Biometrika (2010), **xx**, x, *pp*. 1–24

© 2007 Biometrika Trust

Printed in Great Britain

Partial Envelopes for Efficient Estimation in Multivariate Linear Regression

BY ZHIHUA SU & R. DENNIS COOK

School of Statistics, University of Minnesota, 224 Church St., S.E., Minneapolis, Minnesota 55414, U.S.A.

suzhihua@stat.umn.edu dennis@stat.umn.edu

SUMMARY

We introduce the partial envelope model, which leads to a parsimonious method for multivariate linear regression when some of the predictors are of special interest. It has the potential to achieve massive efficiency gains compared to the standard model in the estimation of the coefficients for the selected predictors. The partial envelope model is a variation on the envelope model proposed by Cook et al. (2010) but, as it focuses on part of the predictors, it has looser restrictions and can further improve efficiency.

We develop maximum likelihood estimation for the partial envelope model and discuss application of the bootstrap. An example is provided to illustrate some of its operating characteristics.

Some key words: Dimension reduction, Envelope model, Grassmann manifolds, Reducing subspaces.

1. INTRODUCTION

Introduced recently by Cook et al. (2010), enveloping is a new approach to multivariate analysis that has the potential to produce very substantial gains in efficiency. Their development was

in terms of the standard multivariate linear model

$$Y = \mu + \beta X + \varepsilon, \tag{1}$$

where $\mu \in \mathbb{R}^r$, the random response $Y \in \mathbb{R}^r$, the fixed predictor vector $X \in \mathbb{R}^p$ is centered to have sample mean 0, and the error vector $\varepsilon \sim N(0, \Sigma)$. They demonstrated that an envelope estimator of the unknown coefficient matrix $\beta \in \mathbb{R}^{r\times p}$ has the potential to achieve massive gains in efficiency relative to the standard estimator of β, and that these gains will be passed on to other tasks like prediction. In this article we propose an extension of envelopes called partial envelopes. Partial envelopes can be focused on selected columns of β and can achieve gains in efficiency beyond those possible by using an envelope. Additionally, the envelope estimator reduces to the standard estimator when $r \leq p$ and β is of rank r, so there is no possibility for efficiency gains in this setting. Partial envelopes remove this restriction, providing gains even when $r \leq p$.

In the next section we review envelopes and envelope estimation. Because this is a new area, our goal is to provide intuition and insight rather than technical details. Additional results for envelopes will be discussed during our extension to partial envelopes given in §3.

2. ENVELOPES

It can happen in the context of model (1) that some linear combinations of Y are immaterial to the regression because their distribution does not depend on X, while other linear combinations of Y do depend on X and are thus material to the regression. In effect, envelopes separate the material and immaterial parts of Y, and thereby allow for gains in efficiency.

Suppose that we can find a subspace $\mathcal{S} \subseteq \mathbb{R}^r$ so that

$$Q_{\mathcal{S}}Y \mid X \sim Q_{\mathcal{S}}Y, \;\; Q_{\mathcal{S}}Y \perp\!\!\!\perp P_{\mathcal{S}}Y \mid X, \tag{2}$$

where '$\sim$' means identically distributed, $P_{(\cdot)}$ projects onto the subspace indicated by its argument and $Q = I_r - P$. For any $\mathcal{S}$ with those properties, $P_{\mathcal{S}}Y$ carries all of the material information and perhaps some immaterial information, while $Q_{\mathcal{S}}Y$ carries just immaterial information. Let $\mathcal{B} = \text{span}(\beta)$. Then (2) holds if and only if $\mathcal{B} \subseteq \mathcal{S}$ and $\Sigma = \Sigma_{\mathcal{S}} + \Sigma_{\mathcal{S}^\perp}$, where $\Sigma_{\mathcal{S}} = \text{var}(P_{\mathcal{S}}Y)$ and $\Sigma_{\mathcal{S}^\perp} = \text{var}(Q_{\mathcal{S}}Y)$ (Cook et al., 2010). However, $\mathcal{S}$ is not necessarily unique because there may be infinitely many subspaces that satisfy these relations in a particular problem. A reducing subspace $\mathcal{S}$ of Σ has the property that $\Sigma\mathcal{S} \subseteq \mathcal{S}$ and $\Sigma\mathcal{S}^\perp \subseteq \mathcal{S}^\perp$ (see, for example, Conway, 1990). Cook et al. (2010) showed that $\mathcal{S}$ is a reducing subspace of Σ if and only if $\Sigma = \Sigma_{\mathcal{S}} + \Sigma_{\mathcal{S}^\perp}$. This enabled them to address the uniqueness issue and ensure that $P_{\mathcal{S}}Y$ contains only material information by defining the minimal subspace to be the intersection of all reducing subspaces of Σ that contain $\mathcal{B}$, which is called the Σ-envelope of $\mathcal{B}$ and denoted as $\mathcal{E}_\Sigma(\mathcal{B})$. Let $u = \dim\{\mathcal{E}_\Sigma(\mathcal{B})\}$. Then

$$\mathcal{B} \subseteq \mathcal{E}_\Sigma(\mathcal{B}), \;\; \Sigma = \Sigma_{\mathcal{E}} + \Sigma_{\mathcal{E}^\perp}, \tag{3}$$

where $\mathcal{E}_\Sigma(\mathcal{B})$ is shortened to $\mathcal{E}$ for subscripts. These relationships establish a unique link between the coefficient matrix β and the covariance matrix Σ of (1), and it is this link that has the potential to produce gains in the efficiency of estimates of β. In particular, Cook et al. (2010) demonstrated that these gains will be massive when $\Sigma_{\mathcal{E}^\perp}$ contains at least one eigenvalue that is substantially larger than the largest eigenvalue of $\Sigma_{\mathcal{E}}$, so that in effect Y contains redundant immaterial information. Model (1) will be called the standard model when (3) is not imposed and called the envelope model when (3) is imposed.

The left-hand panel of Fig. 1 gives a schematic illustration of envelope estimation in a regression with $r = 2$ responses and a single binary predictor X indicating one of two bivariate normal populations represented by ellipses that cover, say, 99% of their distributions. Cook et al. (2010)

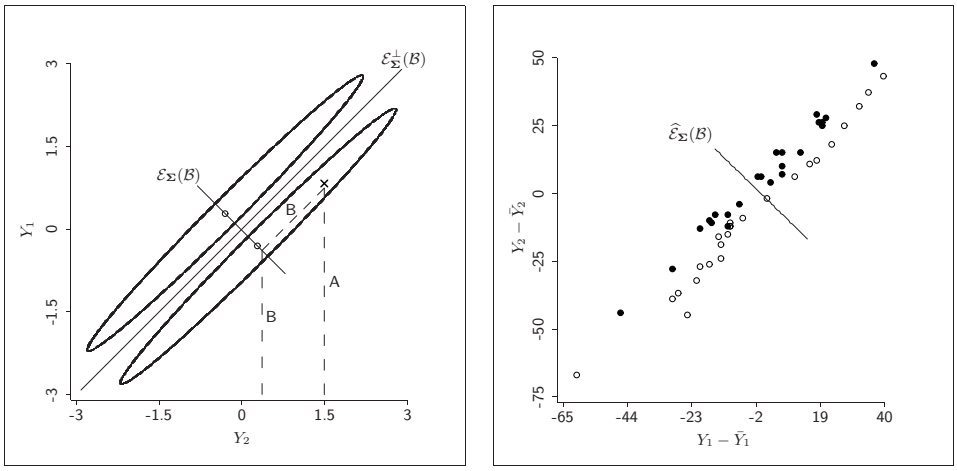

Fig. 1. Variance reduction by envelope estimation. Left-hand panel: Schematic representation of envelope estimation. Right-hand panel: Wheat protein data with points marked as high protein ● and low protein ○.

showed that $\mathcal{E}_\Sigma(\mathcal{B})$ is characterized by the identity $\mathcal{E}_\Sigma(\mathcal{B}) = \sum_{j=1}^k P_j\mathcal{B}$, where $k \leq r$ and P_j is the projection onto the j-th eigenspace of Σ. When Σ has distinct eigenvalues $\mathcal{E}_\Sigma(\mathcal{B})$ could be spanned by any of the 2^r possible subsets of its eigenvectors. In the left-hand panel of Fig. 1, the Σ-envelope of $\beta = (\beta_j) = \mathrm{E}(Y \mid X = 1) - \mathrm{E}(Y \mid X = 0)$ is parallel to the second eigenvector of Σ, and its orthogonal complement, which ties in with the immaterial part of Y, is parallel to the first eigenvector. This setup meets the population requirements (3). Standard inference on the second coordinate β_2 of β is based on marginal data obtained by projecting each data point onto the Y_2 axis, represented by the line segment A from a representative data point marked with an ex, giving rise to the usual two-sample inference methods. In contrast, for envelope inference on β_2 each data point is first projected onto $\mathcal{E}_\Sigma(\mathcal{B})$ and then projected onto the horizontal axis, represented in the plot by the two line segments marked B. Imagining this process for many data points from each population, it can be seen that the two empirical distributions of the projected data from the standard method will have much larger variation than the empirical distributions for the envelope method. The envelope $\mathcal{E}_\Sigma(\mathcal{B})$ is estimated in practice and so will have a degree of wobble that spreads the distribution of the data projected along routes represented by path B.

