Biometrika (2011), **98**, 1, *pp*. 1–29

Advance Access publication on 30 June 2011
Printed in Great Britain

Scaled Envelopes: Scale Invariant and Efficient Estimation in Multivariate Linear Regression

BY R. DENNIS COOK

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church Street SE, Minneapolis, Minnesota, 55455

dennis@stat.umn.edu

AND ZHIHUA SU

Department of Statistics, University of Florida, 102 Griffin-Floyd Hall, Gainesville, Florida, 32611

zhihuasu@stat.ufl.edu

SUMMARY

Efficient estimation of the regression coefficients is a fundamental problem in multivariate linear regression. The envelope model proposed by Cook et al. (2010) was shown to have the potential to achieve substantial efficiency gains by accounting for linear combinations of the response vector that are essentially immaterial to coefficient estimation. This requires in part that the distribution of those linear combinations be invariant to changes in the non-stochastic predictor vector. However, inference based on an envelope is not invariant or equivariant under rescaling of the responses, tending to limit application to responses that are measured in the same or similar units. The efficiency gains promised by envelopes often cannot be realized when the responses are measured in different scales. To overcome this limitation and broaden the scope of envelope methods, we propose a scaled version of the envelope model, which preserves the potential of the original envelope methods to increase efficiency and is invariant to scale changes. Likelihood-based estimators are derived and theoretical properties of the estimators are studied in various circumstances. It is shown that estimating appropriate scales for the responses can produce substantial efficiency gains when the original envelope model offers none. Simulations and an example are given to support the theoretical claims.

Some key words: Dimension reduction, Envelope model, Reducing subspace, Similarity transformation.

1. INTRODUCTION

The standard multivariate linear regression model can be written as

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

where $Y \in \mathbb{R}^r$ is the stochastic response vector, $X \in \mathbb{R}^p$ denotes the vector of non-stochastic predictors centered at 0 in the sample, the error vector $\varepsilon \in \mathbb{R}^r$ has mean 0 and covariance matrix $\Sigma > 0$, $\alpha \in \mathbb{R}^r$ is an unknown vector of intercepts and $\beta \in \mathbb{R}^{r \times p}$ is an unknown matrix of regression coefficients. If X is stochastic, X and Y have a joint distribution, but we still condition on the observed values of X since the predictors are ancillary under model (1). The jth row of the ordinary least squares estimator of β is equal to the coefficient vector from the ordinary least squares regression of the jth element of Y on X ($j = 1, \ldots, r$). Stochastic relationships among the elements of Y are not used in this standard estimator of β. However, the relationships among the elements of Y play a central role in envelope estimation.

The envelope model proposed by Cook et al. (2010) has the potential to yield an estimator of β that is substantially less variable than the ordinary least squares estimator. In many datasets, the distribution of some linear combinations of Y may be invariant to changes in X and uncorrelated with a complementary set of linear combinations. When this occurs, Y can be divided into a material part, whose distribution depends on X, and an immaterial part, whose distribution does not depend on X. The immaterial part of Y contains no information on β, but it induces extraneous variation into the estimation of β via model (1). The envelope model was designed to account for the immaterial response variation, resulting in an estimator of β that may be more efficient than the standard estimator and substantially more efficient when the immaterial variation is substantially greater than the material variation in Y. The envelope estimator of β reduces to the ordinary least squares estimator when there is no immaterial variation in Y.

We define a scale transformation of the response to be of the form $Y \longmapsto AY$, where $A \in \mathbb{R}^{r\times r}$ is a non-singular diagonal matrix. Like principal component analysis, partial least squares and other methods, the envelope model is not invariant or equivariant under scale transformations: if we perform a scale transformation on the responses, the envelope estimator of the new β could reduce to the ordinary least squares estimator. This property tends to limit application of the envelope model to responses that are in the same or similar scales.

In this article we propose a scaled envelope model, which is scale-invariant and can achieve efficiency gains beyond those possible from the original envelope model. This is accomplished by incorporating a scaling matrix into the model and so scale transformations are considered during estimation. Scaling is a common practice in chemometrics and in many other applications.

The following notations and definitions will be used in our discussion. For positive integers a and b, $\mathbb{R}^{a\times b}$ denotes the class of all $a \times b$ matrices. If $A \in \mathbb{R}^{a\times b}$, then $\mathrm{span}(A)$ is the subspace spanned by the columns of A. For a subspace $\mathcal{S}$, $\mathcal{S}^{\perp}$ stands for its orthogonal complement. With $A \in \mathbb{R}^{a\times a}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^{a}$, $A\mathcal{S} = \{As : s \in \mathcal{S}\}$. The spectral norm of a matrix of A is denoted by $\|A\|$ and the Moore–Penrose inverse of A is denoted by $A^{\dagger}$. For a positive definite matrix $\Delta \in \mathbb{R}^{a\times a}$, the inner product in $\mathbb{R}^a$ defined by $\langle x_1, x_2\rangle_\Delta = x_1^T \Delta x_2$ is called the Δ inner product, where x_1 and x_2 are two arbitrary vectors in $\mathbb{R}^a$. The symbol $P_{A(\Delta)}$ is a projection operator onto A or $\mathrm{span}(A)$ in the Δ inner product if A is a space or a matrix, and $P_{A(\Delta)} = A(A^T\Delta A)^\dagger A^T\Delta$ if A is a matrix. We use $Q_{A(\Delta)} = I - P_{A(\Delta)}$. Projection operators employing the identity inner product are written as P_A, i.e., $P_A = P_{A(I)}$, and $Q_A = I - P_A$. The notation $\sim$ means identically distributed, and $\otimes$ stands for the Kronecker product.

2. ENVELOPE MODEL

Following Cook et al. (2010), let $\mathcal{S}$ be a subspace of $\mathbb{R}^r$ with the properties that (i) $Q_\mathcal{S}Y \mid X \sim Q_\mathcal{S}Y$, and (ii) $P_\mathcal{S}Y$ is uncorrelated with $Q_\mathcal{S}Y$ given X. Condition (i) indicates that $Q_\mathcal{S}Y$ carries no marginal information about β, and condition (ii) requires that $Q_\mathcal{S}Y$ does not carry information about β through its conditional correlation with $P_\mathcal{S}Y$. Let $\mathcal{B} = \mathrm{span}(\beta)$. Conditions (i) and (ii) are equivalent to

$$(a)\ \mathcal{B} \subseteq \mathcal{S}, \quad (b)\ \Sigma = P_\mathcal{S}\Sigma P_\mathcal{S} + Q_\mathcal{S}\Sigma Q_\mathcal{S}, \tag{2}$$

where $P_\mathcal{S}\Sigma P_\mathcal{S} = \mathrm{var}(P_\mathcal{S}Y)$ and $Q_\mathcal{S}\Sigma Q_\mathcal{S} = \mathrm{var}(Q_\mathcal{S}Y)$. Following standard terminology in the literature on invariant subspaces and functional analysis (Conway, 1990), the decomposition of Σ shown in (2b) is equivalent to requiring that $\mathcal{S}$ be a reducing subspace of Σ, although this notion of reduction is incompatible with how reduction is usually understood in statistics. The Σ-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ and by the abbreviated version $\mathcal{E}$ if it appears in a subscript, is defined as the intersection of all $\mathcal{S} \subseteq \mathbb{R}^r$ that satisfies condition (2), and thus $\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$ is the subspace of minimal dimension that reduces Σ and contains $\mathcal{B}$. To describe this structure succinctly, we refer to $P_\mathcal{E}Y$ as the part of Y that is material to the estimation of β, and to $Q_\mathcal{E}Y$ as the part of Y that is immaterial to the estimation of β. We call (1) the ordinary envelope model when conditions (2) are imposed. We also refer to it as the envelope model when there is no chance of confusing it with the scaled envelope model of the next section.

Let u denote the dimension of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, let $\Gamma \in \mathbb{R}^{r\times u}$ be an orthogonal basis of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$, and let $\Gamma_0 \in \mathbb{R}^{r\times(r-u)}$ be an orthogonal basis of $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$. The coordinate form of an envelope model can then be written as

$$Y = \alpha + \Gamma\eta X + \varepsilon, \quad \Sigma = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T, \tag{3}$$

where the coefficients $\beta = \Gamma\eta$. The positive definite matrix $\Omega = \mathrm{var}(\Gamma^T Y) \in \mathbb{R}^{u\times u}$ represents the variation in the material part of Y; similarly, $\Omega_0 = \mathrm{var}(\Gamma_0^T Y) \in \mathbb{R}^{(r-u)\times(r-u)}$ represents the variation in the immaterial part. When $u = r$, $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}) = \mathbb{R}^r$, the envelope model reduces to the standard model and there is no gain in efficiency. However, substantial efficiency gains can be obtained when $\|\Gamma_0\Omega_0\Gamma_0^T\| = \|\Omega_0\| \gg \|\Gamma\Omega\Gamma^T\| = \|\Omega\|$.

The parameters in (3) are estimated by maximizing a normal likelihood function. Let $\widetilde{\Sigma}_Y$, $\widetilde{\beta}$ and $\widetilde{\Sigma}_{\mathrm{res}}$ denote the sample covariance matrix of Y, the least squares estimator of β, and the sample covariance matrix of the residuals from the least squares regression of Y on X. The estimator of the envelope subspace is then the span of $\arg\min\{\log|\Gamma^T\widetilde{\Sigma}_{\mathrm{res}}\Gamma| + \log|\Gamma^T\widetilde{\Sigma}_Y^{-1}\Gamma|\}$, where the minimization is over the $r \times u$ Grassmannian (Cook et al., 2010). Let $\widehat{\Gamma}$ be a basis of the estimated envelope subspace. The envelope estimators of the regression coefficients and the error covariance matrix are then $\widehat{\beta} = P_{\widehat{\Gamma}}\widetilde{\beta}$ and $\widehat{\Sigma} = P_{\widehat{\Gamma}}\widetilde{\Sigma}_{\mathrm{res}}P_{\widehat{\Gamma}} + Q_{\widehat{\Gamma}}\widetilde{\Sigma}_Y Q_{\widehat{\Gamma}}$. The forms of the estimators are consistent with the conditions in (2).

