Biometrika (2012), **99**, 3, *pp.* 687–702

Printed in Great Britain

doi: 10.1093/biomet/ass024
Advance Access publication 24 June 2012

Inner envelopes: efficient estimation in multivariate linear regression

By ZHIHUA SU and R. DENNIS COOK

School of Statistics, University of Minnesota, 224 Church St. S.E., Minneapolis, Minnesota 55455, U.S.A.

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

Summary

In this article we propose a new model, called the inner envelope model, which leads to efficient estimation in the context of multivariate normal linear regression. The asymptotic distribution and the consistency of its maximum likelihood estimators are established. Theoretical results, simulation studies and examples all show that the efficiency gains can be substantial relative to standard methods and to the maximum likelihood estimators from the envelope model introduced recently by Cook et al. (2010). Compared to the envelope model, the inner envelope model is based on a different construction and it can produce substantial efficiency gains in situations where the envelope model offers no gains. In effect, inner envelopes open a new frontier to the way in which reducing subspaces can be used to improve efficiency in multivariate problems.

Some key words: Dimension reduction; Envelope model; Grassmann manifold; Reducing subspace.

1. Introduction

The classical multivariate linear regression model is formulated as follows:

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

where $Y \in \mathbb{R}^r$ is a random response vector, $X \in \mathbb{R}^p$ is a nonstochastic vector of predictors centred to have sample mean zero, the errors $\varepsilon \in \mathbb{R}^r$ are normally distributed with mean zero and unknown covariance matrix $\Sigma > 0$, and errors from different samples are independent. The coefficient matrix $\beta \in \mathbb{R}^{r\times p}$ has rank p and is unknown, and $\alpha \in \mathbb{R}^r$ is an unknown intercept. Our interest lies in the estimation of β. We use n to denote the sample size and assume that $p < r < n$.

An important step forward in efficient estimation of β comes from the work of Cook et al. (2010), in which a new class of models, called envelope models, was proposed. Envelope models are based on the idea that a projection of the response vector Y may be immaterial to the goal of estimating β, while still contributing extraneous variation that causes the estimator of β to be more variable than otherwise. Envelope estimation accounts for such extraneous variation, making the estimator of β potentially much more efficient. The partial envelope model proposed by Su & Cook (2011) is a generalization of the envelope model. It allows for a projection of Y that is immaterial to the goal of estimating a subset β_1 of the columns of β. Because they can be tailored, partial envelopes can lead to greater efficiency gains than envelope models for the purpose of estimating β_1.

The efficiency gains offered by envelopes and partial envelopes depend on the presence of immaterial projections of Y. We offer another route to pursue efficiency gains based on

envelopes, so that gains might still be achieved when the entire vector Y is material to the estimation of β. In particular, we (a) propose a new inner envelope model and demonstrate theoretically and by simulation that its maximum likelihood estimator of β can increase efficiency well beyond that available from standard methods and from envelope estimators; and (b) establish consistency results to show that the inner envelope estimators are robust to deviations from normality, which has not been studied previously for envelope models.

The following notation and definitions will be used in our discussion. For positive integers p and q, $\mathbb{R}^{p\times q}$ denotes the class of all $p \times q$ matrices, and $\mathbb{S}^{p\times p}$ denotes the class of all symmetric $p \times p$ matrices. We use P_A to indicate a projection operator onto A or span(A) if A is a space or a matrix, and $Q_A = I - P_A$. The symbol $\sim$ means identically distributed, and $U \perp\!\!\!\perp V \mid X$ indicates the conditional independence of U and V given X. For a subspace $\mathcal{V}$, $\mathcal{V}^\perp$ stands for its orthogonal complement relative to the usual inner product. A basis matrix for $\mathcal{V}$ is any matrix whose columns form a basis for $\mathcal{V}$. The sum of spaces is defined as $\mathcal{V}_1 + \mathcal{V}_2 = \{v_1 + v_2; v_1 \in \mathcal{V}_1, v_2 \in \mathcal{V}_2\}$, and with a matrix $A \in \mathbb{R}^{p\times p}$ and a subspace $\mathcal{V} \subseteq \mathbb{R}^p$, $A\mathcal{V} = \{Av : v \in \mathcal{V}\}$. For matrices $A \in \mathbb{R}^{r\times r}$ and $B \in \mathbb{R}^{r\times r}$, the subspace $\mathcal{S}_d(A, B)$ is the span of $A^{-1/2}$ times the first d eigenvectors of $A^{-1/2}BA^{-1/2}$. The spectral norm of a matrix A is denoted by $\|A\|$ and the Moore–Penrose inverse of A is denoted as $A^\dagger$. The notation $\mathbb{G}^{r\times d}$ is reserved for the Grassmann manifold of dimension d in $\mathbb{R}^r$, which is the set of all d dimensional subspaces in $\mathbb{R}^r$. A matrix $A \in \mathbb{R}^{p\times q}$ is semi-orthogonal when it is column orthogonal, $A^{\mathrm{T}}A = I_q$, and we call A_0 its completion if $(A, A_0) \in \mathbb{R}^{p\times p}$ is an orthogonal matrix. The vector operator, vec, stacks the columns of a matrix into a vector and the vector half operator, vech, stacks elements from the upper triangular or lower triangular part of a symmetric matrix into a vector columnwise.