The asymptotic approximations of the variance of the envelope estimator of β take this wobble into account.

For a numerical illustration of this phenomenon, the right-hand panel of Fig. 1 shows a plot of $r = 2$ responses, the logarithms of near infrared reflectance at two wavelengths in the range 1680-2310 nm, measured on samples from two populations of ground wheat with low and high protein content (24 and 26 samples, respectively). The plotted points resemble the schematic representation in the left-hand panel of Fig. 1, and consequently an envelope analysis can be expected to yield more precise results than the standard analysis. The standard estimate of β_2 is -2·1 with a standard error of 9·4. In contrast, the envelope estimate of β_2 is -4·7 with a standard error of 0·46. The standard and envelope analyses for β_1 are related similarly.

While $\mathcal{E}_\Sigma(\mathcal{B})$ is necessarily spanned by some subset of the eigenvectors of Σ, the maximum likelihood estimator of $\mathcal{E}_\Sigma(\mathcal{B})$ shown in the right-hand panel of Fig. 1 is not spanned by a subset of the eigenvectors of the usual pooled estimator of Σ. This happens because the likelihood will balance the mean and variance conditions in (3), leading away from the sample eigenvectors.

If $r \leq p$ and β has full row rank r, then $\mathcal{B} = \mathbb{R}^r$. Consequently $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$, the envelope estimator of β reduces to the standard estimator of β, and enveloping offers no gains. When $r > p$ it is still possible to have $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$ so again enveloping offers no gains over the standard analysis. In these and other situations the partial envelopes developed in the next section can provide gains over both the standard and envelope estimators of β.

3. PARTIAL ENVELOPES

3·1. *Definition*

A subset of the predictors is often of special interest in multivariate regression, particularly when some predictors correspond to treatments while the remaining predictors are included to

account for heterogeneity among the experimental units. Partial envelopes are designed to focus consideration on the coefficients corresponding to the predictors of interest.

Partition X into to sets of predictors $X_1 \in \mathbb{R}^{p_1}$ and $X_2 \in \mathbb{R}^{p_1}$, $p_1 + p_2 = p$, $p_1 < r$, and conformably partition the columns of β into β_1 and β_2. Then model (1) can be rewritten as $Y = \mu + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where β_1 corresponds to the coefficients of interest. We can now consider the Σ-envelope for $\mathcal{B}_1 = \mathrm{span}(\beta_1)$, leaving β_2 as an unconstrained parameter. This leads to the parametric structure $\mathcal{B}_1 \subseteq \mathcal{E}_\Sigma(\mathcal{B}_1)$ and $\Sigma = P_{\mathcal{E}_1}\Sigma P_{\mathcal{E}_1} + Q_{\mathcal{E}_1}\Sigma Q_{\mathcal{E}_1}$, where $P_{\mathcal{E}_1}$ denotes the projection onto $\mathcal{E}_\Sigma(\mathcal{B}_1)$, which is called the partial envelope for $\mathcal{B}_1$. This is the same as the envelope structure, except the enveloping is relative to $\mathcal{B}_1$ instead of the larger space $\mathcal{B}$. For emphasis we will henceforth refer to $\mathcal{E}_\Sigma(\mathcal{B})$ as the full envelope. Because $\mathcal{B}_1 \subseteq \mathcal{B}$, the partial envelope is contained in the full envelope, $\mathcal{E}_\Sigma(\mathcal{B}_1) \subseteq \mathcal{E}_\Sigma(\mathcal{B})$, which allows the partial envelope to offer gains that may not be possible with the full envelope.

To provide some insights into this setting, let $R_{1|2}$ denote the population residuals from the multivariate linear regression of X_1 on X_2. Then, recalling that we have required X to be centered, the linear model can be re-parameterized as $Y = \mu + \beta_1 R_{1|2} + \beta_2^* X_2 + \varepsilon$, where β_2^* is a linear combination of β_1 and β_2. Next, let $R_{Y|2} = Y - \mu - \beta_2^* X_2$, the population residuals from the regression of Y on X_2 alone. We can now write a linear model involving β_1 alone: $R_{Y|2} = \beta_1 R_{1|2} + \varepsilon$. The partial envelope $\mathcal{E}_\Sigma(\mathcal{B}_1)$ is the same as the full envelope for $\mathcal{B}_1$ in the regression of $R_{Y|2}$ on $R_{1|2}$. In other words, we can interpret partial envelopes in terms of the motivating conditions (2) applied to the regression of $R_{Y|2}$ on $R_{1|2}$. In particular, the schematic representation of Fig. 1 also serves for partial envelopes by reinterpreting Y_1 and Y_2 as the two coordinates of $R_{Y|2}$ and reinterpreting the binary predictor as $R_{1|2}$. The classical added variable plot (Cook & Weisberg, 1982) is simply a plot of the sample version $\hat{R}_{Y|2}$ of $R_{Y|2}$ versus the sample version $\hat{R}_{1|2}$ of $R_{1|2}$.

3·2. *Maximum likelihood estimators*

Because we have centered the predictors, the maximum likelihood estimator of μ is simply $\hat{\mu} = \bar{Y}$. The estimators of the remaining parameters require the estimator of $\mathcal{E}_\Sigma(\mathcal{B}_1)$. Let $S_{Y|2}$ denote the sample covariance matrix of $\hat{R}_{Y|2}$, let $S_{R|2}$ denote the sample covariance matrix of the residual vectors from the regression of $\hat{R}_{Y|2}$ on $\hat{R}_{1|2}$ and let $|A|_0$ denote the product of the non-zero eigenvalues of the square matrix $A \geq 0$. Then, as shown in Appendix 1, the maximum likelihood estimator of $\mathcal{E}_\Sigma(\mathcal{B}_1)$ for a fixed dimension u_1 is

$$\widehat{\mathcal{E}}_\Sigma(\mathcal{B}_1) = \arg\min_{\mathcal{S} \in \mathcal{G}(u_1, r)} \{\log |P_{\mathcal{S}} S_{R|2} P_{\mathcal{S}}|_0 + \log |Q_{\mathcal{S}} S_{Y|2} Q_{\mathcal{S}}|_0\}, \tag{4}$$

where $\mathcal{G}(u_1, r)$ denotes the Grassmann manifold of dimension u_1 in $\mathbb{R}^r$. The Grassmann manifold $\mathcal{G}(u_1, r)$ is the set of all u_1-dimensional subspaces in $\mathbb{R}^r$. The maximum likelihood estimator of the full envelope $\mathcal{E}_\Sigma(\mathcal{B})$ is obtained using the objective function on the right hand side of (4) except $S_{R|2}$ is replaced with the sample covariance matrix S_R of the residual vectors from the fit of the standard model, $S_{Y|2}$ is replaced with the sample covariance matrix S_Y of the observed response vectors, u_1 is replaced with $u = \dim\{\mathcal{E}_\Sigma(\mathcal{B})\}$ and $\mathcal{S}$ is reinterpreted as an argument representing the full envelope (Cook et al., 2010). We see from this result that $\widehat{\mathcal{E}}_\Sigma(\mathcal{B}_1)$ is the same as the estimator of the full envelope applied in the context of the working model $\hat{R}_{Y|2} = \beta_1 \hat{R}_{1|2} + \varepsilon$.