Figure 1 provides a graphical illustration of the working mechanism of the envelope model. In both panels, the two ellipses represent two populations. The predictor $X \in \mathbb{R}^1$ is an indicator variable taking values 0 or 1 to denote the different populations, Y_1 and Y_2 are two responses representing two characteristics of the populations, and β is the difference between the two population means. The left panel represents the analysis under the standard model. For inference on

β_2, the second element of β, a data point y is directly projected onto the Y_2 axis following the dashed line marked A. The two curves in the left panel stand for the two projected distributions from the two populations. There is considerable overlap between the two projected distributions, so it may take a large sample size to infer that $\beta_2 \neq 0$ in a least squares analysis. The right panel presents the analysis under the envelope model. Cook et al. (2010) proved that $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is spanned by some subset of the eigenvectors of Σ. In this case, the eigenvector corresponding to the smaller eigenvalue of Σ provides all the material information, since the distribution of Y does not depend on X in the direction of $\mathcal{E}_{\mathbf{\Sigma}}^{\perp}(\mathcal{B})$, which corresponds to the other eigenvector of Σ and to the immaterial information. So $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is spanned by the second eigenvector of Σ and $u = 1$. For inference on β_2 under the envelope model, a data point y is first projected onto $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ to remove the immaterial information $Q_\Gamma y$ and simultaneously extract the material information $P_\Gamma y$, which is then projected onto the Y_2 axis following the dashed lines marked B. The two curves at the bottom stand for the projected distributions for the two populations, which are now well separated. This indicates that by accounting for the immaterial information, the envelope model achieves substantial efficiency gains compared to the standard model.

3. SCALED ENVELOPE MODEL

3·1. *Motivation*

The ordinary envelope model (3) is not invariant or equivariant under linear transformations of the response. In particular, suppose that we rescale Y by multiplication by a non-singular diagonal matrix A. Let $Y_N = AY$ denote the new response, let $\widehat{\beta}$ and $\widehat{\Sigma}$ denote the estimators of β and Σ based on the envelope model for Y on X, and let $\widehat{\beta}_N$ and $\widehat{\Sigma}_N$ denote the estimators of β and Σ based on the envelope model for Y_N on X. Then we do not generally have invariance, i.e., $\widehat{\beta}_N = \widehat{\beta}$, $\widehat{\Sigma}_N = \widehat{\Sigma}$, or equivariance, i.e., $\widehat{\beta}_N = A\widehat{\beta}$, $\widehat{\Sigma}_N = A\widehat{\Sigma}A$. In fact, the dimension of

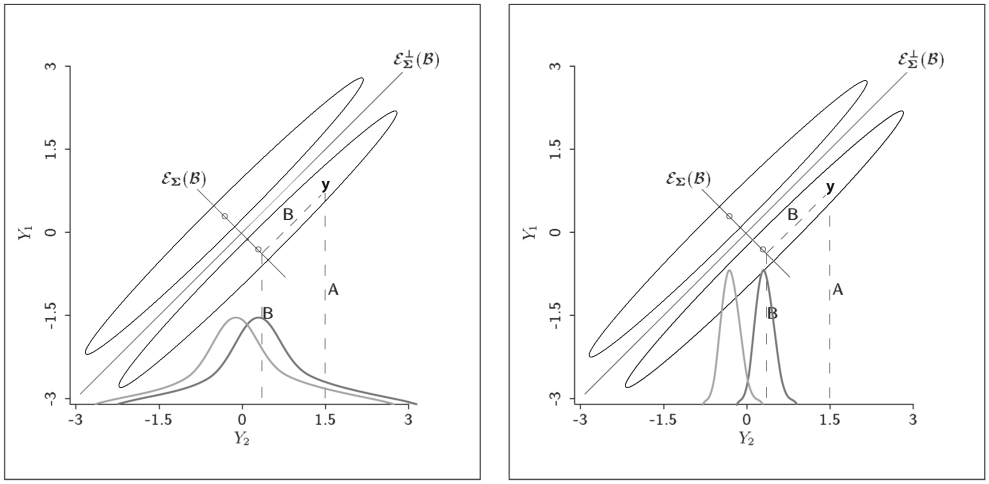

Fig. 1: Left panel: Inference on β_2 under the standard model. Right panel: Inference on β_2 under the envelope model.

the envelope subspace may change because of the transformation. We illustrate this using the example in Fig. 1. Suppose we multiply Y_2 by 2 and leave Y_1 unchanged, so A is a 2×2 diagonal matrix with diagonal elements 1 and 2. The distribution of $AY \mid X$ is displayed in Fig. 2. We denote the two eigenvectors of the new covariance matrix Σ_N as v_1 and v_2 and let $\mathcal{B}_N = \mathrm{span}(\beta_N)$ as marked in the left panel. Since $\mathcal{B}_N$ aligns with neither v_1 nor v_2, the envelope is two dimensional: $\mathcal{E}_{\mathbf{\Sigma}_N}(\mathcal{B}_N) = \mathbb{R}^2$. In this case, all linear combinations of Y are material to the regression, the envelope model is the same as the standard model and no efficiency gains are achieved.

The scaled envelope model as described formally in §3·2 seeks a rescaling that converts Fig. 2 to Fig. 1, performs the envelope estimation as in the right panel of Fig. 1, and then transforms the estimators back to the original scales, which is the scale in Fig. 2. This process results in the material part of Y being represented as $AP_\Gamma A^{-1}Y$, while it is represented as $P_\Gamma Y$ in an envelope analysis. In linear algebra, the transformation matrices $AP_\Gamma A^{-1}$ and P_Γ are said to

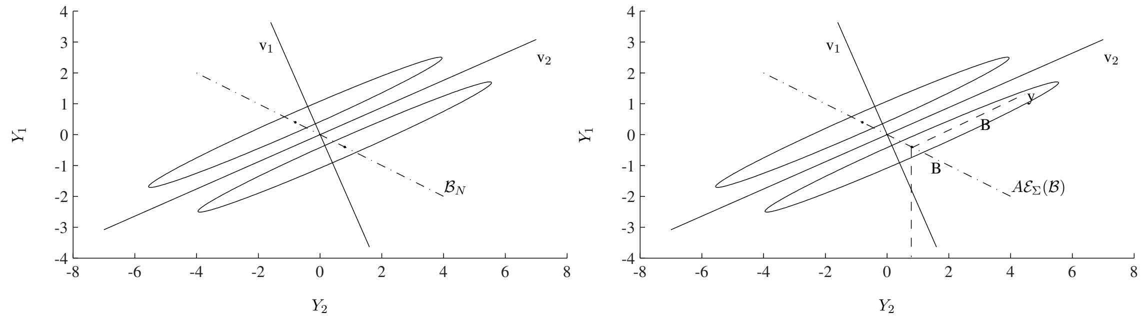

Fig. 2: Left panel: Example of the dimension of the envelope subspace changing under response rescaling. Right panel: Inference on β_2 under the scaled envelope model.

be similar: an $s \times s$ matrix M is similar to an $s \times s$ matrix N if there exists an $s \times s$ non-singular matrix T such that $N = TMT^{-1}$ (e.g., Harville, 2008). When M represents a linear transformation from an s-dimensional linear space $\mathcal{V}$ to $\mathcal{V}$, N is the matrix representation of the same linear transformation but under another basis of $\mathcal{V}$, and T^{-1} is the matrix representation of the change of basis. Therefore the process $AP_\Gamma A^{-1}$ is the same as treating A^{-1} as a similarity transformation to represent P_Γ in original coordinate system as $AP_\Gamma A^{-1}$. This process can be represented by the two line segments marked B in the right panel of Fig. 2. Additional discussion is given in §4·2.

This process also has another interpretation. As $AP_\Gamma A^{-1} = P_{A\Gamma(A^{-2})}$, the first line segment marked B in the right panel of Fig. 2 can also be considered as the projection onto the space spanned by $A\Gamma$ but in the A^{-2} inner product. In other words, the scaled envelope first projects the data onto $A\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ in the A^{-2} inner product. After this projection, the data point is projected onto the Y_2 axis in the original scales, as represented by the second line segment marked B in Fig. 2. Again, the projected distributions for the two populations have a very good separation, which illustrates the efficiency gains obtained by using scaled envelopes.

From the previous discussion, we notice that $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ can be very different after the response transformation, even the dimension of $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ can change. However, $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B})$ is equivariant under orthogonal transformations $Y \to \Psi Y$ of the response, where Ψ is an orthogonal matrix. In this case $\mathcal{E}_{\boldsymbol{\Sigma}_N}(\mathcal{B}_N) = \Psi\mathcal{E}_{\boldsymbol{\Sigma}}(\mathcal{B})$, where $\Sigma_N = \Psi\Sigma\Psi$ is the new error covariance matrix, and $\mathcal{B}_N = \mathrm{span}(\beta_N)$ with $\beta_N = \Psi\beta$ being the new regression coefficients.