2. ENVELOPES

Suppose there is a subspace $\mathcal{S} \subseteq \mathbb{R}^p$ that has the following two properties,

$$\text{(2a) } Q_{\mathcal{S}}Y \mid X \sim Q_{\mathcal{S}}Y, \quad \text{(2b) } Q_{\mathcal{S}}Y \perp\!\!\!\perp P_{\mathcal{S}}Y \mid X. \tag{2}$$

Property (2a) means that the distribution of $Q_{\mathcal{S}}Y$ does not depend on X, so marginally $Q_{\mathcal{S}}Y$ carries no information about β. Property (2b) means that $Q_{\mathcal{S}}Y$ is conditionally independent of $P_{\mathcal{S}}Y$ given X and thus $Q_{\mathcal{S}}Y$ cannot convey information about β through an association with $P_{\mathcal{S}}Y$. The properties given at (2) are equivalent to the single condition $Q_{\mathcal{S}}Y \mid (P_{\mathcal{S}}, X) \sim Q_{\mathcal{S}}Y$. This structure then implies that the projection $Q_{\mathcal{S}}Y$ is immaterial to the estimation of β. All of the immaterial information in Y can be obtained by finding the smallest subspace $\mathcal{S}$ that satisfies the requirements in (2).

Let $\mathcal{B} = \text{span}(\beta)$. Cook et al. (2010) showed that the pair of conditions (2) is equivalent to the pair of conditions

$$\text{(3a) } \mathcal{B} \subseteq \mathcal{S}, \quad \text{(3b) } \Sigma = P_{\mathcal{S}}\Sigma P_{\mathcal{S}} + Q_{\mathcal{S}}\Sigma Q_{\mathcal{S}}. \tag{3}$$

Condition (3b) holds if and only if $P_{\mathcal{S}}Y$ and $Q_{\mathcal{S}}Y$ are uncorrelated given X, and it is equivalent to requiring that $\mathcal{S}$ be a reducing subspace of Σ. Together these conditions imply that we can obtain all of the immaterial information by selecting $\mathcal{S}$ 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 by $\mathcal{E}_\Sigma(\mathcal{B})$; $\mathcal{E}_\Sigma(\mathcal{B})$ is shortened to $\mathcal{E}$ when used as a subscript. The projection of Y that is immaterial to the estimation of β is then given uniquely by $Q_{\mathcal{E}}Y$, while $P_{\mathcal{E}}Y$ is material to the same purpose.

Let $u = \dim\{\mathcal{E}_\Sigma(\mathcal{B})\}$, and let $\Gamma \in \mathbb{R}^{r\times u}$ and $\Gamma_0 \in \mathbb{R}^{r\times(r-u)}$ denote semi-orthogonal basis matrices for $\mathcal{E}_\Sigma(\mathcal{B})$ and $\mathcal{E}_\Sigma^\perp(\mathcal{B})$. The coordinate form of the envelope model can now be obtained by imposing conditions (3) on the standard model (1):

$$Y = \alpha + \Gamma\eta X + \varepsilon, \quad \Sigma = \Gamma\Omega\Gamma^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}, \tag{4}$$

where $\eta \in \mathbb{R}^{u\times p}$ holds the coordinates of β relative to Γ, and $\Omega \in \mathbb{R}^{u\times u}$ and $\Omega_0 \in \mathbb{R}^{(r-u)\times(r-u)}$ are positive definite matrices. As can be seen in model (4), $\mathcal{E}_\Sigma(\mathcal{B})$ links the mean and covariance structures and it is this link that provides the efficiency gains. The gains can be great when variation of the immaterial data $\Gamma_0^{\mathrm{T}}Y$ is substantially larger than that of the material data $\Gamma^{\mathrm{T}}Y$; for instance, when $\|\Omega\| \ll \|\Omega_0\|$ (Cook et al., 2010). A schematic showing how an envelope increases efficiency was given by Su & Cook (2011).