The maximum likelihood estimator $\hat{\beta}_1$ of β_1 is the projection onto $\widehat{\mathcal{E}}_\Sigma(\mathcal{B}_1)$ of the estimator of β_1 from the standard model. The maximum likelihood estimator $\hat{\beta}_2$ of β_2 is the coefficient matrix from the ordinary least squares fit of the residuals $Y - \bar{Y} - \hat{\beta}_1 X_1$ on X_2. If X_1 and X_2 are orthogonal then $\hat{\beta}_2$ reduces to the maximum likelihood estimator of β_2 from the standard

model. The maximum likelihood estimator $\widehat{\Sigma}$ of Σ is

$$\widehat{\Sigma} = \hat{P}_{\mathcal{E}_1} S_{R|2} \hat{P}_{\mathcal{E}_1} + \hat{Q}_{\mathcal{E}_1} S_{Y|2} \hat{Q}_{\mathcal{E}_1},$$

where $\hat{P}_{\mathcal{E}_1}$ denotes the projection operator for $\widehat{\mathcal{E}}_\Sigma(\mathcal{B}_1)$, $\widehat{\Sigma}_{\mathcal{E}_1} = \hat{P}_{\mathcal{E}_1} S_{R|2} \hat{P}_{\mathcal{E}_1}$ is the estimated covariance matrix for the material part of Y and $\widehat{\Sigma}_{\mathcal{E}_1^\perp} = \hat{Q}_{\mathcal{E}_1} S_{Y|2} \hat{Q}_{\mathcal{E}_1}$ is the estimated covariance matrix for the immaterial part of Y.

3·3. *Asymptotic distributions*

In this section we give the asymptotic distributions of $n^{1/2}(\hat{\beta}_j - \beta_j)$, $j = 1, 2$. The results are conveniently expressed in terms of a coordinate version of the partial envelope model. Let $\Gamma \in \mathbb{R}^{r \times u_1}$ be a semi-orthogonal matrix, $\Gamma^T \Gamma = I_{u_1}$, whose columns form a basis for $\mathcal{E}_\Sigma(\mathcal{B}_1)$, let $(\Gamma, \Gamma_0) \in \mathbb{R}^{r \times r}$ be an orthogonal matrix and let $\eta \in \mathbb{R}^{u_1 \times p_1}$ be the coordinates of β_1 in terms of the basis matrix Γ. Then

$$Y = \mu + \Gamma \eta X_1 + \beta_2 X_2 + \varepsilon, \;\; \Sigma = \Sigma_{\mathcal{E}_1} + \Sigma_{\mathcal{E}_1^\perp} = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T, \tag{5}$$

where $\Omega \in \mathbb{R}^{u_1 \times u_1}$ and $\Omega_0 \in \mathbb{R}^{(r-u_1)\times(r-u_1)}$ are positive definite matrices that serve as coordinates of $\Sigma_{\mathcal{E}_1}$ and $\Sigma_{\mathcal{E}_1^\perp}$ relative to the basis matrices Γ for $\mathcal{E}_\Sigma(\mathcal{B}_1)$ and its orthogonal complement.

In preparation for the limiting distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$, let Δ denote the limit as the sample size $n \to \infty$ of the sample covariance matrix of X, and partition $\Delta = (\Delta_{jk})$ according to the partitioning of X, $j, k = 1, 2$. Let $\Delta_{1|2} = \Delta_{11} - \Delta_{12}\Delta_{22}^{-1}\Delta_{21}$, with $\Delta_{2|1}$ defined similarly by interchanging the subscripts. The matrix $\Delta_{1|2}$ is constructed in the same way as the covariance matrix for the conditional distribution of $X_1 \mid X_2$ when X is normally distributed, although here X is fixed. For $A \in \mathbb{R}^{p_1 \times p_1}$, define

$$M(A) = \eta A \eta^T \otimes \Omega_0^{-1} + \Omega \otimes \Omega_0^{-1} + \Omega^{-1} \otimes \Omega_0 - 2 I_{u_1(r-u_1)}.$$

Let $\hat{\beta}_1(\Gamma)$, $\hat{\beta}_1(\eta)$ and $\hat{\beta}_1(\beta_2)$ denote the maximum likelihood estimators of β_1 when Γ, η and β_2 are known. If a random vector w_n has the property that $n^{1/2}(w_n - \alpha)$ converges in distribution to a $N(0, A)$ then we will describe the asymptotic variance of w_n as $\mathrm{avar}(n^{1/2}w_n) = A$. The limiting distributions of $\hat{\beta}_1$ and $\hat{\beta}_2$ are stated in the following proposition; justification is given in Appendix 2.

PROPOSITION 1. *Under model (5),* $n^{1/2}\{\mathrm{vec}(\hat{\beta}_j) - \mathrm{vec}(\beta_j)\}$, $j = 1, 2$, *converge in distribution to normal random vectors with mean 0 and covariance matrices*

$$\begin{aligned}
\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\} &= \Delta_{1|2}^{-1} \otimes \Sigma_{\mathcal{E}_1} + (\eta^T \otimes \Gamma_0)M^{-1}(\Delta_{1|2})(\eta \otimes \Gamma_0^T), \\
&= \mathrm{avar}[n^{1/2}\,\mathrm{vec}\{\hat{\beta}_1(\Gamma)\}] + \mathrm{avar}[n^{1/2}\,\mathrm{vec}\{Q_{\mathcal{E}_1}\hat{\beta}_1(\eta)\}]; \\
\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_2)\} &= \{\Delta_{2|1} \otimes \Sigma^{-1} - (\Delta_{21}\eta^T \otimes \Gamma_0\Omega_0^{-1})M^{-1}(\Delta_{11})(\eta\Delta_{12} \otimes \Omega_0^{-1}\Gamma_0^T)\}^{-1}, \\
&= \{\Delta_{2|1} \otimes \Sigma^{-1} - (\Delta_{21} \otimes \Gamma_0\Omega_0^{-1}\Gamma_0^T)\mathrm{avar}\{\hat{\beta}_1(\beta_2)\}(\Delta_{12} \otimes \Gamma_0\Omega_0^{-1}\Gamma_0^T)\}^{-1},
\end{aligned}$$

where $\mathrm{avar}[n^{1/2}\,\mathrm{vec}\{\hat{\beta}_1(\Gamma)\}] = \Delta_{1|2}^{-1} \otimes \Sigma_{\mathcal{E}_1}$ *and* $\mathrm{avar}[n^{1/2}\,\mathrm{vec}\{Q_{\mathcal{E}_1}\hat{\beta}_1(\eta)\}]$ *is defined implicitly.*

Several comments on this proposition are in order. We first consider regressions in which $\Delta_{12} = 0$. This will arise when X_1 and X_2 are asymptotically uncorrelated or have been chosen by design to be orthogonal. Because X is non-random, this condition can always be forced without an inferential cost by replacing X_1 with $\hat{R}_{1|2}$. As discussed in §3·1, this replacement alters the definition of β_2 but does not alter the parameter of interest β_1. When $\Delta_{12} = 0$, $\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_2)\} = \Delta_{22} \otimes \Sigma^{-1}$, which is the same as the asymptotic covariance matrix for the estimator of β_2 from the standard model. And $\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\}$ reduces to the asymptotic covariance matrix for the full envelope estimator of β_1 in the model $\hat{R}_{Y|2} = \beta_1 X_1 + \varepsilon$. No longer requiring that $\Delta_{12} = 0$, we can carry out asymptotic inference for β_1 based on a par-

tial envelope by using the full envelope for β_1 in the model $\hat{R}_{Y|2} = \beta_1 \hat{R}_{1|2} + \varepsilon$. If we choose $\beta_1 = \beta$, so $X_1 = X$, β_2 is nil and $\Delta_{1|2} = \Delta$, then $\operatorname{avar}\{n^{1/2} \operatorname{vec}(\hat{\beta}_1)\}$ reduces to the asymptotic covariance matrix for the full envelope estimator of β derived by Cook et al. (2010).