3·2. *Model Formulation*

To represent a rescaling formally, we introduce a diagonal matrix $\Lambda =$diag$\{1, \lambda_2, \ldots, \lambda_r\} \in \mathbb{R}^{r\times r}$ with $\lambda_i > 0$ for $i = 2, \ldots, r$, such that $Y_N = \Lambda^{-1}Y$ follows an envelope model with the dimension of the envelope subspace $\mathcal{E}_{\Lambda^{-1}\boldsymbol{\Sigma}\Lambda^{-1}}(\Lambda^{-1}\mathcal{B})$ equal to u. Consequently, $\Lambda^{-1}\mathcal{B} \subseteq \mathrm{span}(\Gamma)$, and $\Lambda^{-1}\Sigma\Lambda^{-1} = P_\Gamma\Lambda^{-1}\Sigma\Lambda^{-1}P_\Gamma + Q_\Gamma\Lambda^{-1}\Sigma\Lambda^{-1}Q_\Gamma$, where $\Gamma \in \mathbb{R}^{r\times u}$ is now an orthogonal basis of $\mathcal{E}_{\Lambda^{-1}\boldsymbol{\Sigma}\Lambda^{-1}}(\Lambda^{-1}\mathcal{B})$, and $\Gamma_0 \in \mathbb{R}^{r\times(r-u)}$ is a completion of Γ.

The coordinate form of the scaled envelope model is then

$$Y = \alpha + \Lambda\Gamma\eta X + \epsilon, \;\; \Sigma = \Lambda\Gamma\Omega\Gamma^T\Lambda + \Lambda\Gamma_0\Omega_0\Gamma_0^T\Lambda. \tag{4}$$

The coefficients $\beta = \Lambda\Gamma\eta$, where $\eta = \Gamma^T\Lambda^{-1}\beta \in \mathbb{R}^{u\times p}$, and the positive definite matrices $\Omega = \mathrm{var}(\Gamma^T\Lambda^{-1}Y) = \Gamma^T\Lambda^{-1}\Sigma\Lambda^{-1}\Gamma \in \mathbb{R}^{u\times u}$ and $\Omega_0 = \mathrm{var}(\Gamma_0^T\Lambda^{-1}Y) = \Gamma_0^T\Lambda^{-1}\Sigma\Lambda^{-1}\Gamma_0 \in \mathbb{R}^{(r-u)\times(r-u)}$. Setting the first element of Λ to 1 is necessary for the scaling parameters to be identifiable. Otherwise we can multiply Λ by an arbitrary constant c and multiply η by its reciprocal $1/c$. Computation is facilitated when Λ is identifiable, but this is not necessary for efficient estimation of β, as discussed in §4·3.

3·3. *Parameter count*

With a scaled envelope model of dimension u, we need r parameters for α, $(r-1)$ parameters for Λ, pu parameters for η, $u(u+1)/2$ parameters for Ω, and $(r-u)(r-u+1)/2$ parameters for Ω_0. We cannot estimate Γ, but only its span, so $u(r-u)$ parameters are needed for $\mathrm{span}(\Gamma) = \mathcal{E}_{\Lambda^{-1}\boldsymbol{\Sigma}\Lambda^{-1}}(\Lambda^{-1}\mathcal{B})$. Then the total number of parameters is $N(u) = 2r - 1 +$

$pu + r(r+1)/2$. Compared to an envelope model with the same dimension, the scaled envelope model has $r-1$ additional parameters because of the diagonal scaling matrix Λ.

4. ESTIMATORS AND THEIR PROPERTIES

4·1. Maximum likelihood estimation when Λ is known

As background, we first discuss estimation when Λ is known. In this case, we transform the response Y in (4) to $\Lambda^{-1}Y$ and write the resulting ordinary envelope model as

$$\Lambda^{-1}Y = \alpha_o + \Gamma\eta X + \epsilon_o, \quad \mathrm{var}(\epsilon_o) = \Sigma_o = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T. \tag{5}$$

This leads to scaled envelope estimators $\widehat{\beta}_\Lambda$ and $\widehat{\Sigma}_\Lambda$ of β and Σ, when Λ is known: first transform Y to $\Lambda^{-1}Y$ and estimate $\beta_o = \Gamma\eta$ and Σ_o from model (5) following Cook et al. (2010). Then $\widehat{\beta}_\Lambda = \Lambda\widehat{\beta}_o$ and $\widehat{\Sigma}_\Lambda = \Lambda\widehat{\Sigma}_o\Lambda$.

Model (5) is just an ordinary envelope model with response $\Lambda^{-1}Y$. We use the subscript o to stand for quantities from this model, which occur within the context of the scaled envelope model, to distinguish it from the ordinary envelope model (3) when $\Lambda = I_r$. For instance, $\beta_o = \Gamma\eta$. It will be seen later that calculations based on model (5) are informative ingredients for the scaled envelope model.

4·2. Maximum likelihood estimation

In this section, we assume for the purpose of developing estimators of β and Σ that the errors ε in (4) are normally distributed. Normality is not required for the definition of scaled envelopes, but this assumption results in estimators that perform well when normality does not hold, as discussed in §6·2.

Suppose that the observed data (X_i, Y_i) $(i = 1, \ldots, n)$, are independent, and n is the sample size. Let $\bar{Y}$ denote the sample mean of Y. Then the maximum likelihood estimators $\widehat{\Gamma}$ and $\widehat{\Lambda}$ of

Γ and Λ can be obtained by minimizing the objective function,

$$L(\Lambda, \Gamma) = \log|\Gamma^T \Lambda^{-1} \widetilde{\Sigma}_{\text{res}} \Lambda^{-1} \Gamma| + \log|\Gamma^T \Lambda \widetilde{\Sigma}_Y^{-1} \Lambda \Gamma|. \tag{6}$$

Technical details are given in Appendix A.

The maximum likelihood estimators of the rest of the parameters are as follows: $\widehat{\Gamma}_0$ can be any orthogonal basis of the orthogonal complement of $\text{span}(\widehat{\Gamma})$, $\widehat{\alpha} = \bar{Y}$, $\hat{\eta} = \widehat{\Gamma}^T \widehat{\Lambda}^{-1} \widetilde{\beta}$, $\widehat{\Omega} = \widehat{\Gamma}^T \widehat{\Lambda}^{-1} \widetilde{\Sigma}_{\text{res}} \widehat{\Lambda}^{-1} \widehat{\Gamma}$, $\widehat{\Omega}_0 = \widehat{\Gamma}_0^T \widehat{\Lambda}^{-1} \widetilde{\Sigma}_Y \widehat{\Lambda}^{-1} \widehat{\Gamma}_0$, $\widehat{\beta} = \widehat{\Lambda} \widehat{P}_\Gamma \widehat{\Lambda}^{-1} \widetilde{\beta}$, and

$$\begin{aligned}\widehat{\Sigma} &= \widehat{\Lambda} \widehat{P}_\Gamma \widehat{\Lambda}^{-1} \widetilde{\Sigma}_{\text{res}} \widehat{\Lambda}^{-1} \widehat{P}_\Gamma \widehat{\Lambda}^T + \widehat{\Lambda} \widehat{P}_{\Gamma_0} \widehat{\Lambda}^{-1} \widetilde{\Sigma}_Y \widehat{\Lambda}^{-1} \widehat{P}_{\Gamma_0} \widehat{\Lambda} \\ &= \widehat{\Lambda} \widehat{\Gamma} \widehat{\Omega} \widehat{\Gamma}^T \widehat{\Lambda}^T + \widehat{\Lambda} \widehat{\Gamma}_0 \widehat{\Omega}_0 \widehat{\Gamma}_0^T \widehat{\Lambda}.\end{aligned}$$

The forms of $\widehat{\beta}$ and $\widehat{\Sigma}$ reveal the working process of estimation under the scaled envelope model, as introduced in §3·1. For instance, consider $\widehat{\beta} = \widehat{\Lambda} \widehat{P}_\Gamma \widehat{\Lambda}^{-1} U^T F (F^T F)^{-1}$, where U is the $n \times r$ matrix whose i-th row is $(Y_i - \bar{Y})^T$, and F is the $n \times p$ matrix whose i-th row is X_i^T $(i = 1, \ldots, n)$. The response is first rescaled $Y \to \widehat{\Lambda}^{-1} Y$ and centered to get $\widehat{\Lambda}^{-1} U^T$ and then ordinary envelope estimation is performed using the rescaled response to get $\widehat{P}_\Gamma \widehat{\Lambda}^{-1} U^T F (F^T F)^{-1}$. After that the estimator is transformed back to the original scales to get $\widehat{\beta}$. This confirms the discussion in §3·1: the scaled envelope model transforms Y to $\widehat{\Lambda} \widehat{P}_\Gamma \widehat{\Lambda}^{-1} Y$, and the process $\widehat{\Lambda} \widehat{P}_\Gamma \widehat{\Lambda}^{-1}$ is the same as treating $\widehat{\Lambda}^{-1}$ as a similarity transformation to the original scale of Y_N.

4·3. Parameter identifiability

In our experience, the objective function (6) nearly always has a unique pair $\{\widehat{\Lambda}, \text{span}(\widehat{\Gamma})\}$ as the global minimizer. However, occasionally we may find that Λ and $\text{span}(\Gamma)$ are not identifiable. When this happens, the objective function will typically be flat along some directions, and any value may be returned in those directions. But this potential non-uniqueness is not an issue, as the parameters that we are interested in are β and Σ. Proposition 1 ensures that the maximizers in β and Σ with respect to the log-likelihood function are in fact uniquely defined. This implies that we will get the same estimators $\widehat{\beta}$ and $\widehat{\Sigma}$ whether the global minimizer $\{\widehat{\Lambda}, \mathrm{span}(\widehat{\Gamma})\}$ is unique or not, which is also confirmed in our numerical experiments.