The partial envelope model (Su & Cook, 2011) is a generalization of the envelope model that was designed for regressions in which the coefficients of some predictors are of special interest. Partition X into $X_1 \in \mathbb{R}^{p_1}$ and $X_2 \in \mathbb{R}^{p_2}$, $p_1 + p_2 = p$, with corresponding partition of $\beta = (\beta_1, \beta_2)$, where $\beta_1 \in \mathbb{R}^{r\times p_1}$ and $\beta_2 \in \mathbb{R}^{r\times p_2}$, and suppose that the goal is to estimate the coefficients β_1 of X_1. With this change of objective, the logic underlying the development of the partial envelope model parallels that for the envelope model. Let $\mathcal{E}_\Sigma(\mathcal{B}_1)$ denote the smallest reducing subspace of Σ that contains $\mathcal{B}_1 = \mathrm{span}(\beta_1)$, let $u_1 = \dim\{\mathcal{E}_\Sigma(\mathcal{B}_1)\}$, and let $\Gamma \in \mathbb{R}^{r\times u_1}$ and $\Gamma_0 \in \mathbb{R}^{r\times(r-u_1)}$ be semi-orthogonal basis matrices for $\mathcal{E}_\Sigma(\mathcal{B}_1)$ and $\mathcal{E}_\Sigma^{\perp}(\mathcal{B}_1)$. Then the partial envelope model is

$$Y = \alpha + \Gamma\eta X_1 + \beta_2 X_2 + \varepsilon, \quad \Sigma = \Gamma\Omega\Gamma^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}, \tag{5}$$

where $\eta \in \mathbb{R}^{u_1\times p_1}$, $\Omega \in \mathbb{R}^{u_1\times u_1}$ and $\Omega_0 \in \mathbb{R}^{(r-u_1)\times(r-u_1)}$. If $X_1 = X$, then $\mathcal{B}_1 = \mathcal{B}$, $\mathcal{E}_\Sigma(\mathcal{B}_1) = \mathcal{E}_\Sigma(\mathcal{B})$ and model (5) is the same as model (4). This partial envelope model focuses on the relatively narrow goal of estimating β_1 and consequently has the potential to achieve efficiency gains beyond those available by using the envelope model (4). This happens because $\mathcal{E}_\Sigma(\mathcal{B}_1)$ is often a proper subset of $\mathcal{E}_\Sigma(\mathcal{B})$, so more immaterial information is ruled out by partial envelopes which makes them more efficient.

The envelope model relies on the presence of immaterial data to achieve efficiency gains. However, if $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$ then there is no immaterial information and enveloping offers no gains. The inner envelope model introduced in § 3 has the potential to achieve efficiency gains when envelopes offer no benefits.

3. Inner envelopes

3·1. *Motivation*

In the absence of immaterial information $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$, the envelope model (4) reduces to the standard model (1) and no efficiency gains are achieved. However, in some regressions we may still be able to improve efficiency by relaxing the base requirements (2). Consider a multivariate regression in which the response can be decomposed into its projections onto three orthogonal subspaces $\mathcal{S}_1$, $\mathcal{S}_2$ and $\mathcal{S}_3$ of $\mathbb{R}^r$, $Y = P_1Y + P_2Y + P_3Y$, with the properties that

$$\text{(6a) } P_3Y \mid X \sim P_3Y, \quad \text{(6b) } P_1Y \perp\!\!\!\perp (P_2Y, P_3Y) \mid X, \tag{6}$$

where $P_j = P_{\mathcal{S}_j}$ $(j = 1, 2, 3)$. The distribution of $P_3Y \mid X$ is independent of X, while the distributions of $P_1Y \mid X$ and $P_2Y \mid X$ are allowed to depend on X. If we could find such a decomposition with $P_2Y = 0$ then (6) would reduce to (2) with $P_{\mathcal{S}} = P_1$ and $Q_{\mathcal{S}} = P_3$, and we could employ an envelope model (4). Otherwise, P_2Y represents a confounder whose distribution depends on X and is correlated with P_3Y.