We next turn to a comparison of the partial envelope estimator $\hat{\beta}_1$ of β_1 and the full envelope estimator $\hat{\beta}_{1\mathrm{e}}$, where the subscript e added to a statistic means computation based on the full envelope. The next proposition provides a comparison of $\operatorname{avar}\{n^{1/2} \operatorname{vec}(\hat{\beta}_1)\}$ and $\operatorname{avar}\{n^{1/2} \operatorname{vec}(\hat{\beta}_{1\mathrm{e}})\}$ at two extremes; justification is provided in Appendix 2.

PROPOSITION 2. *If* $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$ *then* $\operatorname{avar}\{n^{1/2} \operatorname{vec}(\hat{\beta}_1)\} \leq \operatorname{avar}\{n^{1/2} \operatorname{vec}(\hat{\beta}_{1\mathrm{e}})\}$. *If* $\mathcal{E}_\Sigma(\mathcal{B}) = \mathcal{E}_\Sigma(\mathcal{B}_1)$ *then* $\operatorname{avar}\{n^{1/2} \operatorname{vec}(\hat{\beta}_1)\} \geq \operatorname{avar}\{n^{1/2} \operatorname{vec}(\hat{\beta}_{1\mathrm{e}})\}$.

At one extreme, this proposition tells us that when the full envelope is $\mathbb{R}^r$, which is the situation that motivated this work, the covariance matrix for the partial envelope estimator will never be greater than the full envelope estimator, which is the same as the standard estimator. At the other extreme, if the full and partial envelopes are the same, then the partial envelope estimator of β_1 will never do better than the full envelope estimator of the same quantity. The following lemma will be helpful in developing some intuition for this conclusion.

LEMMA 1. *Let* $\mathcal{B}_2 = \operatorname{span}(\beta_2)$. *Then* $\mathcal{E}_\Sigma(\mathcal{B}) = \mathcal{E}_\Sigma(\mathcal{B}_1) + \mathcal{E}_\Sigma(\mathcal{B}_2)$.

This lemma says that the full envelope is the sum of the partial Σ-envelopes for $\mathcal{B}_1$ and $\mathcal{B}_2$. The dimension of $\mathcal{E}_\Sigma(\mathcal{B}_1) \cap \mathcal{E}_\Sigma(\mathcal{B}_2)$ can vary from 0 to the minimum of the two. But $\mathcal{E}_\Sigma(\mathcal{B}) = \mathcal{E}_\Sigma(\mathcal{B}_1)$ if and only if $\mathcal{E}_\Sigma(\mathcal{B}_2) \subseteq \mathcal{E}_\Sigma(\mathcal{B}_1)$. This means that $\mathcal{E}_\Sigma(\mathcal{B}_2)$ can contain information on $\mathcal{E}_\Sigma(\mathcal{B}_1)$. This information is not used by the partial envelope, but is used by the full envelope and consequently the full envelope estimator may have the smaller asymptotic variance.

Beyond the conclusions in Proposition 2, we were unable to develop practically ponderable conditions for characterizing when the asymptotic variance of the partial envelope estimator of

β_1 is less than that of the full envelope estimator of β_1. A straightforward course in practice is to simply compute and compare the asymptotic variances.

3·4. *Selecting dimension for envelopes*

The dimension u_1 of the partial envelope is essentially a model selection parameter, and standard methods can be used to aid in its choice. We briefly describe two methods in this section. The first is based on sequential hypothesis testing.

The hypothesis $u_1 = d$, $d < p_1$, can be tested by using the likelihood ratio statistic $\Lambda(d) = 2\{\widehat{L}(r) - \widehat{L}(d)\}$, where

$$\widehat{L}(d) = -(nr/2)\{1 + \log(2\pi)\} - (n/2)\log|\hat{P}_{\mathcal{E}_1} S_{R|2} \hat{P}_{\mathcal{E}_1}|_0 - (n/2)\log|\hat{Q}_{\mathcal{E}_1} S_{Y|2} \hat{Q}_{\mathcal{E}_1}|_0$$

denotes the maximum value of the likelihood for the partial envelope model with $u_1 = d$ (See Appendix A1). When $d = r$ the partial envelope model reduces to the standard model and thus

$$\widehat{L}(r) = -(nr/2)\{1 + \log(2\pi)\} - (n/2)\log|S_R|$$

is the value of the maximized log likelihood for the standard model. Following standard likelihood theory, under the null hypothesis $\Lambda(d)$ is distributed asymptotically as a chi-squared random variable with $p_1(r - d)$ degrees of freedom. The test statistic $\Lambda(d)$ can be used in a sequential scheme to choose u_1: Starting with $d = 0$ and using a common test level, choose u_1 to be the first hypothesized value that is not rejected.

The dimension of the partial envelope could also be determined by using an information criterion:

$$\widehat{u}_1 = \arg\min_d \{-2\widehat{L}(d) + h(n)g(d)\},$$

where $h(n)$ is equal to $\log n$ for Bayes information criterion and is equal to 2 for Akaike's information criterion, and $g(d)$ is the number of real parameters in the partial envelope model,

$$g(d) = r + d(r-d) + dp_1 + rp_2 + d(d+1)/2 + (r-d)(r-d+1)/2.$$

Subtracting $g(d)$ from the number of parameters $r + pr + r(r+1)/2$ for the standard model gives the degrees of freedom for $\Lambda(d)$ mentioned previously.

3·5. *Computing*

Computational algorithms for Grassmann optimization typically require that the objective function be written in terms of semi-orthogonal basis matrices rather than projection operators (Edelman et al., 1999; Liu et al., 2004). Since eigenvalues are invariant under cyclic permutations, (4) can be expressed equivalently as

$$\begin{aligned}\widehat{\mathcal{E}}_\Sigma(\mathcal{B}_1) &= \operatorname{span}\{\arg\min\,(\log|G^T S_{R|2} G| + \log|H^T S_{Y|2} H|)\} \\ &= \operatorname{span}\{\arg\min\,(\log|G^T S_{R|2} G| + \log|G^T S_{Y|2}^{-1} G|)\},\end{aligned}$$

where the minimization is over semi-orthogonal matrices $G \in \mathbb{R}^{r\times u_1}$ and (G, H) is an orthogonal matrix. We adapted Lippert's sg_min 2·4·1 computer code (www-math.mit.edu/lippert/sgmin.html) to perform this numerical optimization. That program offers several optimization methods including Newton–Raphson iteration on Grassmann manifolds with analytic first and numerical second derivatives of the objective function, and we have found it to be stable and reliable. However, like most numerical methods for optimizing nonlinear objective functions, it requires good starting values to facilitate convergence and avoid any lurking local optima. A standard way to deal with multiple local optima is to use Newton–Raphson iteration, beginning with a $\sqrt{n}$ consistent starting value. An estimator that is one Newton–Raphson iteration step away from a $\sqrt{n}$ consistent estimator may be sufficient because it is asymptot-

ically equivalent to the maximum likelihood estimator (see, for example, Small et al., 2000), although we prefer to always iterate until convergence. In the remainder of this section we describe methods for determining starting values. Equally important, our descriptions should also serve to highlight additional structure of the envelope model. Because these are new methods we describe them in the context of the full envelope, understanding that they apply equally to partial envelopes by replacing S_R and S_Y with $S_{R|2}$ and $S_{Y|2}$. We assume that $u = \dim\{\mathcal{E}_\Sigma(\mathcal{B})\}$ is known, and we use the notation of (5) to describe the coordinate model for the full envelope when β_2 is nil:

$$Y = \mu + \Gamma\eta X + \varepsilon, \;\; \Sigma = \Sigma_{\mathcal{E}} + \Sigma_{\mathcal{E}^\perp} = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T, \tag{6}$$

where $\Omega \in \mathbb{R}^{u\times u}$ and $\Omega_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are positive definite matrices that serve as coordinates of $\Sigma_{\mathcal{E}}$ and $\Sigma_{\mathcal{E}^\perp}$ relative to the basis matrices Γ for $\mathcal{E}_\Sigma(\mathcal{B})$ and its orthogonal complement.