Following Henderson & Searle (1979), the operator vec: $\mathbb{R}^{a\times b} \to \mathbb{R}^{ab}$ stacks the columns of a matrix, and the operator vech: $\mathbb{R}^{a\times a} \to \mathbb{R}^{a(a+1)/2}$ stacks the lower triangular part of a symmetric matrix. Then we combine the constituent parameters Λ, η, Γ, Ω and Ω_0 in the scaled envelope models (4) into the vector $\phi = \{\lambda^T, \mathrm{vec}(\eta)^T, \mathrm{vec}(\Gamma)^T, \mathrm{vech}(\Omega)^T, \mathrm{vech}(\Omega_0)^T\}^T = (\lambda^T, \phi_o^T)^T$, where $\phi_0 = \{\mathrm{vec}(\eta)^T, \mathrm{vec}(\Gamma)^T, \mathrm{vech}(\Omega)^T, \mathrm{vech}(\Omega_0)^T\}^T$ contains the constituent parameters from model (5) and $\lambda = (\lambda_2, \ldots, \lambda_r)^T$ is the vector of the 2nd to the rth diagonal elements of Λ. Let L denote the $r^2 \times (r-1)$ matrix with columns $e_j \otimes e_j$, where $e_j \in \mathbb{R}^r$ contains a 1 in the j-th position and 0's elsewhere, $j = 2, \ldots, r$. Then, for later use, $\lambda = L^T \mathrm{vec}(\Lambda)$. As $\beta = \Lambda\Gamma\eta = \Lambda\beta_o$ and $\Sigma = \Lambda(\Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T)\Lambda = \Lambda\Sigma_o\Lambda$, β and Σ are both functions of ϕ.

PROPOSITION 1. *Assume that model (4) has independent but not necessarily normal errors with finite second moments, and that* $n^{-1}\sum_{i=1}^n X_i X_i^T > 0$. *Then* $\beta(\phi)$ *and* $\Sigma(\phi)$ *are identifiable and* $\widehat{\beta}$ *and* $\widehat{\Sigma}$ *are uniquely defined.*

Proposition 1 says that even when ϕ is not identifiable, β and Σ are identifiable. Further, we can get unique estimators $\widehat{\beta} = \beta(\widehat{\phi})$ and $\widehat{\Sigma} = \Sigma(\widehat{\phi})$. This provides the foundation for our discussion of the asymptotic distribution and consistency of $\widehat{\beta}$ and $\widehat{\Sigma}$ in §4·4 and §4·5. The proof of Proposition 1 is included in Appendix B.

Although Λ and $\mathrm{span}(\Gamma)$ are not of particular interest, a discussion of identifiability may result in a better understanding of the scaled envelope model (4). In the supplementary material, we show that under some weak conditions, Λ is identifiable if and only if $\mathrm{span}(\Gamma)$ is identifiable.

4·4. Asymptotic distribution

In this section, we give the asymptotic distribution of the scaled envelope estimator $\{\text{vec}(\widehat{\beta})^T, \text{vech}(\widehat{\Sigma})^T\}^T$ under normality. Several definitions are needed in preparation for the result. The contraction matrix $C_r \in \mathbb{R}^{r(r+1)/2 \times r^2}$ and the expansion matrix $E_r \in \mathbb{R}^{r^2 \times r(r+1)/2}$ link the vec and vech operators: for any symmetric matrix $A \in \mathbb{R}^{r\times r}$, $\text{vec}(A) = E_r \text{vech}(A)$, and $\text{vech}(A) = C_r \text{vec}(A)$. Let $\Sigma_X = \lim_{n\to\infty} n^{-1} \sum_{i=1}^n X_i X_i^T$, and let p_{ii} denote the ith diagonal element of the projection matrix P_F, where F was defined in §4·2.

We write the asymptotic covariance matrix in terms of quantities designated with subscripts o that stem from model (5), which has response $\Lambda^{-1}Y$, and one quantity that depends on Λ. We next describe these constructions. The gradient matrix $G_o = \partial\{\text{vec}(\beta_o)^T, \text{vech}(\Sigma_o)^T\}^T / \partial\phi_o^T$ for model (5) has dimension $\{pr + r(r+1)/2\} \times \{pu + r(r+1)/2\}$ and is equal to (Cook et al., 2010)

$$\begin{pmatrix} I_p \otimes \Gamma & \eta^T \otimes I_r & 0 & 0 \\ 0 & 2C_r(\Gamma\Omega \otimes I_r - \Gamma \otimes \Gamma_0\Omega_0\Gamma_0^T) & C_r(\Gamma \otimes \Gamma)E_u & C_r(\Gamma_0 \otimes \Gamma_0)E_{r-u} \end{pmatrix}.$$

The Fisher information for $\{\text{vec}(\beta_o)^T, \text{vech}(\Sigma_o)^T\}^T$ from model (5) is the $\{rp + r(r+1)/2\} \times \{rp + r(r+1)/2\}$ block diagonal matrix $J_o = \text{bdiag}\{\Sigma_X \otimes \Sigma_o^{-1}, 2^{-1}E_r^T(\Sigma_o^{-1} \otimes \Sigma_o^{-1})E_r\}$, where $\text{bdiag}(\cdot)$ indicates a block diagonal matrix with the diagonal blocks as arguments. Let $h_o = \{(\beta_o \otimes I_r), 2(\Sigma_o \otimes I_r)C_r^T\}^T$, which is the gradient component $h_o = \partial\{\text{vec}(\beta)^T, \text{vech}(\Sigma)^T\}^T/\partial\Lambda$ for the scaled model (4) evaluated at $\Lambda = I_r$. Let $A_o = Q_{G_o(J_o)} h_o L$ and let $D_\Lambda = \text{bdiag}\{I_p \otimes \Lambda, C_r(\Lambda \otimes \Lambda)\text{E}_r\}$, which is a block diagonal matrix with the same dimensions as J_o. Of the quantities defined here, only D_Λ depends on Λ.

The gradient matrix $H = \partial\{\text{vec}(\beta)^T, \text{vech}(\Sigma)^T\}^T/\partial\phi^T$ for the scaled envelope model (4) has dimension $\{pr + r(r+1)/2\} \times \{r - 1 + pu + r(r+1)/2\}$ and can be represented as $H =$

$\{D_\Lambda h_o(I_r \otimes \Lambda^{-1})L, D_\Lambda G_o\}$. The Fisher information J under the scaled envelope model can be obtained by replacing Σ_o with Σ in J_o, $J = \mathrm{bdiag}\{\Sigma_X \otimes \Sigma^{-1}, 2^{-1}E_r^T(\Sigma^{-1} \otimes \Sigma^{-1})E_r\}$.

PROPOSITION 2. *Under model (4) with normal errors, assume that* $\max_{i \le n} p_{ii} \to 0$ *as* $n \to \infty$. *Then* $\sqrt{n}[\{\mathrm{vec}(\widehat{\beta}) - \mathrm{vec}(\beta)\}^T, \{\mathrm{vech}(\widehat{\Sigma}) - \mathrm{vech}(\Sigma)\}^T]^T$ *converges in distribution to a normal random vector with mean zero and covariance matrix*

$$V = H(H^TJH)^\dagger H^T = D_\Lambda\{A_o(A_o^TJ_oA_o)^\dagger A_o^T\}D_\Lambda + D_\Lambda\{G_o(G_o^TJ_oG_o)^\dagger G_o^T\}D_\Lambda = V_1 + V_2,$$

where $V_1 = D_\Lambda\{A_o(A_o^TJ_oA_o)^\dagger A_o^T\}D_\Lambda$ *and* $V_2 = D_\Lambda\{G_o(G_o^TJ_oG_o)^\dagger G_o^T\}D_\Lambda$.

The proof of Proposition 2 is included in Appendix B. Since $J^{-1} - H(H^TJH)^\dagger H^T = J^{-1/2}Q_{J^{1/2}H}J^{-1/2} \ge 0$, it follows that $V \le J^{-1}$, where J^{-1} is the asymptotic covariance matrix of $\{\mathrm{vec}(\widetilde{\beta})^T, \mathrm{vech}(\widetilde{\Sigma}_{\mathrm{res}})^T\}^T$. Consequently,

COROLLARY 1. *Assume that the conditions in Proposition 2 hold. Then the scaled envelope model (4) is asymptotically more efficient than or as efficient as the standard model (1) in estimating* β *and* Σ.

The factor $G_o(G_o^TJ_oG_o)^\dagger G_o^T$ that occurs in V_2 is the asymptotic covariance matrix for the ordinary envelope estimator of $\{\mathrm{vec}(\widehat{\beta}_o), \mathrm{vech}(\widehat{\Sigma}_o)\}$ under model (5) (Cook et al., 2010). Consequently, V_2 is the asymptotic covariance of $\{\mathrm{vec}(\widehat{\beta}_\Lambda), \mathrm{vech}(\widehat{\Sigma}_\Lambda)\}$ under the scaled envelope model assuming that Λ is known. This implies that V_1 can then be interpreted as the asymptotic cost of estimating Λ; that is, the part of V that is due to the estimation of Λ. Since $\mathrm{tr}(V_1V_2^{-1})$ does not depend on Λ, the relative cost of estimating Λ is constant in Λ, although it can depend on the other parameters in the model.

These asymptotic results are for the estimators of β and Σ jointly. The regression coefficients β are often of special interest in practice, so we next focus on this aspect of the regression. The following notational convention will facilitate the discussion. If $\sqrt{n}(T - \theta)$ converges in distribution to a random variable with mean 0 and variance A, we write the asymptotic variance of T as $\mathrm{avar}(\sqrt{n}T) = A$.