Nevertheless, the structure in (6) still allows for efficiency gains. Condition (6a) implies that $\beta = P_1\beta + P_2\beta$. If we can estimate $P_1\beta$ with greater precision than possible with the standard model and estimate $P_2\beta$ with about the same precision, then overall, we may get better efficiency in estimating β. This is the basic idea for the inner envelope model. To ensure that there is no projection of P_1Y whose distribution is independent of X, we require that $\mathcal{S}_1 \subseteq \mathcal{B}$. Condition

(6b) holds if and only if $\mathcal{S}_1$ is a reducing subspace of Σ. Consequently, we obtain the following constraints, which parallel those in (3):

$$\text{(7a) } \mathcal{S}_1 \subseteq \mathcal{B}, \quad \text{(7b) } \Sigma = P_1 \Sigma P_1 + Q_1 \Sigma Q_1. \tag{7}$$

To achieve the most gains subject to this structure, we maximize $\dim\{\text{span}(\mathcal{S}_1)\}$ so that the greatest part of β can be estimated efficiently. We call the largest reducing subspace of Σ contained within $\mathcal{B}$ the inner Σ-envelope, as formally defined in § 3·2.

3·2. *Definitions*

DEFINITION 1. *Let $M \in \mathbb{S}^{r\times r}$. The inner M-envelope of the subspace $\mathcal{V} \subseteq \mathbb{R}^r$, denoted by $\mathcal{IE}_M(\mathcal{V})$, is the reducing subspace of M with maximal dimension that is contained within $\mathcal{V}$.*

The existence of inner envelopes is ensured because the space with only one element span(0) is a reducing subspace of M that is contained within $\mathcal{V}$. We next state two characterizing propositions.

PROPOSITION 1. *Let $M \in \mathbb{S}^{r\times r}$. Then $\mathcal{IE}_M(\mathcal{V}) = \sum_i \mathcal{V}_i$, where the sum is over all reducing subspaces $\mathcal{V}_i$ of M that are contained in $\mathcal{V}$.*

PROPOSITION 2. *Let $M \in \mathbb{S}^{r\times r}$. Then $\mathcal{IE}_M(\mathcal{V}) = \mathcal{E}_M^{\perp}(\mathcal{V}^{\perp})$.*

Proposition 1 is a natural consequence of the definition, which states that the inner envelope contains all the reducing subspaces of Σ that are contained in $\mathcal{B}$. Proposition 2 builds a connection between inner envelopes and envelopes; that is, an inner M-envelope of a subspace is the same as the orthogonal complement of the M-envelope of its orthogonal complement. Propositions 1 and 2 ensure that the inner envelope is uniquely defined as the largest subspace $\mathcal{S}_1$ that satisfies (7). Proofs of these propositions are in the Appendix.

Like the envelope model, the coordinate form of the inner envelope model is expressed in terms of semi-orthogonal basis matrices $\Gamma_1 \in \mathbb{R}^{r\times d}$ and $\Gamma_0 \in \mathbb{R}^{r\times(r-d)}$ for $\mathcal{IE}_\Sigma(\mathcal{B})$ and $\mathcal{IE}_\Sigma^{\perp}(\mathcal{B})$. We shorten $\mathcal{IE}_\Sigma(\mathcal{B})$ to $\mathcal{IE}$ for subscripts and let $d = \dim\{\mathcal{IE}_\Sigma(\mathcal{B})\}$. Then we can write $\beta = P_{\mathcal{IE}}\beta + Q_{\mathcal{IE}}\beta = \Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}}$, where $B \in \mathbb{R}^{(r-d)\times(p-d)}$ is a semi-orthogonal matrix so that $\Gamma_0 B$ is a semi-orthogonal basis matrix for $Q_{\mathcal{IE}}\mathcal{B}$, $\eta_1^{\mathrm{T}} \in \mathbb{R}^{d\times p}$ and $\eta_2^{\mathrm{T}} \in \mathbb{R}^{(p-d)\times p}$. Written in terms of the motivating conditions (6) and (7), $\mathcal{S}_1 = \mathcal{IE}_\Sigma(\mathcal{B}) = \text{span}(\Gamma_1)$, $\mathcal{S}_2 = \text{span}(\Gamma_0 B)$ and $\mathcal{S}_3 = \text{span}^{\perp}(\Gamma_1, \Gamma_0 B)$ with dimensions d, $p-d$ and $r-p$. Condition (7a) implies that $(\eta_1, \eta_2) \in \mathbb{R}^{p\times p}$ has full rank. The inner envelope model can now be stated in full as

$$Y = \alpha + (\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})X + \varepsilon, \quad \Sigma = \Gamma_1\Omega_1\Gamma_1^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}, \tag{8}$$

where $\Omega_1 \in \mathbb{R}^{d\times d}$ and $\Omega_0 \in \mathbb{R}^{(r-d)\times(r-d)}$ are positive definite matrices. If $d = 0$, then $\mathcal{IE}_\Sigma(\mathcal{B}) = \text{span}(0)$ and (8) reduces to the standard model. If $d = p$, then $\mathcal{IE}_\Sigma(\mathcal{B}) = \mathcal{B}$ and (8) reduces to the envelope model $\mathcal{IE}_\Sigma(\mathcal{B}) = \mathcal{E}_\Sigma(\mathcal{B})$.