The starting values that we use are functions of S_R, S_Y and the ordinary least squares estimator B of β. The asymptotic behavior of these statistics is summarized in the following lemma. Because they are standard moment estimators its proof seems straightforward and is omitted.

LEMMA 2. *The sample matrices B, S_Y and S_R are $\sqrt{n}$ consistent estimators of their population counterparts β, $\Sigma_Y = \Gamma(\Omega + \eta\Delta\eta^T)\Gamma^T + \Sigma_{\mathcal{E}^\perp}$ and $\Sigma = \Sigma_{\mathcal{E}} + \Sigma_{\mathcal{E}^\perp}$.*

We see from this lemma that the eigenvectors of both Σ_Y and Σ will be in either $\mathcal{E}_\Sigma(\mathcal{B})$ or its orthogonal complement, but from one of these matrices alone there is no way to tell in which subspace an eigenvector lies. However, we can tell by using Σ_Y and Σ together, which is the role of the partially maximized log likelihood:

LEMMA 3. *Under the envelope model (6),*

$$\mathcal{E}_\Sigma(\mathcal{B}) = \operatorname{span}\{\arg\min_{V(\Sigma)} (\log|G^T\Sigma G| + \log|G^T\Sigma_Y^{-1}G|)\},$$

where $\min_{V(\Sigma)}$ *means that the minimum is taken over all subsets of* u *eigenvectors of* Σ.

This lemma says that we can always find $\mathcal{E}_\Sigma(\mathcal{B})$ in the population by minimizing the log likelihood function over all subsets of u eigenvectors of Σ. Its proof is in Appendix 2. Based on this, we determine our first set of starting values G_0 as the best set of u eigenvectors of S_R,

$$G_0 = \arg\min_{V(S_R)} (\log|G^T S_R G| + \log|G^T S_Y^{-1} G|),$$

When the number of subsets is too large for this to be practical, we have used a sequential scheme that involves starting with a randomly selected subset of u eigenvectors of S_R and then updating each eigenvector in the subset from among those remaining, continuing for two or three iterations through the entire subset of u eigenvectors.

The starting value G_0 works well when the signal is sufficiently large, but may perform poorly when the signal is small or r is large. In those cases we have found it useful to update G_0 by using the ordinary squares estimator B of β and Krylov subspaces. First, project B onto $\mathrm{span}(G_0)$ to update B, $B_0 = P_{G_0} B$, and then form the updated estimator of Σ,

$$S_{R,0} = (U - FB_0^T)^T (U - FB_0^T)/n,$$

where U is the $n \times r$ matrix with rows $(Y - \bar{Y})^T$ and F is the $n \times p$ matrix with rows X^T. Next, let $K_0 = (B_0, S_{R,0}B_0, S_{R,0}^2 B_0, \ldots)$, where the column dimension of K_0 must be at least u. The new starting value G_1 then consists of the first u left singular vectors of K_0. The population rationale for the final step comes from Cook et al. (2007) who showed that there always exists an integer k, which is bounded by the number of eigenspaces of Σ, such that $\mathcal{E}_\Sigma(\mathcal{B}) = \mathrm{span}(\beta, \Sigma\beta, \ldots, \Sigma^{k-1}\beta)$.

To illustrate the behaviour of the starting value G_1 we simulated data from model (6) with $r = 100$ responses, $p = 3$ predictors and an envelope of dimension $u = 4$. Let $W_1 \in \mathbb{R}^{r \times u}$,

$W_2 \in \mathbb{R}^{u\times p}$ and $W_3 \in \mathbb{R}^{u\times u}$ be matrices of independent uniform $(0,1)$ random variables. Then the various matrices in the model were generated as follows $\Gamma = W_1(W_1^TW_1)^{-1/2}$, $\eta = W_2$, $\Omega = 0{\cdot}01 \times W_3W_3^T$, $\Omega_0 = 0{\cdot}64 \times I_{r-u}$, and $\mu = 0$. The predictors were also generated as vectors of independent uniform random variables, centered to have mean 0 in the sample. To gain a feeling for the sizes of the various matrices involved in this computation, we take the expectations E_w with respect to the matrices W_j used in their construction: $\mathrm{tr}\{\mathrm{E}_w(\Sigma_{\mathcal{E}})\} \approx 0{\cdot}05$, $\mathrm{tr}\{\mathrm{E}_w(\Sigma_{\mathcal{E}^\perp})\} \approx 61$ and $\mathrm{tr}\{\mathrm{E}_w(\Gamma\eta\Delta\eta^T\Gamma^T)\} \approx 0{\cdot}33$. From this we see that the variation $\Sigma_{\mathcal{E}}$ in the material part of Y will tend to be small relative to the variation $\Sigma_{\mathcal{E}^\perp}$ in the immaterial part of Y, while the signal is larger than $\Sigma_{\mathcal{E}}$ but is still small relative to $\Sigma_{\mathcal{E}^\perp}$. Recalling the discussion of §2, this is the kind of setting in which envelopes can provide substantial gains in efficiency.

To illustrate the advantages of the proposed starting values, we generated $n = 1000$ observations from the above simulation model and started iteration at G_1 and a randomly selected starting value $G^* = Z(Z^TZ)^{-1/2}$, where $Z \in \mathbb{R}^{r\times u}$ is a matrix of independent standard normal random variables. The two starting values converged to the same solution, the maximum angle between $\widehat{\mathcal{E}}_\Sigma(\mathcal{B})$ and $\mathcal{E}_\Sigma(\mathcal{B})$ being only about 6 degrees. However, starting at G^* required about twice the number of iterations to reach convergence as when starting at G_1, and the log likelihood increased about $8,000$ units when starting from G^*, but increased only 3 units when starting from G_1. We repeated this numerical experiment with $n = 300$ observations, which is a fairly small sample size in view of the number of responses. In that case, starting from G_1 converged to a solution $\widehat{\mathcal{E}}_\Sigma(\mathcal{B})$ that was only about 12 degrees away the true subspace $\mathcal{E}_\Sigma(\mathcal{B})$. The algorithm also converged when starting from G^*, but it reached a local solution that was about 87 degrees away from the true subspace. Generally, our experience indicates that random starts are not very helpful since they tend to reach local maxima.

4. Asymptotic approximations and the bootstrap

In this section we report a few of our results from a simulation study to investigate the accuracy of the asymptotic variance $\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\}$ presented in Proposition 1. We simulated data from model (5) with $r = 10$, $p = 10$, $p_1 = 1$, $\mu = 0$ $\eta_1 = 1$ and the elements of $\Gamma \in \mathbb{R}^{10\times 1}$ and $\beta_2 \in \mathbb{R}^{10\times 9}$ selected once at the outset as independent standard normal variables. For each sample, the elements in $X \in \mathbb{R}^{10}$ were generated as 0 or 10 each with probability $1/2$. The covariance matrix Σ had one small eigenvalue 0·0006 with corresponding eigenvector Γ, eight intermediate eigenvalues between 0·40 and 51, and one large eigenvalue of about 986. The actual variance of $\hat{\beta}_1$ was estimated as the sample variance of the estimates $\hat{\beta}_1^{(k)}$, $k = 1, \ldots, 200$, from 200 replications of the simulation scenario for each sample size. We also estimated the sample variance of $\hat{\beta}_1$ based on 200 residual bootstrap samples from one of the 200 replications. The results are shown in Fig. 2 for the four values of u that were used to construct the estimators. For clarity, we let u_0 denote the true value of u. In the simulation model $u_0 = 1$.