The asymptotic variance $\mathrm{avar}\{\sqrt{n}\mathrm{vec}(\widehat{\beta})\}$ of the scaled envelope estimator of β is the upper $pr \times pr$ diagonal block of V, $\mathrm{avar}\{\sqrt{n}\mathrm{vec}(\widehat{\beta})\} = (I_{pr}, 0)V_1(I_{pr}, 0)^T + \mathrm{avar}\{\sqrt{n}\mathrm{vec}(\widehat{\beta}_\Lambda)\}$, where $(I_{pr}, 0)$ has dimension $pr \times \{pr + r(r+1)/2\}$.

COROLLARY 2. *Assume that the conditions in Proposition 2 hold and that* $\Sigma_o = \sigma^2 I_r$, *so* $\Sigma = \sigma^2\Lambda^2$. *Then* $\mathrm{avar}\{\mathrm{vec}(\widehat{\beta})\} = \mathrm{avar}\{\mathrm{vec}(\widehat{\beta}_\Lambda)\} = \mathrm{avar}\{\mathrm{vec}(\widetilde{\beta})\}$, *where, as defined previously,* $\widetilde{\beta}$ *denotes the ordinary least squares estimator of* β *from the standard model (1).*

This corollary says that in the special case where the scaled responses $\Lambda^{-1}Y$ have error covariance matrix $\Sigma_o = \sigma^2 I_r$, the asymptotic variance of the scale envelope estimator $\widehat{\beta}$ is the same as that of the scaled envelope estimator $\widehat{\beta}_\Lambda$ when Λ is known, which is the same as the asymptotic variance of the ordinary least squares estimator from the standard model. Consequently, scaling offers no gains and, since $\mathrm{avar}\{\mathrm{vec}(\widehat{\beta})\} = (I_{pr}, 0)V_1(I_{pr}, 0)^T + \mathrm{avar}\{\sqrt{n}\mathrm{vec}(\widehat{\beta}_\Lambda)\} \leq \mathrm{avar}\{\mathrm{vec}(\widetilde{\beta})\}$, there is also no asymptotic cost of estimating Λ for the ultimate goal of estimating β, $(I_{pr}, 0)V_1(I_{pr}, 0)^T = 0$. However, in other cases there can be considerable gain in pursuing scaling, particularly when $\|\Omega_0\| \gg \|\Omega\|$. These results are illustrated in §6.

4·5. Consistency

As the scaled envelope estimators are obtained using the normal likelihood as an objective function, a natural question is on the consistency of these estimators when the normality assumption fails. The next proposition gives conditions for $\sqrt{n}$ consistency of $\widehat{\beta}$ and $\widehat{\Sigma}$.

PROPOSITION 3. *Assume that model (4) has independent but not necessary normal errors with mean zero and finite fourth moments, and that* $\max_{i \leq n} p_{ii} \to 0$ *as* $n \to \infty$. *Then*

$$\sqrt{n}\{(\mathrm{vec}(\widehat{\beta})^T, \mathrm{vech}(\widehat{\Sigma})^T)^T - (\mathrm{vec}(\beta)^T, \mathrm{vech}(\Sigma)^T)^T\}$$

is asymptotically normally distributed, and $\widehat{\beta}$ *and* $\widehat{\Sigma}$ *are* $\sqrt{n}$ *consistent estimators of* β *and* Σ.

The assumption on p_{ii} is the same condition that Huber (1973) used to establish consistency for the standard model estimator $\mathrm{vec}(\widetilde{\beta})$, which basically requires that the maximum leverage goes to zero as $n \to \infty$. Additionally, in finite samples the estimators are robust to moderate departure from normality as demonstrated in the simulations in §6·2. The proof of Proposition 3 is included in Appendix B.

5. SELECTION OF u

Likelihood-based methods, such as the Akaike information criterion AIC, the Bayesian information criterion BIC, or other information criteria, can be used to select the dimension u for the scaled envelope model. Non-parametric methods as cross validation or permutation tests (Cook & Yin, 2001) can also be used to select u. We will use BIC in data examples, but will discuss properties of both AIC and BIC.

The AIC estimator of u is $\arg\min -2\hat{L}(u) + 2N(u)$, where the minimum is taken over the set of integers $0, 1, \ldots, r$, $N(u) = 2r - 1 + pu + r(r+1)/2$ is the number of parameters, as discussed in §3·3, and $\hat{L}(u)$ is the maximized log likelihood under the scaled envelope model with dimension u,

$$\hat{L}(u) = -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\widetilde{\Sigma}_Y| - \frac{n}{2}\log|\widehat{\Gamma}^T\widehat{\Lambda}^{-1}\widetilde{\Sigma}_{\mathrm{res}}\widehat{\Lambda}^{-1}\widehat{\Gamma}| - \frac{n}{2}\log|\widehat{\Gamma}^T\widehat{\Lambda}\widetilde{\Sigma}_Y^{-1}\widehat{\Lambda}\widehat{\Gamma}|.$$

Here $\mathrm{span}(\widehat{\Gamma})$ and $\widehat{\Lambda}$ are maximum likelihood estimators for $\mathcal{E}_{\Lambda^{-1}\mathbf{\Sigma}\Lambda^{-1}}(\Lambda^{-1}\mathcal{B})$ and Λ under the scaled envelope model. BIC works similarly, except its objective function is $-2\hat{L}(u) + \log(n)N(u)$.

In univariate linear regression, the asymptotic properties of AIC and BIC have been studied in detail. Briefly, if the true model is among the candidate models, BIC selects the true model with probability approaching 1 as $n \to \infty$ (Yang, 2005), and AIC will have positive probability of selecting models that properly include the true model (Nishii, 1984). These properties can be generalized straightforwardly to multivariate linear regression. The next proposition gives the properties of AIC and BIC in the framework of the scaled envelope model. The candidate set is the set of scaled envelope models having dimensions varying from 0 to r.

PROPOSITION 4. *Under the scaled envelope model (4) assuming normal errors, if there is one and only one true model in the candidate set, as $n \to \infty$,* BIC *will select the true model with probability tending to 1, and* AIC *will select a model that at least contains the true model.*

The proof of Proposition 4 is similar to the proof in Nishii (1984): Scaled envelope models with dimension smaller than the true model introduce bias into the mean function that dominates the penalty term asymptotically, and scaled envelope models with dimension larger than the true model have larger penalty terms which will be not selected by BIC but selected by AIC with positive probabilities.

6. SIMULATIONS AND DATA EXAMPLE

6·1. Computing

Given u, to estimate the scales Λ and $\mathrm{span}(\Gamma)$, we apply an alternating algorithm to (6). We can start with $\Lambda = I_r$ or any reasonable guess, and our numerical experience suggests that the alternating algorithm is not sensitive to the choice of starting values. When Λ is specified, $\Lambda^{-1}Y$ follows an envelope model with mean $\Gamma\eta X$ and covariance matrix $\Sigma_o = \Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T$. When Γ is specified, Λ can be estimated by minimizing (6) using a standard optimization algorithm. We continue the process until the absolute value of the percentage increment of (6) between two consecutive iterations is less than a pre-specified value.

6·2. *Simulations*

A simulation study was conducted to compare the scaled envelope estimator with the standard model estimator on finite sample size performance. We simulated data from model (4), with $r = 10$, $u = 5$ and $p = 5$. The elements in X were generated once as independent $N(0, 5)$ random variables, but the analysis was still conditioned on their observed values. We took $\Omega = \sigma^2 I_5$ and $\Omega_0 = \sigma_0^2 I_5$. The matrix η was generated as a 5×5 matrix of independent $N(0, 2)$ random variables, and Γ was obtained by orthogonalizing a 10×5 matrix of independent $U(0, 1)$ random variables. The scale matrix Λ was a diagonal matrix with diagonal elements 1, $2^{0\cdot5}$, 2^1, $2^{1\cdot5}$, …, $2^{4\cdot5}$. We took σ^2 as 0.25 and σ_0^2 as 5 and 25. The sample sizes were 100, 200, 300, 500, 800, 1200, and 200 replicates were generated for each sample size. With each sample size, the standard deviation of each element in $\widehat{\beta}$ over the replicates is computed, which we call the actual standard deviations of the elements in $\widehat{\beta}$. We also computed the bootstrap standard deviations by bootstrapping the residuals 200 times.

We applied the ordinary envelope model to the data and inferred that $u = 10$, so the envelope estimator is the same as the standard estimator, and no efficiency gains were offered. The scaled envelope model effectively removed the immaterial part of Y relative to X, and obtained efficiency gains compared to the standard model, both asymptotically and with finite sample sizes. The scaled envelope model was fitted according to the discussion in §6·1. The left panel of Fig. 3 plots the standard deviations of a selected element in $\widehat{\beta}$ with $\sigma_0^2 = 5$. We took the logarithm of both the sample size and the standard deviation to linearize their relationship. The simulations for the right panel were based on the same setting as for the left panel, except $\sigma_0^2 = 25$. With sample size larger than 200, the efficiency gain remains roughly constant as sample size increases, and it is also about the same as the asymptotic difference between the scaled envelope

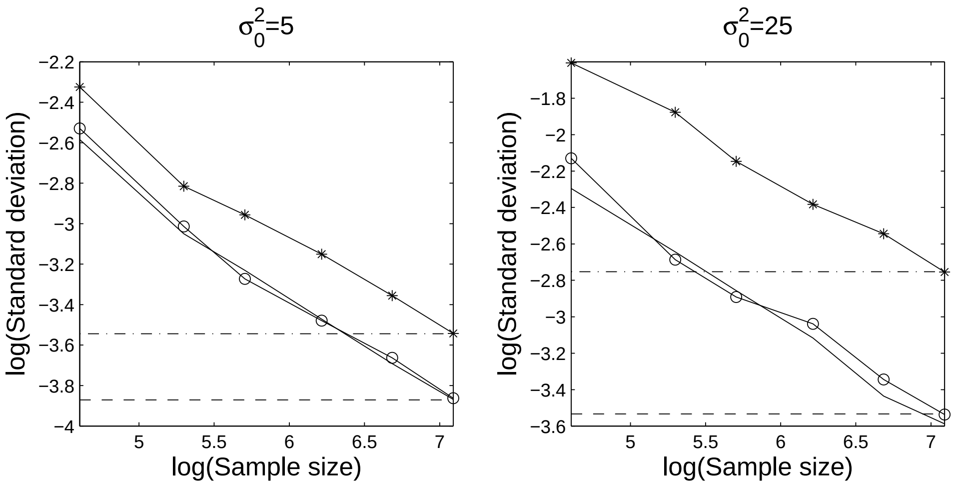

Fig. 3: Logarithmic comparison of the scaled envelope estimators and standard model estimators: — the actual standard deviation of scaled envelope estimators; –∗– actual standard deviation of standard model estimators; –○– bootstrap standard deviation of the scaled envelope estimators; – – asymptotic standard deviation of scaled envelope estimators; – · – asymptotic standard deviation of the standard model estimators.

estimator and the least squares estimator. Figure 3 suggests that the bootstrap standard deviation is a good estimator of the actual standard deviation.