For example, suppose we are comparing group means for three multivariate normal populations on $r = 3$ characteristics Y_1, Y_2 and Y_3. The predictor $X \in \mathbb{R}^2$ is composed of two group indicators x_i $(i = 1, 2)$, each taking value 1 for the ith group, and 0 otherwise. The coefficient matrix $\beta = (\beta_1, \beta_2)$ is a 3×2 matrix, with $\beta_1 = E(Y \mid x_1 = 1, x_2 = 0) - E(Y \mid x_1 = 0, x_2 = 0)$ and $\beta_2 = E(Y \mid x_1 = 0, x_2 = 1) - E(Y \mid x_1 = 0, x_2 = 0)$. Let $\lambda_1 < \lambda_2 < \lambda_3$ be the three distinct eigenvalues of Σ with corresponding eigenvectors v_1, v_2 and v_3. Then $\Sigma = \lambda_1 v_1 v_1^{\mathrm{T}} + \lambda_2 v_2 v_2^{\mathrm{T}} + \lambda_3 v_3 v_3^{\mathrm{T}}$. If β_1 aligns with v_1 and if $\beta_2 = a v_2 + b v_3$, where $a \neq 0$ and $b \neq 0$, then $\mathcal{IE}_\Sigma(\mathcal{B}) = \text{span}(v_1)$ and $\mathcal{IE}_\Sigma^{\perp}(\mathcal{B}) = \text{span}(v_2, v_3)$. The envelope model (4) reduces to the standard model in this illustration since $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^3$. However, the partial envelope $\mathcal{E}_\Sigma(\mathcal{B}_1)$ coincides with the inner envelope $\mathcal{IE}_\Sigma(\mathcal{B})$.

3·3. *Maximum likelihood estimators*

We assume that the data (Y_i, X_i) are independent observations on $Y \mid X = X_i$ $(i = 1, \ldots, n)$. The derivation of the maximum likelihood estimators is based on the coordinate version of the inner envelope model (8). As X is centred, the maximum likelihood estimator of α is $\bar{Y}$, the sample mean of the Y_is. As shown in the Appendix, given a fixed dimension d, a semi-orthogonal basis matrix $\hat{\Gamma}_1$ for the maximum likelihood estimator of $\mathcal{IE}_\Sigma(\mathcal{B})$ can be obtained by minimizing the following function of G_1 over the Grassmann manifold $\mathbb{G}^{r\times d}$,

$$\log|G_1^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}G_1| + \log|G_1^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}^{-1}G_1| + \sum_{i=p-d+1}^{r-d}\log\{1+\tilde{\lambda}_i(G_0)\}, \tag{9}$$

where $G_1 \in \mathbb{R}^{r\times d}$ is a semi-orthogonal matrix, $(G_1, G_0) \in \mathbb{R}^{r\times r}$ is an orthogonal matrix, and $\hat{\Sigma}_{\mathrm{fit}}$ and $\hat{\Sigma}_{\mathrm{res}}$ are the sample covariance matrices of the fitted vectors and residual vectors from the ordinary least squares fit of Y on X. The $\tilde{\lambda}_i(G_0)$s are the ordered, descending eigenvalues of $(G_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}G_0)^{-1/2}(G_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{fit}}G_0)(G_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}G_0)^{-1/2}$. For later use, we denote its matrices of ordered eigenvectors and eigenvalues as $\tilde{V}(G_0)$ and $\tilde{\Lambda}(G_0) = \mathrm{diag}\{\tilde{\lambda}_1(G_0), \cdots, \tilde{\lambda}_{r-d}(G_0)\}$, and let $\tilde{K}(G_0) = \mathrm{diag}\{0, \cdots, 0, \tilde{\lambda}_{p-d+1}(G_0), \cdots, \tilde{\lambda}_{r-d}(G_0)\}$. After obtaining $\hat{\Gamma}_1$, $\hat{\Gamma}_0$ is constructed as any semi-orthogonal basis matrix for $\mathrm{span}^{\perp}(\hat{\Gamma}_1)$. The maximum likelihood estimators of the remaining parameters are as given in the following list. In preparation, let F denote the $n \times p$ matrix with ith row X_i^{T}, let U be the $n \times r$ matrix with ith row $(Y_i - \bar{Y})^{\mathrm{T}}$, and let $\hat{\beta}_{\mathrm{sm}}$ denote the maximum likelihood estimator of β under the standard model. Then