The vertical axis of each of the four panels in Fig. 2 is the standard deviation for one element of $\hat{\beta}_1$ and the horizontal axis is the sample size. The results shown in Fig. 2 illustrate the general conclusions that we reached from our simulation study. Because Γ corresponds to a relatively small eigenvalue, we expected that the asymptotic variability of the envelope estimator with $u = u_0 = 1$ would be much smaller than that of the standard estimator. That expectation is confirmed by the results shown in the first panel of Fig. 2. That plot also shows that $\mathrm{avar}(n^{1/2}\hat{\beta}_1)$ can give a very good approximation of the actual variance when $u = u_0$. The remaining panels in Fig. 2 show that the envelope estimator can still give substantial gains over the standard estimator when $u > u_0$. This typically happens when the estimated envelope avoids the larger eigenvalues of Σ, as was the case in the simulation. Nevertheless, when $u > u_0$ the actual variability of the envelope estimator can be substantially larger than $\mathrm{avar}(n^{1/2}\hat{\beta}_1)$. The residual bootstrap is a

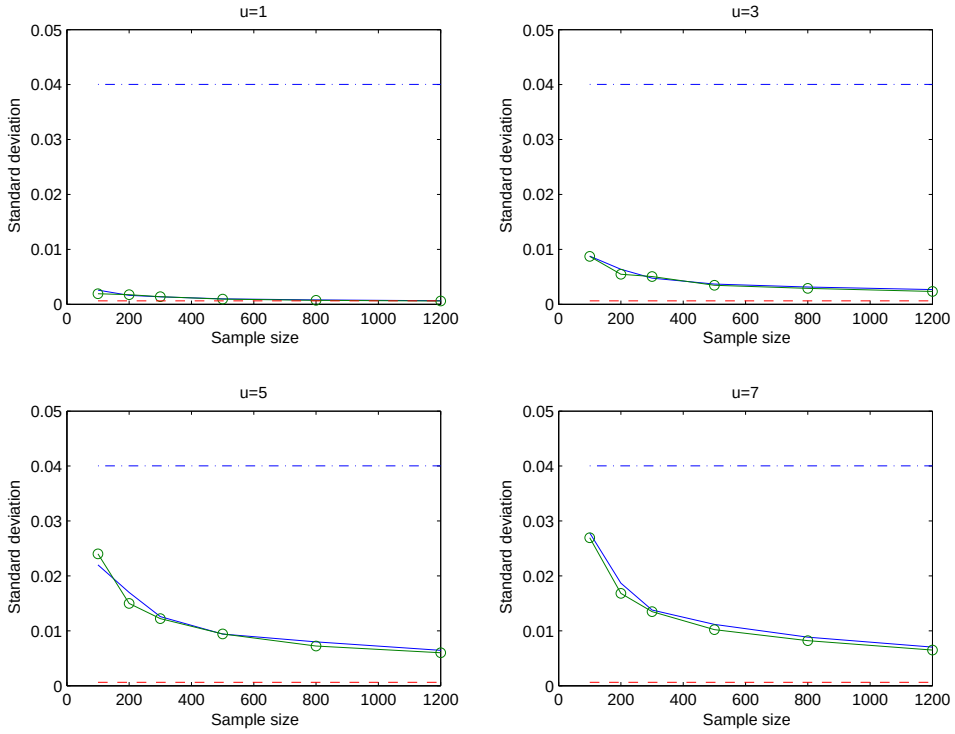

Fig. 2. Simulation results on the asymptotic standard deviation of an element of $\hat{\beta}_1$. The horizontal dashed line at about 0·04 marks the standard deviation of the standard model estimator and the dashed line just above the horizontal axis marks the asymptotic standard deviation of the envelope model estimator. The solid line corresponds to the estimated actual standard deviation of the envelope estimator and the line marked with ∘ corresponds to the bootstrap standard deviation.

reliable method for estimating the actual variance of $\hat{\beta}_1$, regardless of the relation between u and u_0.

We also studied how the error distribution might affect the performance of the envelope estimator. The simulation scenario was identical to that described for Fig. 2, except that ε was generated as $\Sigma^{1/2}\epsilon$, where the elements of ϵ were independent and identically distributed standard normal, $t_6/(3/2)^{1/2}$, $12^{1/2}\{U(0,1) - 0 \cdot 5\}$ or $(\chi^2_4 - 4)/\sqrt{8}$ random variables. The results shown in Fig. 3 indicate that the performance of the envelope estimator is quite robust.

5. EXAMPLE

This section is devoted to an example that illustrates aspects of partial envelopes. The dataset is from Johnson & Wichern (2007) and is on properties of pulp fibers and the paper made from

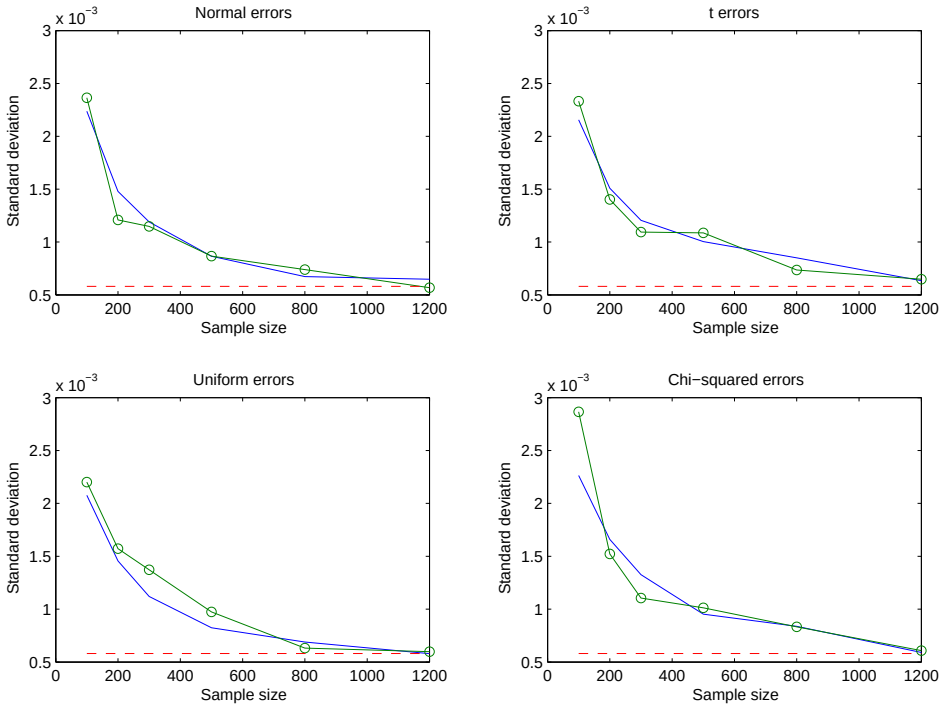

Fig. 3. Simulation results for four error distributions. The contents of the plots are as described for Fig. 2, except the standard deviation of the standard model estimator is not shown.

them. The reason for choosing this example is its richness reflecting multiple results within the same context. The data has 62 measurements on four paper properties: breaking length, elastic modulus, stress at failure and burst strength. The predictors are three properties of fiber: arithmetic fiber length, long fiber fraction and fine fiber fraction.

First we fitted an envelope model to all the predictors. Likelihood ratio testing suggested $u = 2$. The ratio of the asymptotic standard deviation from the standard model to that from the envelope model was computed for each element in β, the range is 0·98 to 1·10, with an average of 1·03. This suggests that we do not gain much efficiency by fitting the envelope. The reason is apparent from the estimated structure of Σ: the eigenvalues of $\hat{\Sigma}_{\mathcal{E}}$ are 4·9532 and 0·0143 while the eigenvalues for $\hat{\Sigma}_{\mathcal{E}^\perp}$ are 0·1007 and 0·0060. So the part of Y that is material to X is no less variable than the immaterial part, and not much efficiency is gained from enveloping β.

Next we fitted the partial envelope models to each column of β. We started with the fine fiber fraction. Likelihood ratio testing selected $u_1 = 1$. The asymptotic standard deviation ratios between the standard model and the partial envelope model for the elements in the third column of β are 63·59, 6·79, 10·40 and 7·49. Substantial reduction is thus achieved when attention is focused on fine fiber fraction since the part of Y that is material to this predictor is much less variable than the immaterial part. A close look at $\hat{\Sigma}$ reveals that $\hat{\Sigma}_{\mathcal{E}_1}$ has eigenvalue 0·0149 while $\hat{\Sigma}_{\mathcal{E}_1^\perp}$ has eigenvalues 11·0981, 0·1008 and 0·0070.