Table 1 provides the mean and standard deviation of 200 estimated scales with $\sigma_0^2 = 5$. The results for $\sigma_0^2 = 25$ are similar. From the table, we find that our algorithm is quite stable.

Figure 4 presents the asymptotic behavior of the scaled envelope estimators under non-normal errors. We performed the same simulations as in the right panel of Fig. 3, except the errors were generated as centered and consistently scaled t_6, $U(0,1)$, and χ_4^2 random variables to represent distributions with longer tails, shorter tails and skewness. We used six degrees of freedom for the t distribution to ensure the existence of fourth moments, as required by Proposition 3. Figure 4 does not show notable differences caused by the different error distributions, so we conclude that a moderate departure from normality does not much affect the results. With non-normal

Table 1: Mean of base 2 logarithms of the diagonal elements in $\widehat{\Lambda}$, the number in parentheses are their standard deviations, $\sigma_0^2 = 5$.

n	100	500	1200
$\log_2 \hat{\lambda}_2$	$0{\cdot}50$ $(0{\cdot}073)$	$0{\cdot}50$ $(0{\cdot}032)$	$0{\cdot}50$ $(0{\cdot}020)$
$\log_2 \hat{\lambda}_3$	$0{\cdot}99$ $(0{\cdot}085)$	$1{\cdot}00$ $(0{\cdot}039)$	$1{\cdot}00$ $(0{\cdot}022)$
$\log_2 \hat{\lambda}_4$	$1{\cdot}50$ $(0{\cdot}067)$	$1{\cdot}50$ $(0{\cdot}029)$	$1{\cdot}50$ $(0{\cdot}019)$
$\log_2 \hat{\lambda}_5$	$2{\cdot}00$ $(0{\cdot}051)$	$2{\cdot}00$ $(0{\cdot}024)$	$2{\cdot}00$ $(0{\cdot}016)$
$\log_2 \hat{\lambda}_6$	$2{\cdot}50$ $(0{\cdot}062)$	$2{\cdot}50$ $(0{\cdot}029)$	$2{\cdot}50$ $(0{\cdot}017)$
$\log_2 \hat{\lambda}_7$	$2{\cdot}99$ $(0{\cdot}065)$	$3{\cdot}00$ $(0{\cdot}029)$	$3{\cdot}00$ $(0{\cdot}019)$
$\log_2 \hat{\lambda}_8$	$3{\cdot}50$ $(0{\cdot}055)$	$3{\cdot}50$ $(0{\cdot}023)$	$3{\cdot}50$ $(0{\cdot}016)$
$\log_2 \hat{\lambda}_9$	$3{\cdot}99$ $(0{\cdot}057)$	$4{\cdot}00$ $(0{\cdot}025)$	$4{\cdot}00$ $(0{\cdot}016)$
$\log_2 \hat{\lambda}_{10}$	$4{\cdot}50$ $(0{\cdot}054)$	$4{\cdot}50$ $(0{\cdot}025)$	$4{\cdot}50$ $(0{\cdot}016)$

errors, the estimator is no longer the maximum likelihood estimator, but efficiency gains are still realized.

As discussed following Proposition 2, the asymptotic variance of $\mathrm{vec}(\widehat{\beta})$ depends on $(I_{pr}, 0)V_1(I_{pr}, 0)^T$, the cost of estimating the scaling parameters, and $\mathrm{avar}\{\sqrt{n}\mathrm{vec}(\widehat{\beta}_\Lambda)\}$, the asymptotic variance of $\mathrm{vec}(\widehat{\beta})$ assuming that Λ is known. Fig. 5 displays the relative cost $C = \mathrm{tr}^{1/2}[(I_{pr}, 0)V_1(I_{pr}, 0)^T \mathrm{avar}^{-1}\{\sqrt{n}\mathrm{vec}(\widehat{\beta}_\Lambda)\}]$ in different settings. We used the same model as the one used to generate the left panel of Fig. 3. While σ_0 was fixed at $\sqrt{5}$, we evaluated the relative cost with σ equal to $0{\cdot}1$, $0{\cdot}2$, $0{\cdot}5$, 1, $\sqrt{5}$, 5 and 10. We also multiplied the original η by $0{\cdot}25$, 1 and 4 to represent different signal levels. Fig. 5 indicates that the relative cost is lower with a stronger signal and less discrepancy between σ and σ_0. It confirms Corollary 2 that when

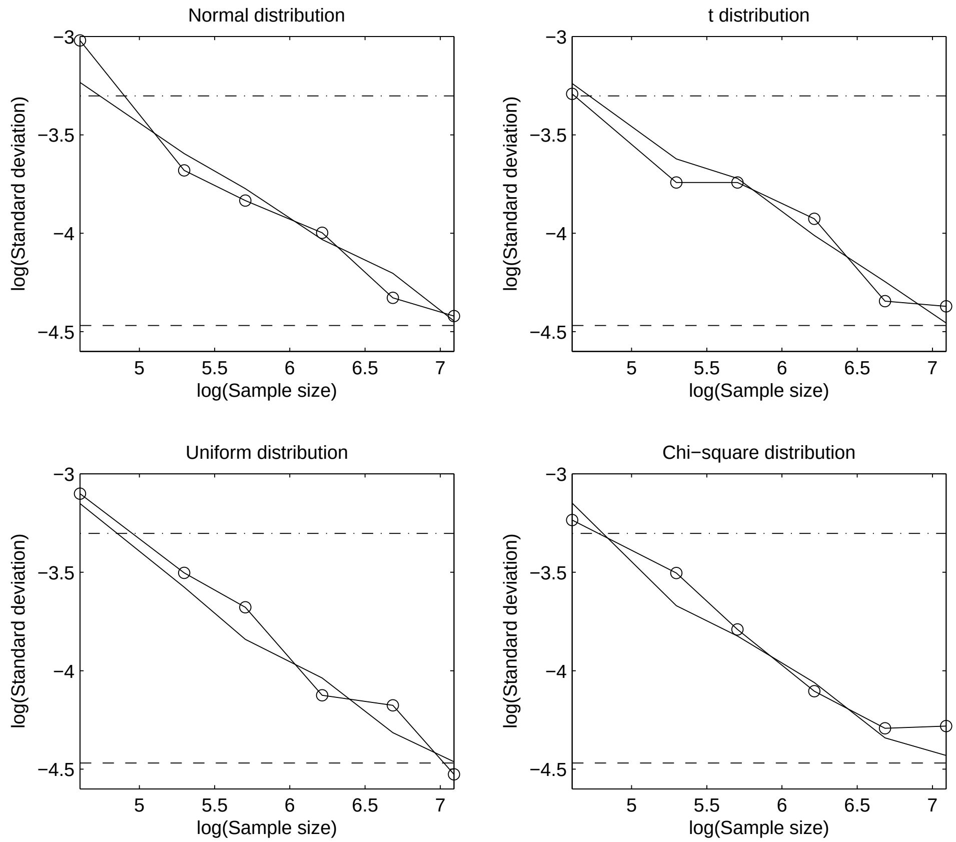

Fig. 4: Comparison of the scaled envelope estimators with normal, t_6, $U(0,1)$, and χ^2_4 errors. The line marks are the same as those in Fig. 3.

$\sigma = \sigma_0$, there is no relative cost in estimating Λ. The relative cost is the highest when the gain from scaled envelopes is the greatest, $\sigma \ll \sigma_0$. It is the lowest when there is little to gain from using scaled envelopes, $\sigma \approx \sigma_0$.

