared bias norms given in their Table 1 relate to $\boldsymbol{\beta} = (\sqrt{10}, \ldots, \sqrt{10})^T$. Assuming that each element of $\sum_{i=1}^{200} \boldsymbol{\beta}_{\text{em}}^i - \boldsymbol{\beta}$ has the same order of magnitude, the element-wise bias they report in the worst case ($\sigma_0 = 2$, $n = 20$) is only about 1% of the common element ($\sqrt{10}$) of $\boldsymbol{\beta}$. A 1% bias will often be swamped by variation, and may be unimportant for the scientific substance of a data analysis. Nevertheless, we do not doubt that in general bias correction can be beneficial for dimension reduction and for enveloping, particularly when the sample size is small or moderate – as it is for classical inference.

6.2. Heterscedasticity

Lue and Su bring up the important issue of heteroscedasticity. Their simulation results illuminate the following point. In any linear regression model, the linear coefficient cannot fully recover a function of X in the variance of the error, unless that function depends only on the effective predictor. More specifically, consider the model

$$\mathbf{Y} = \boldsymbol{\beta} X + \mathbf{F}(\boldsymbol{\gamma}^T X)\boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{r\times p}$, $\boldsymbol{\gamma} \in \mathbb{R}^{p\times s}$, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, $\mathbf{F} : \mathbb{R}^s \to \mathbb{R}^{r\times r}$, and $\boldsymbol{\varepsilon} \perp\!\!\!\perp X$. In this model, unless $\text{span}(\boldsymbol{\gamma}) \subseteq \text{span}(\boldsymbol{\beta}^T)$, no consistent estimator of $\boldsymbol{\beta}$ can fully recover $\text{span}(\boldsymbol{\gamma})$. Lue and Su's model (1) is a special case of (1) with

$$\boldsymbol{\beta} = (\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_1 + \mathbf{e}_2, \mathbf{0}, \mathbf{0}, \mathbf{0})^T, \quad \boldsymbol{\gamma} = \mathbf{e}_3.$$

Hence $\text{span}(\boldsymbol{\gamma}) \perp \text{span}(\boldsymbol{\beta})$, which falls in the described scenario. At the same time, multivariate response SIR (mrSIR) is capable of recovering the directions in the variance. We suspect that the same trend displayed in Table 1 of Lue and Su's comments would hold even if the true $\boldsymbol{\beta}$ were used in place of the MLE under the envelope model. In this case, since $\text{span}(\boldsymbol{\gamma}) \perp \text{span}(\boldsymbol{\beta})$, information about $\boldsymbol{\gamma}$ can only be found in the residuals.

Another interesting point related to Lue and Su's discussion is how to construct an envelope model when the conditional covariance matrix $\text{var}(\mathbf{Y}|X)$ depends on X. In this case a reducing subspace $\mathcal{S}$ of $\text{var}(\mathbf{Y}|X)$ would also depend on X. At present we do not yet have an answer to this question. The notion underlying our envelope proposal is to link the regression mean structure with the error covariance structure – positing that the former depends on X, while the latter does not. This allows us to use $\boldsymbol{\Sigma} = \text{var}(\mathbf{Y}|X)$ (constant with X) to frame and focus inference on $\boldsymbol{\beta}$, increasing efficiency under appropriate conditions. Allowing $\boldsymbol{\Sigma} = \text{var}(\mathbf{Y}|X)$ to vary with X would put us in a much more general setting; intriguingly, this may both reduce and increase the gains accrued by linking regression mean and error covariance. We consider this an open and possibly promising avenue for further research.

References

Chen, X., Zou, C. and Cook, R. D. (2010). Coordinate-independent sparse sufficient dimension reduction and variable selection. *Ann. Statist.*, to appear.

Chung, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. Roy. Statist. Soc. Ser. B* **72**, 3-25.

Cook, R. D., Li, B. and Chiaromonte, F. (2007). Dimension reduction without matrix inversion. *Biometrika* **94**, 569-584.

Johnson, R. A. and Wichern, W. A. (2008). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Small, C. G., Wang, J. and Yang, Z. (2000). Eliminating multiple root problems in estimation. *Statist. Sci.* **15**, 313-332.

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: dennis@stat.umn.edu

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.

E-mail: bing@stat.psu.edu

Department of Statistics, The Pennsylvania State University, University Park PA, 16802, USA.

E-mail: chiaro@stat.psu.edu

School of Statistics, University of Minnesota, Minneapolis, MN 55455, USA.

E-mail: suzhihua@stat.umn.edu

(Received March 2010; accepted March 2010)