As we indicated in Section 2·4·2, the actual variance can be estimated by the bootstrap variance. A simulation with 200 bootstrap replicates was run to investigate the actual variance of $\hat{\beta}_1$. Under the partial envelope model, although the bootstrap standard deviations for the elements in $\hat{\beta}_1$ are 9·70, 2·29, 2·57 and 1·49 times as large as their asymptotic counterparts, they are still 6·56, 2·97, 4·05 and 5·03 times the size of the asymptotic standard deviations for the standard model.

Next the partial envelope was fitted to arithmetic length and we inferred that $u_1 = 0$. This means that with the other two predictors present, paper properties are invariant to the change in arithmetic length. The test of the hypothesis $u_1 = 0$ under the partial envelope model is equivalent to the F-test of the hypothesis $\beta_1 = 0$ under the standard model.

Finally, we applied the partial envelope model to the long fiber fraction. The estimated envelope had dimension two and it was only a small angle apart from the envelope model we fitted in the first place. The standard deviation ratios between the envelope model and partial envelope model for the second column of β are 0·9985, 0·9994, 0·9980 and 1·0046. This illustrates our statement in Proposition 2 that when $\mathcal{E}_\Sigma(\mathcal{B}) = \mathcal{E}_\Sigma(\mathcal{B}_1)$, the partial envelope model cannot outperform the envelope model. By Lemma 1, $\mathcal{E}_\Sigma(\mathcal{B})$ can be decomposed into $\mathcal{E}_\Sigma(\mathcal{B}_1) + \mathcal{E}_\Sigma(\mathcal{B}_2) + \mathcal{E}_\Sigma(\mathcal{B}_3)$, where $\mathcal{B}_i$ represents the space spanned by the i-th column of β.

Recall the dimensions of $\mathcal{E}_\Sigma(\mathcal{B})$, $\mathcal{E}_\Sigma(\mathcal{B}_1)$, $\mathcal{E}_\Sigma(\mathcal{B}_2)$ and $\mathcal{E}_\Sigma(\mathcal{B}_3)$ were inferred to be 2, 0, 2 and 1 respectively. Then $\mathcal{E}_\Sigma(\mathcal{B}_3)$ is forced to lie within $\mathcal{E}_\Sigma(\mathcal{B}_2)$, and the angle between the sample version of the two is around 8 degrees.

This example illustrates situations in which we will or will not expect to get significant reduction from fitting the partial envelope model. Basically, when $\hat{\Sigma}_{\mathcal{E}_1^\perp}$ has at least one large eigenvalue, massive reduction in variance is a typical result from applying the partial envelope model. But if a large eigenvalue is associated with $\hat{\Sigma}_{\mathcal{E}_1}$, we may achieve no noticeable reduction. In the application context, we found that partial envelopes significantly reduced the standard errors of the coefficients of fine fiber fraction.

While envelopes convert equivariantly under symmetric linear transformations of the response that commute with Σ, they are not equivariant for all linear transformations (Cook et al., 2010). Similarly, a partial envelope may not convert equivariantly under scale changes of the response and for this reason it may be advantageous to choose commensurate scales. Nevertheless, as illustrated in this example, useful results are often obtained using the original measurement scales.

ACKNOWLEDGEMENT

We are grateful to the editor and two referees for their insightful suggestion and comments that helped us improve the paper. This work was supported in part by a grant from the U.S. National Science Foundation.

APPENDIX 1: MAXIMUM LIKELIHOOD ESTIMATORS

As described in §3·2, $\mathcal{E}_\Sigma(\mathcal{B}_1)$ is the same as the full envelope in the model $\hat{R}_{Y|2} = \beta_1 \hat{R}_{1|2} + \varepsilon$. Following the derivation in §4·2 of Cook et al. (2010) with their Y and X replaced by $\hat{R}_{Y|2}$ and $\hat{R}_{1|2}$, $\widehat{\mathcal{E}}_\Sigma(\mathcal{B}_1)$

can be obtained by minimizing the following function over $\mathcal{S} \in \mathcal{G}(u_1, r)$,

$$\log|P_{\mathcal{S}} S_{R|2} P_{\mathcal{S}}|_0 + \log|Q_{\mathcal{S}} S_{Y|2} Q_{\mathcal{S}}|_0, \tag{A1}$$

where $S_{R|2} = U^T Q_{F_2} Q_{F^*} Q_{F_2} U/n$, $S_{Y|2} = U^T Q_{F_2} U/n$, $F^* = Q_{F_2} F_1/n$, F_1 is the $n \times p_1$ matrix with rows X_1^T, F_2 is the $n \times p_2$ matrix with rows X_2^T and U is the $n \times r$ matrix with rows $(Y - \bar{Y})^T$.

After getting $\hat{P}_{\mathcal{E}_1}$ from optimizing (*A*1), $\hat{\beta}_1 = \hat{P}_{\mathcal{E}_1} \tilde{\beta}_1$, where $\tilde{\beta}_1$ is the ordinary maximum likelihood estimator of the coefficients for X_1. Let $\hat{\Gamma}$ be a semi-orthogonal basis for $\widehat{\mathcal{E}}_\Sigma(\mathcal{B}_1)$. Then $\hat{\eta} = \hat{\Gamma}^T \hat{\beta}_1$, $\hat{\Omega} = \hat{\Gamma}^T S_{R|2} \hat{\Gamma}$, $\hat{\Omega}_0 = \hat{\Gamma}_0^T S_{Y|2} \hat{\Gamma}_0$, $\hat{\Sigma}_1 = \hat{\Gamma}\hat{\Omega}\hat{\Gamma}^T$ and $\hat{\Sigma}_2 = \hat{\Gamma}_0 \hat{\Omega}_0 \hat{\Gamma}_0^T$. Having derived $\hat{\beta}_1$, the maximum likelihood estimator of β_2 is $\hat{\beta}_2 = (U - F_1 \hat{\beta}_1^T)^T F_2 (F_2^T F_2)^{-1}$. Substitute all the above estimators into the log likelihood function, with a fixed dimension of the envelope d, the maximized log likelihood is equal to

$$\widehat{L}(d) = -(nr/2)\{1 + \log(2\pi)\} - (n/2)\log|\hat{P}_{\mathcal{E}_1} S_{R|2} \hat{P}_{\mathcal{E}_1}|_0 - (n/2)\log|\hat{Q}_{\mathcal{E}_1} S_{Y|2} \hat{Q}_{\mathcal{E}_1}|_0.$$

APPENDIX 2: PROOFS

Proof of Proposition 1. Because of the over-parameterization in (5), we use Proposition 4·1 in Shapiro (1986) to derive the asymptotic distributions. For simplicity, we denote $\mathrm{vec}(\beta_2)$, $\mathrm{vec}(\eta)$, $\mathrm{vec}(\Gamma)$, $\mathrm{vech}(\Omega)$ and $\mathrm{vech}(\Omega_0)$ as ϕ_0, ϕ_1, ϕ_2, ϕ_3 and ϕ_4, respectively, and then we combine them into the vector $\phi = (\phi_0^T, \phi_1^T, \phi_2^T, \phi_3^T, \phi_4^T)^T$. Here vec and vech are the "vector" and "vector-half" operators defined by Henderson & Searle (1979). Let

$$h(\phi) = \begin{pmatrix} \mathrm{vec}(\beta_2) \\ \mathrm{vec}(\beta_1) \\ \mathrm{vech}(\Sigma) \end{pmatrix} = \begin{pmatrix} \mathrm{vec}(\beta_2) \\ \mathrm{vec}(\Gamma\eta) \\ \mathrm{vech}(\Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T) \end{pmatrix} \equiv \begin{pmatrix} h_0(\phi) \\ h_1(\phi) \\ h_2(\phi) \end{pmatrix}.$$

Then $n^{1/2}(\hat{h}-h)$ converges in distribution to $N(0,S_0)$, where $S_0 = H(H^TJH)^\dagger H^T$ and $H = (\partial h_i/\partial\phi_j^T)$ is the gradient matrix ($i=0,1,2$; $j=0,1,2,3,4$),

$$H = \begin{pmatrix} I_{rp_2} & 0 & 0 & 0 & 0 \\ 0 & I_{p_1}\otimes\Gamma & \eta^T\otimes\Gamma_0 & 0 & 0 \\ 0 & 0 & 2C_r(\Gamma\Omega\otimes\Gamma_0 - \Gamma\otimes\Gamma_0\Omega_0) & C_r(\Gamma\otimes\Gamma)E_{u_1} & C_r(\Gamma_0\otimes\Gamma_0)E_{(r-u_1)} \end{pmatrix}.$$