6·3. Data example

For this illustration we used a data set from Johnson & Wichern (2007) on the performance of a firm's sales staff. Fifty sales persons were selected at random and their performance was measured on growth of sales, profitability of sales, and new account sales. The selected sales

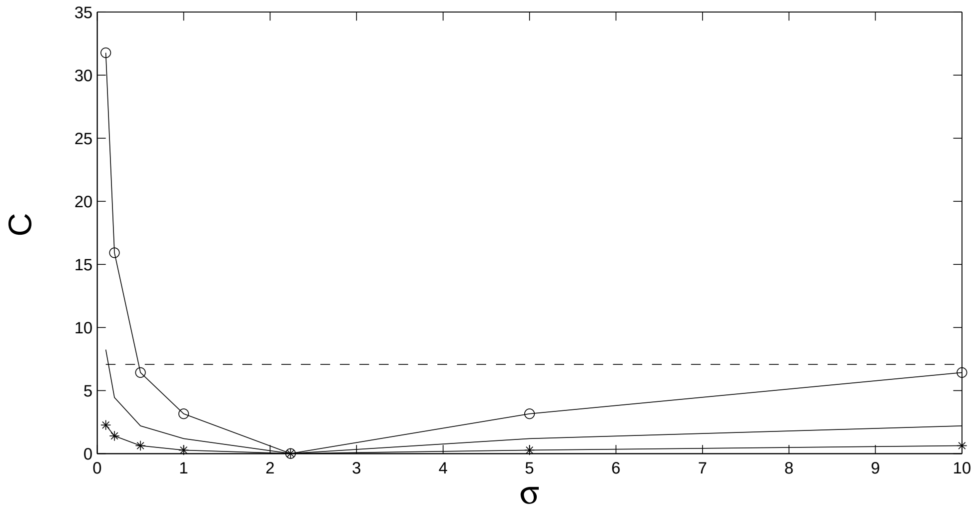

Fig. 5: Relative cost C versus the variation of the material part of $\Lambda^{-1}Y$, $\sigma = \surd\|\Omega\|$. -o-, — and -*- correspond to η multiplied by $0{\cdot}25$, 1 and 4 respectively. The horizontal line is at $C = \surd(pr)$, which corresponds to equal costs, $(I_{pr}, 0)V_1(I_{pr}, 0)^T = \text{avar}\{\surd n \text{vec}(\widehat{\beta}_\Lambda)\}$.

staff also took four tests that measured creativity, mechanical reasoning, abstract reasoning and mathematical ability. Scores were recorded for these tests. We considered how sales performance X affects test scores Y, yielding $r = 4$ and $p = 3$, and compared the standard errors of the ordinary least squares estimator $\widetilde{\beta}$ to the standard errors of the scaled envelope estimator $\widehat{\beta}$ by using the fractions $f_{ij} = 1 - \hat{\text{avar}}^{1/2}(\surd n\widetilde{\beta}_{ij})/\hat{\text{avar}}^{1/2}(\surd n\widehat{\beta}_{ij})$, where the subscripts i, j indicate the elements of the estimator of β. The standard errors of the ordinary least squares estimators and the ordinary envelope estimators were compared in the same way.

We first fitted an ordinary envelope model to the data and BIC suggested that $u = 3$. Compared to $\widetilde{\beta}$, the standard deviations of the elements in the ordinary envelope estimator were $1{\cdot}0\%$ to $28{\cdot}7\%$ smaller, $0{\cdot}01 \le f_{ij} \le 0{\cdot}287$. A sample size of about $n = 100$ observations would be needed to reduce the standard error of the ordinary least squares estimator by $28{\cdot}7\%$, so using the ordinary envelope estimator is roughly equivalent to doubling the sample size for inference on some elements of β with the ordinary least squares estimator.

When the scaled envelope model was fitted to the data, BIC suggested that $u = 2$. The scale transformation matrix Λ was estimated with diagonal elements $1, 0{\cdot}97, 0{\cdot}81$ and $1{\cdot}70$. Compared to $\widetilde{\beta}$, the standard deviations of the elements in the scaled envelope estimator were $12{\cdot}7\%$ to $68{\cdot}2\%$ smaller, $0{\cdot}127 \leq f_{ij} \leq 0{\cdot}682$, which is a significant improvement over the gains provided by the ordinary envelope model. For instance, a sample size of about $n = 500$ observations would be needed to reduce the standard error of the ordinary least squares estimator by 68%. These gains are reflected by the estimates of $\|\Omega_0\|$ and $\|\Omega\|$: $\|\widehat{\Omega}\| = 1{\cdot}10$ and $\|\widehat{\Omega}_0\| = 13{\cdot}17$.

7. Discussion

By introducing a scaling parameter for each response, the scaled envelope estimator broadens the effective scope of envelope constructions, and can bring efficiency gains that are not offered by the ordinary envelope estimator. While scaled envelopes are applicable in any multivariate linear regression where (1) is a useful model, we have found them particularly serviceable when the ordinary envelope offers only modest gains. The specific estimation procedure proposed here should give good results when the error distribution does not deviate substantially from the multivariate normal; otherwise, a different, perhaps robust, estimator may be desirable. Although rare, we have observed the alternating algorithm described in §6·1 can get caught in a local minimum, resulting in a modified estimator that does not maximize the likelihood-based objective function and that might then be less efficient than the ordinary least squares estimator. Fortunately, this can be studied by using the bootstrap to compare performance, so the issue is trackable in practice.

The partial envelope model was proposed by Su and Cook (2011) for efficient estimation of a part of β when a subset of the predictors is of special interest. Under model (1), divide $X \in \mathbb{R}^p$ into $X_1 \in \mathbb{R}^{p_1}$ and $X_2 \in \mathbb{R}^{p_2}$ with $p_1 + p_2 = p$, so that $Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$, where X_1 is of main interest, $\beta_1 \in \mathbb{R}^{r\times p_1}$ and $\beta_2 \in \mathbb{R}^{r\times p_2}$. Instead of enveloping β, we can envelop only the key parameter β_1. Again we can divide Y into a material part and an immaterial part, but the distribution of the immaterial part is now invariant to changes in X_1, instead of invariant to changes in X as under the envelope model. Let $\mathcal{B}_1 = \mathrm{span}(\beta_1)$. Then the smallest reducing subspace $\mathcal{S}$ of Σ that satisfies $\mathcal{B}_1 \subseteq \mathcal{S}$ and $\Sigma = P_{\mathcal{S}}\Sigma P_{\mathcal{S}} + Q_{\mathcal{S}}\Sigma Q_{\mathcal{S}}$ is called a partial Σ-envelope of $\mathcal{B}_1$, which is denoted by $\mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}_1)$. Model (1) is called partial envelope model when these conditions are imposed with $\mathcal{S} = \mathcal{E}_{\mathbf{\Sigma}}(\mathcal{B}_1)$. Compared with the envelope model, the partial envelope model is more flexible in application and is often more efficient for the purpose of estimating β_1.

Scaling can be incorporated with a partial envelope model as follows. Given a dimension u_1, we can find a scale transformation Λ, such that $\Lambda^{-1}\mathcal{B}_1 \subseteq \mathrm{span}(\Gamma)$, $\Lambda^{-1}\Sigma\Lambda^{-1} = P_\Gamma\Lambda^{-1}\Sigma\Lambda^{-1}P_\Gamma + Q_\Gamma\Lambda^{-1}\Sigma\Lambda^{-1}Q_\Gamma$, where Λ is a diagonal matrix having positive diagonal elements and first element equal to 1, and $\Gamma \in \mathbb{R}^{r\times u_1}$ is an orthogonal basis of the partial $\Lambda^{-1}\Sigma\Lambda^{-1}$-envelope of $\Lambda^{-1}\mathcal{B}_1$. We call (1) the scaled partial envelope model if the preceding two conditions are imposed. The estimation of the parameters and the asymptotic distribution of the estimators can be developed in parallel to the scaled envelope model. Compared to the scaled envelope model, as $\mathcal{B}_1 \subseteq \mathcal{B}$, it is very likely that we come up with a smaller envelope subspace, and achieves greater efficiency gains for the purpose of estimating β_1.

The inner envelope model, introduced in Su & Cook (2012), uses a different construction from the envelope model and can achieve efficient estimation of β even when there is no immaterial information in the data. A scale invariant version of the inner envelope model can be developed similarly, although the procedure will be more complicated.

We confined our discussion to the class of scaling transformations represented by diagonal matrices, but depending on the application envelope methodology might also be developed for other classes of transformations. In signal processing for example, correlated signals Z that follow an envelope model might become mixed to $Y = AZ$, where A is not diagonal but is constrained to fall into a restricted class of transformations like matrices with constant diagonal and off diagonal entries.

Acknowledgement

We are grateful to the editor and three referees for their insightful suggestions and comments that helped us improve the paper. Research for this article was supported in part by a Doctoral Dissertation Fellowship from the Graduate School, University of Minnesota, and by grants from the U.S. National Science Foundation.

APPENDIX

Appendix A: Maximum Likelihood Estimators

The maximum likelihood estimator of α is $\bar{Y}$. Then, with the dimension of the $\Lambda^{-1}\Sigma\Lambda^{-1}$-envelope of $\Lambda^{-1}\mathcal{B}$ fixed at u, the log-likelihood function L_1 is

$$
\begin{aligned}
L_1 &= -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}\{(U - F\beta^T)\Sigma^{-1}(U - F\beta^T)^T\} && \text{(A1)}\\
&= -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\operatorname{tr}[\Sigma^{-1}\{n\widetilde{\Sigma}_{\mathrm{res}} + (\widetilde{\beta} - \beta)F^TF(\widetilde{\beta}^T - \beta^T)\}] && \text{(A2)}\\
&= -\frac{nr}{2}\log(2\pi) - n\log|\Lambda| - \frac{n}{2}\log|\Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T|\\
&\quad -\frac{1}{2}\operatorname{tr}\{(U\Lambda^{-1} - F\eta^T\Gamma^T)(\Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T)^{-1}(U\Lambda^{-1} - F\eta^T\Gamma^T)^T\}. && \text{(A3)}
\end{aligned}
$$