$$\begin{aligned}
\hat{\eta}_1^{\mathrm{T}} &= \hat{\Gamma}_1^{\mathrm{T}}\hat{\beta}_{\mathrm{sm}},\\
\hat{\Omega}_1 &= (U\hat{\Gamma}_1 - F\hat{\eta}_1)^{\mathrm{T}}(U\hat{\Gamma}_1 - F\hat{\eta}_1)/n,\\
\hat{\Omega}_0 &= \hat{\Gamma}_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\hat{\Gamma}_0 + (\hat{\Gamma}_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\hat{\Gamma}_0)^{1/2}\tilde{V}(\hat{\Gamma}_0)\tilde{K}(\hat{\Gamma}_0)\tilde{V}(\hat{\Gamma}_0)^{\mathrm{T}}(\hat{\Gamma}_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\hat{\Gamma}_0)^{1/2},\\
\mathrm{span}(\hat{B}) &= \hat{\Omega}_0\mathcal{S}_{p-d}(\hat{\Omega}_0, \hat{\Gamma}_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{fit}}\hat{\Gamma}_0),\\
\hat{\eta}_2^{\mathrm{T}} &= (\hat{B}^{\mathrm{T}}\hat{\Omega}_0^{-1}\hat{B})^{-1}\hat{B}^{\mathrm{T}}\hat{\Omega}_0^{-1}\hat{\Gamma}_0^{\mathrm{T}}\hat{\beta}_{\mathrm{sm}},\\
\hat{\beta} &= \hat{\Gamma}_1\hat{\eta}_1^{\mathrm{T}} + \hat{\Gamma}_0\hat{B}\hat{\eta}_2^{\mathrm{T}} = P_{\hat{\Gamma}_1}\hat{\beta}_{\mathrm{sm}} + \hat{\Gamma}_0 P_{\hat{B}(\hat{\Omega}_0^{-1})}\hat{\Gamma}_0^{\mathrm{T}}\hat{\beta}_{\mathrm{sm}},\\
\hat{\Sigma} &= \hat{\Gamma}_1\hat{\Omega}_1\hat{\Gamma}_1^{\mathrm{T}} + \hat{\Gamma}_0\hat{\Omega}_0\hat{\Gamma}_0^{\mathrm{T}}.
\end{aligned}$$

Details of the development are provided in the Appendix. The structure of $\hat{\beta}$ is consistent with the discussion in § 3·1. It has two terms, the first obtained by projecting $\hat{\beta}_{\mathrm{sm}}$ onto the estimated inner envelope $\mathrm{span}(\hat{\Gamma}_1)$. For the second term, if we multiply both sides of model (8) by Γ_0^{T}, we have a reduced rank regression with coefficient $B\eta_2^{\mathrm{T}}$ and covariance matrix Ω_0. By a result of Cook & Forzani (2008), and with Γ_0 replaced by its estimator $\hat{\Gamma}_0$, the estimator of the coefficients in this reduced rank regression has the form $P_{\hat{B}(\hat{\Omega}_0^{-1})}\hat{\Gamma}_0^{\mathrm{T}}\hat{\beta}_{\mathrm{sm}}$. Then $\hat{\Gamma}_0\hat{B}\hat{\eta}_2^{\mathrm{T}}$ has the form $\hat{\Gamma}_0 P_{\hat{B}(\hat{\Omega}_0^{-1})}\hat{\Gamma}_0^{\mathrm{T}}\hat{\beta}_{\mathrm{sm}}$. When $d = 0$, we have $\hat{\beta} = \hat{\beta}_{\mathrm{sm}}$.

3·4. *Consistency of the maximum likelihood estimators without normality*

Recalling the discussion in § 2, the definition of an envelope does not require normality. While the maximum likelihood estimators were derived under the normality assumption, a natural concern is for the robustness of the estimators when this assumption fails. In this section, we will show that $\hat{\beta}$ and $\hat{\Sigma}$ are Fisher consistent and also $\sqrt{n}$ consistent with minimal constraints on the error distribution.