The Fisher information for $\{\mathrm{vec}(\beta_2)^T, \mathrm{vec}(\beta_1)^T, \mathrm{vech}(\Sigma)^T\}^T$ in the standard model is

$$J = \begin{pmatrix} \Delta_{22}\otimes\Sigma^{-1} & \Delta_{21}\otimes\Sigma^{-1} & 0 \\ \Delta_{12}\otimes\Sigma^{-1} & \Delta_{11}\otimes\Sigma^{-1} & 0 \\ 0 & 0 & \frac{1}{2}E_r^T(\Sigma^{-1}\otimes\Sigma^{-1})E_r \end{pmatrix},$$

where $C_r \in \mathbb{R}^{r(r+1)/2\times r^2}$ and $E_r \in \mathbb{R}^{r^2\times r(r+1)/2}$ provide the contraction and expansion matrices for the vec and vech operators: for any symmetric $r\times r$ matrix A, $\mathrm{vech}(A) = C_r\mathrm{vec}(A)$ and $\mathrm{vec}(A) = E_r\mathrm{vech}(A)$.

The asymptotic variances for $\hat{\beta}_1$ and $\hat{\beta}_2$ are the first two diagonal blocks of S_0. After some matrix multiplication, we have

$$\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\} = \Delta_{1|2}^{-1}\otimes\Sigma_{\mathcal{E}_1} + (\eta^T\otimes\Gamma_0)M^{-1}(\Delta_{1|2})(\eta\otimes\Gamma_0^T),$$

$$\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_2)\} = \{\Delta_{2|1}\otimes\Sigma^{-1} - (\Delta_{21}\eta^T\otimes\Gamma_0\Omega_0^{-1})M^{-1}(\Delta_{11})(\eta\Delta_{12}\otimes\Omega_0^{-1}\Gamma_0^T)\}^{-1}. \quad \square$$

Proof of Proposition 2. When $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$, the full envelope model is the same as the standard multivariate linear model. Then $\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_{1\mathrm{e}})\} = \Delta_{22}^{-1}\otimes\Sigma$, which is the upper left $rp_2\times rp_2$ block of J^{-1}. From the proof of Proposition 1, $\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\}$ is the upper left $rp_2\times rp_2$ block of $H(H^TJH)^\dagger H^T$. Since $J^{\frac{1}{2}}\{J^{-1} - H(H^TJH)^\dagger H^T\}J^{\frac{1}{2}} = Q_{J^{\frac{1}{2}}H} \geq 0$, $J^{-1} \geq H(H^TJH)^\dagger H^T$. So $\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_{1\mathrm{e}})\} \geq \mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\}$.

If $\mathcal{E}_\Sigma(\mathcal{B}) = \mathcal{E}_\Sigma(\mathcal{B}_1)$, the full and partial envelope models have the same envelope, and then the parameters Γ, Γ_0, Ω, Ω_0 are the same in both models. We write η as $(\eta_1,\ \eta_2)$. Since $\mathrm{vec}(\beta_1) =$

$[(I_{p_1},\ 0) \otimes I_r]\,\mathrm{vec}(\beta)$,

$$\begin{aligned}\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_{1\mathrm{e}})\} &= (I_{p_1},\ 0) \otimes I_r \mathrm{avar}[n^{1/2}\,\mathrm{vec}(\hat{\beta})](I_{p_1},\ 0)^T \otimes I_r \\ &= \Delta_{11}^{-1} \otimes \Gamma\Omega\Gamma^T + (\eta_1^T \otimes \Gamma_0) M_{(\Delta)}^{-1} (\eta_1 \otimes \Gamma_0^T),\end{aligned}$$

where $M_{(\Delta)} = \eta\Delta\eta^T \otimes \Omega_0^{-1} + \Omega \otimes \Omega_0^{-1} + \Omega^{-1} \otimes \Omega_0 - 2I_u \otimes I_{r-u}$. And

$$\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\} = \Delta_{11}^{-1} \otimes \Gamma\Omega\Gamma^T + (\eta_1^T \otimes \Gamma_0) M_{\Delta_{11}}^{-1} (\eta_1 \otimes \Gamma_0^T),$$

where $M_{\Delta_{11}} = \eta_1\Delta_{11}\eta_1^T \otimes \Omega_0^{-1} + \Omega \otimes \Omega_0^{-1} + \Omega^{-1} \otimes \Omega_0 - 2I_{u_1} \otimes I_{r-u_1}$. As stated in §3·3, we assume $\Delta_{12} = 0$ without loss of generality. Then $\eta\Delta\eta^T = \eta_1\Delta_{11}\eta_1^T + \eta_2\Delta_{22}\eta_2^T \geq \eta_1\Delta_{11}\eta_1^T$, so we have $\mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_1)\} \geq \mathrm{avar}\{n^{1/2}\,\mathrm{vec}(\hat{\beta}_{1\mathrm{e}})\}$ because the other terms are the same. □

Proof of Lemma 3. Let Γ be a semi-orthogonal basis for $\mathcal{E}_\Sigma(\mathcal{B})$, and let G be an $r \times u$ semi-orthogonal matrix with G_0 a basis of the orthogonal complement of its span. Then we have

$$\log|G^T\Sigma G| + \log|G^T\Sigma_Y^{-1}G| = \log|G^T\Sigma G| + \log|G_0^T\Sigma_Y G_0| + \log|\Sigma_Y^{-1}|.$$

Since $\Sigma_Y = \Sigma + \Gamma\eta\Delta\eta^T\Gamma^T$,

$$\begin{aligned}&\log|G^T\Sigma G| + \log|G_0^T\Sigma_Y G_0| \\ &= \log|G^T\Sigma G| + \log|G_0^T\Sigma G_0| + \log|I_{r-u} + G_0^T\Gamma\eta\Delta^{-\frac{1}{2}}(G_0^T\Sigma G_0^T)^{-1}\Delta^{-\frac{1}{2}}\eta^T\Gamma^T G_0| \\ &= \log|\Sigma| + \log|I_{r-u} + G_0^T\Gamma\eta\Delta^{-\frac{1}{2}}(G_0^T\Sigma G_0)^{-1}\Delta^{-\frac{1}{2}}\eta^T\Gamma^T G_0|.\end{aligned}$$

The objective function takes its minimum at $\mathrm{span}(G) = \mathrm{span}(\Gamma)$, because it makes the second term zero, otherwise that term will be positive. As Γ is a subset of u eigenvectors of Σ, we can search all the subsets of u eigenvectors of Σ to get the minima. □

REFERENCES

CONWAY, J. (1990). *A Course in Functional Analysis*. New York: Springer.

COOK, R., LI, B. & CHIAROMONTE, F. (2007). Dimension reduction in regression without matrix inversion. *Biometrika* **94**, 569.

COOK, R., LI, B. & CHIAROMONTE, F. (2010). Envelope models for parsimonious and efficient multivariate linear regression. *Statist. Sinica* **20**, 927–1010.

COOK, R. & WEISBERG, S. (1982). *Residuals and Influence in Regression*. New York: Chapman and Hall.

EDELMAN, A., ARIAS, T. & SMITH, S. (1999). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20**, 303–353.

HENDERSON, H. & SEARLE, S. (1979). Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics. *Can. J. Statist.* **7**, 65–81.

JOHNSON, R. & WICHERN, D. (2007). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall.

LIU, X., SRIVASTAVA, A. & GALLIVAN, K. (2004). Optimal linear representations of images for object recognition. *IEEE Trans. Pattern. Anal.* **26**, 662–666.

SHAPIRO, A. (1986). Asymptotic theory of overparameterized structural models. *J. Am. Statist. Assoc.* **81**, 142–149.

SMALL, C., WANG, J. & YANG, Z. (2000). Eliminating multiple root problems in estimation. *Statist. Sci.* **15**, 313–332.