Here (A1), (A2) and (A3) are three versions of the likelihood function: (A1) is a general form with the observed data and parameters β and Σ; (A2) replaces the observed data in (A1) with sufficient statistics $\widetilde{\beta}$ and $\widetilde{\Sigma}_{\text{res}}$; and (A3) rewrites (A1) in terms of the constituent parameters. (A3) has the same form as the log-likelihood function from the envelope model, except we have the extra term $-n\log|\Lambda|$ and the response is $\Lambda^{-1}Y$. Thus, maximizing over all constituent parameters except Λ and Γ, we get the partially maximized form

$$
\begin{aligned}
L_2(\Lambda,\Gamma) &= -\frac{nr}{2}\log(2\pi) - n\log|\Lambda| - \frac{n}{2}\log|\Gamma^T\Lambda^{-1}\widetilde{\Sigma}_{\text{res}}\Lambda^{-1}\Gamma| - \frac{n}{2}\log|\Gamma_0^T\Lambda^{-1}\widetilde{\Sigma}_Y\Lambda^{-1}\Gamma_0| \\
&= -\frac{nr}{2}\log(2\pi) - n\log|\Lambda| - \frac{n}{2}\log|\Gamma^T\Lambda^{-1}\widetilde{\Sigma}_{\text{res}}\Lambda^{-1}\Gamma| - \frac{n}{2}\log|\Lambda^{-1}\widetilde{\Sigma}_Y\Lambda^{-1}| \\
&\quad -\frac{n}{2}\log|\Gamma^T\Lambda\widetilde{\Sigma}_Y^{-1}\Lambda\Gamma| \\
&= -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\widetilde{\Sigma}_Y| - \frac{n}{2}\log|\Gamma^T\Lambda^{-1}\widetilde{\Sigma}_{\text{res}}\Lambda^{-1}\Gamma| - \frac{n}{2}\log|\Gamma^T\Lambda\widetilde{\Sigma}_Y^{-1}\Lambda\Gamma|.
\end{aligned}
$$

Appendix B: Proofs

Proof of Proposition 1. We apply Proposition 3.1 in Shapiro (1986) to prove this proposition, and we will match our notations with Shapiro's during the discussion. For better distinction, we add a subscript s to Shapiro's notation. The θ_s in Shapiro's context is our $\phi = \{\lambda^T, \text{vec}(\eta)^T, \text{vec}(\Gamma)^T, \text{vech}(\Omega)^T, \text{vech}(\Omega_0)^T\}^T$. Shapiro's $\hat{x}_s$ corresponds to our $\{\text{vec}(\widetilde{\beta})^T, \text{vech}(\widetilde{\Sigma}_{\text{res}})^T\}^T$, and Shapiro's ξ_s is $\{\text{vec}(\beta)^T, \text{vech}(\Sigma)^T\}^T$ in our context. The discrepancy function F_s is our log likelihood function, except we omit a constant factor n.

$$
\begin{aligned}
F_s = L_1/n &= -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\text{tr}\{(U - F\beta^T)\Sigma^{-1}(U-F\beta^T)^T/n\} \\
&= -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\text{tr}[\Sigma^{-1}\{n\widetilde{\Sigma}_{\text{res}} + (\widetilde{\beta}-\beta)(F^TF/n)(\widetilde{\beta}^T - \beta^T)\}].
\end{aligned}
$$

As F_s is constructed under normal likelihood function, it satisfies the conditions 1– 4 in §3 of Shapiro (1986). Shapiro's Δ_s is the gradient matrix $\partial\xi_s/\partial\theta_s$, which is the same as H in our context. Let $e = U - F\beta^T$, Shapiro's $V_s = \text{bdiag}\{(F^TF/n)\otimes\Sigma^{-1}, E_r^T(\Sigma^{-1}\otimes\Sigma^{-1})E_r/2\}$ is 1/2 times the Hessian matrix $\partial^2 F_s/\partial\xi_s\partial\xi_s^T$ evaluated at (ξ_s, ξ_s). As we assume $\sum_{i=1}^n X_iX_i^T/n > 0$, V_s is full rank and rank($\Delta_s^T V_s \Delta_s$)=rank(Δ_s). Therefore, all conditions in Proposition 3.1 are satisfied, and the maximizers $\widehat{\beta}$ and $\widehat{\Sigma}$ are uniquely defined. □

Proof of Proposition 3. Since Proposition 2 is a special case of Proposition 3, we prove Proposition 3 first. As we have over-parameterization in Γ, we apply Proposition 4·1 in Shapiro (1986) to establish the proof. The conditions for Proposition 4·1 are the same as Proposition 3·1 in Shapiro, except with an additional assumption that $n^{1/2}(\hat{x}_s - \xi_s)$ is asymptotically normal. We have shown that all the conditions in Shapiro's Proposition 3·1 are satisfied as we discussed in the proof of our Proposition 1. The condition on p_{ii} guarantees that the asymptotic distribution of $n^{1/2}\{(\mathrm{vec}(\widetilde{\beta})^T, \mathrm{vech}(\widetilde{\Sigma}_{\mathrm{res}})^T)^T - (\mathrm{vec}(\beta)^T, \mathrm{vech}(\Sigma)^T)^T\}$ is multivariate normal, so the additional assumption is also satisfied. Therefore from Proposition 4·1 of Shapiro (1986) and using Shapiro's notation, the asymptotic variance has the from $\Delta_s(\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T V_s \Gamma_s V_s \Delta_s (\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T$, where Shapiro's Γ_s is the asymptotic variance of $\{(\mathrm{vec}(\widetilde{\beta})^T, \mathrm{vech}(\widetilde{\Sigma}_{\mathrm{res}})^T\}^T$. □

Proof of Proposition 2. The proof of Proposition 2 starts with the asymptotic covariance matrix $\Delta_s(\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T V_s \Gamma_s V_s \Delta_s (\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T$ given at the end of Proposition 3. With the additional assumption of normality, Shaprio's $\Gamma_s = V_s^{-1}$. Therefore the asymptotic covariance matrix has the form $\Delta_s(\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T$, which is $V = H(H^T J H)^\dagger H^T$ in our notation. In the rest of the proof , which involves involves simplifying V, we use only our notation.

We directly calculated $H = \partial\{\mathrm{vec}(\beta)^T, \mathrm{vech}(\Sigma)^T\}^T / \partial\phi^T = \{D_\Lambda h_o (I_p \otimes \Lambda^{-1}) L, D_\Lambda G_o\} = (H_1, H_2)$, where H_1 and H_2 are defined implicitly to simplify subsequent expressions. Since V is invariant under full rank linear transformations of the columns of H, we next transform the columns of H by the non-singular matrix

$$T = \begin{pmatrix} I_{r-1} & 0 \\ -(H_2^T J H_2)^\dagger H_2^T J H_1 & I_{r(r+1)/2} \end{pmatrix}.$$

Then $HT = (Q_{H_2(J)}H_1, H_2)$ and $T^T H^T JHT = \text{bdiag}(H_1^T Q_{H_2(J)}^T JQ_{H_2(J)}H_1, G_o^T J_o G_o)$. Then by straightforward algebra we have

$$V = HT(T^T H^T JHT)^\dagger T^T H^T = J^{-1/2}PJ^{-1/2} + D_\Lambda G_o(G_o^T J_o G_o)^\dagger G_o^T D_\Lambda^T,$$

where P is the projection onto the span of $J^{1/2}Q_{H_2(J)}H_1$. The second term on the right of the last expression is the same as V_2 stated in the proposition. The first term can be expressed as V_1 by using the identities $Q_{H_2(J)}H_1 = D_\Lambda Q_{G_o(J_o)}D_\Lambda^{-1}H_1 = D_\Lambda Q_{G_o(J_o)}h_o L\Lambda_1^{-1} = D_\Lambda A_o \Lambda_1^{-1}$, where $\Lambda_1 = \text{diag}(\lambda_2, \ldots, \lambda_r)$. □

Proof of Corollary 2. It follows from the discussion §5·2 in Cook et al. (2010) that, in under model (5), $\text{avar}\{\sqrt{n}\text{vec}(\widehat{\beta}_o)\} = \Sigma_X^{-1} \otimes \Sigma_o$, and consequently $\text{avar}\{\sqrt{n}\text{vec}(\widehat{\beta}_\Lambda)\} = \text{avar}\{\sqrt{n}\text{vec}(\Lambda\widehat{\beta}_o)\} = \Sigma_X^{-1} \otimes \Lambda\Sigma_o\Lambda_o = \Sigma_X^{-1} \otimes \Sigma = \text{avar}\{\sqrt{n}\text{vec}(\widetilde{\beta})\}$. Equality with $\text{avar}\{\sqrt{n}\text{vec}(\widehat{\beta})\}$ will follow if we show that $(I_{pr}, 0)Q_{H_2(J)}H_1 = 0$. Equivalently, we need to show that $(I_{pr}, 0)H_2(H_2 JH_2)^\dagger H_2^T JH_1 = (I_{pr}, 0)H_1$, which holds if and only if $(I_{pr}, 0)D_\Lambda G_o(G_o^T J_o G_o)^\dagger G_o^T D_\Lambda^T JH_1 = (I_{pr}, 0)H_1$. Cook et al. (2010) show that $(I_{pr}, 0)G_o(G_o^T J_o G_o)^\dagger G_o^T$ is a row block matrix with first block block $\Sigma_X^{-1} \otimes \Sigma_o$ and second block 0. The rest of the proof follows by carrying out the necessary algebra. □

REFERENCES

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

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

COOK, R. D. & YIN, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics* **43**, 147–199.

HARVILLE, D. A. (2008). *Matrix Algebra from a Statistician's Perspective*. New York: Springer-Verlag.

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

HUBER, P. J. (1973). Robust regression: asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

JOHNSON, R. A. & WICHERN, D. W. (2007). *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall, 6th ed.

NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *The Annals of Statistics* , 758–765.

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

SU, Z. & COOK, R. (2012). Inner envelopes: Efficient estimation in multivariate linear regression. *Biometrika* **99**, 687–702.

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937–950.

[*Received January* 2011. *Revised June* 2011]