PROPOSITION 3. *Under the inner envelope model* (8)*, assume that the errors are independent but not necessarily normal and have finite second moments. Then the following quantities converge in probability:*

$$\hat{\Sigma}_Y \to \Sigma_Y = \Gamma_1\Omega_1\Gamma_1^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}} + (\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})\Sigma_X(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}},$$
$$\hat{\Sigma}_{\mathrm{fit}} \to \Sigma_{\mathrm{fit}} = (\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})\Sigma_X(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}},$$
$$\hat{\Sigma}_{\mathrm{res}} \to \Sigma_{\mathrm{res}} = \Gamma_1\Omega_1\Gamma_1^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}},$$

where Σ_{fit}*,* Σ_{res} *and* Σ_Y *are the population versions of* $\hat{\Sigma}_{\mathrm{fit}}$*,* $\hat{\Sigma}_{\mathrm{res}}$*, and the sample covariance matrix of* Y $\hat{\Sigma}_Y$ *and* Σ_X *is the limit as* $n \to \infty$ *of the sample covariance matrix of* X*.*

The objective function (9) is equivalent to

$$L_{\mathrm{inner}}(G_1) = \log|G_1^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}G_1| + \log|G_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}G_0| + \sum_{i=p-d+1}^{r-d}\log\{1+\tilde{\lambda}_i(G_0)\}.$$

As n increases, $L_{\mathrm{inner}}(G_1)$ converges in probability to

$$\tilde{L}_{\mathrm{inner}}(G_1) = \log|G_1^{\mathrm{T}}\Sigma_{\mathrm{res}}G_1| + \log|G_0^{\mathrm{T}}\Sigma_{\mathrm{res}}G_0| + \sum_{i=p-d+1}^{r-d}\log\{1+\breve{\lambda}_i(G_0)\},$$

which is the population version of $L_{\mathrm{inner}}(G_1)$, and $\breve{\lambda}_i$ denotes the ith eigenvalue of $(G_0^{\mathrm{T}}\Sigma_{\mathrm{res}}G_0)^{-1/2}(G_0^{\mathrm{T}}\Sigma_{\mathrm{fit}}G_0)(G_0^{\mathrm{T}}\Sigma_{\mathrm{res}}G_0)^{-1/2}$.

PROPOSITION 4. *Assume that the conditions in Proposition* 3 *hold, and further assume that the subspace which minimizes* $\tilde{L}_{\mathrm{inner}}$ *is unique. Then* $\Gamma_1 = \arg\min_{G_1}\tilde{L}_{\mathrm{inner}}(G_1)$*, where* Γ_1 *is any semi-orthogonal basis matrix for* $\mathcal{IE}_\Sigma(\mathcal{B})$*.*

The preceding proposition says that even when the errors are not normally distributed, the maximum likelihood estimator of $\mathcal{IE}_\Sigma(\mathcal{B})$ is Fisher consistent. This proposition lays the foundation for the Fisher consistency of $\hat{\beta}$ and $\hat{\Sigma}$, which is stated in Theorem 1.

THEOREM 1. *Assume that the conditions in Proposition* 4 *hold. Then* $\hat{\beta}$ *and* $\hat{\Sigma}$ *are Fisher consistent estimators of* β *and* Σ*.*

In the next theorem we describe asymptotic properties of $\hat{\beta}$ and $\hat{\Sigma}$. In preparation, we denote the ith diagonal element of the projection matrix $P_F = F(F^{\mathrm{T}}F)^{-1}F^{\mathrm{T}}$ as p_{ii}, and we require that

$$\max_{i\leqslant n} p_{ii} \to 0, \quad n \to \infty \tag{10}$$

for establishing consistency (Huber, 1973). Condition (10) is a condition on the explanatory design and it means that maximum leverage tends to zero as $n \to \infty$.

THEOREM 2. *If the errors* ε *have finite fourth moments and that* (10) *holds, then*

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

is asymptotically normally distributed, and $\hat{\beta}$ *and* $\hat{\Sigma}$ *are* $n^{1/2}$ *consistent estimators of* β *and* Σ*.*

Theorem 2 justifies the consistency of $\hat{\beta}$ and $\hat{\Sigma}$ without assuming any error distribution and points out that the convergence rate is $\sqrt{n}$. Proofs of Propositions 3, 4 and Theorems 1, 2 are in the Appendix.

3·5. *Asymptotic distributions with normality*

In this section, we give the asymptotic distribution for $\{\mathrm{vec}^{\mathrm{T}}(\hat{\beta}), \mathrm{vech}^{\mathrm{T}}(\hat{\Sigma})\}^{\mathrm{T}}$ assuming normal errors and, as the general form of the asymptotic variance of $\hat{\beta}$ seems too complicated to interpret directly, we will look at a special case to provide some intuition.

In preparation for stating the limiting distribution, we use avar to denote an asymptotic covariance matrix; that is, if $\sqrt{n}(\hat{\theta} - \theta) \to N(0, A)$, then $\mathrm{avar}(\sqrt{n}\hat{\theta}) = A$. Following Henderson & Searle (1979), $C_r \in \mathbb{R}^{r(r+1)/2 \times r^2}$ 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 $\mathrm{vec}(A) = E_r \mathrm{vech}(A)$, $K_{st} \in \mathbb{R}^{st \times st}$ is the commutation matrix and for matrix $A \in \mathbb{R}^{s \times t}$, $\mathrm{vec}(A^{\mathrm{T}}) = K_{st}\mathrm{vec}(A)$.

We use Proposition 4.1 of Shapiro (1986) to account for the overparameterization in Γ_1 and to obtain the asymptotic distribution for $\{\mathrm{vec}^{\mathrm{T}}(\hat{\beta}), \mathrm{vech}^{\mathrm{T}}(\hat{\Sigma})\}^{\mathrm{T}}$. Although the details are different, the proof parallels Theorem 5.1 of Cook et al. (2010) and so is omitted.

THEOREM 3. *Under model* (8),

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\hat{\beta}) - \mathrm{vec}(\beta) \\ \mathrm{vech}(\hat{\Sigma}) - \mathrm{vech}(\Sigma) \end{pmatrix} \to N_{rp+r(r+1)/2}(0, V_0)$$

in distribution, where $V_0 = H(H^{\mathrm{T}}JH)^{\dagger}H^{\mathrm{T}}$, J *is the Fisher information under the standard model* (1)

$$J = \begin{pmatrix} \Sigma_X \otimes \Sigma^{-1} & 0 \\ 0 & 2^{-1}E_r^{\mathrm{T}}(\Sigma^{-1} \otimes \Sigma^{-1})E_r \end{pmatrix}$$

and H *is the gradient matrix* $\partial\{\mathrm{vec}^{\mathrm{T}}(\hat{\beta}), \mathrm{vech}^{\mathrm{T}}(\hat{\Sigma})\}^{\mathrm{T}}/\partial\phi^{\mathrm{T}}$, *where*

$$\phi = \{\mathrm{vec}(\eta_1)^{\mathrm{T}}, \mathrm{vec}(\eta_2)^{\mathrm{T}}, \mathrm{vec}(B)^{\mathrm{T}}, \mathrm{vec}(\Gamma_1)^{\mathrm{T}}, \mathrm{vech}(\Omega_1)^{\mathrm{T}}, \mathrm{vech}(\Omega_0)^{\mathrm{T}}\}^{\mathrm{T}},$$

and H *equals*

$$\begin{pmatrix} I_p \otimes \Gamma_1 & I_p \otimes \Gamma_0 B & \eta_2 \otimes \Gamma_0 & \eta_1 \otimes I_r - (\eta_2 B^{\mathrm{T}}\Gamma_0^{\mathrm{T}} \otimes \Gamma_1)K_{rd} & 0 & 0 \\ 0 & 0 & 0 & 2C_r(\Gamma_1\Omega_1 \otimes I_r - \Gamma_1 \otimes \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}) & C_r(\Gamma_1 \otimes \Gamma_1)E_d & C_r(\Gamma_0 \otimes \Gamma_0)E_{r-d} \end{pmatrix}.$$

As $H(H^{\mathrm{T}}JH)^{\dagger}H^{\mathrm{T}} \leqslant J^{-1}$, *the inner envelope estimators are always asymptotically less variable than the standard estimators.*

Let V be the asymptotic variance of the standard estimators, $V = J^{-1}$. Then $V^{-1/2}(V - V_0)V^{-1/2} = Q_{J^{1/2}H} \geqslant 0$, and this indicates that the inner envelope estimators reduce the asymptotic variance by a fraction of $Q_{J^{1/2}H}$.

The asymptotic variance of $\hat{\beta}$ is the upper $rp \times rp$ block of V_0. We were unable to simplify the expression to a ponderable form. So instead, we look into a special case that gives a relatively simpler form to interpret. Assume that $\Omega_1 = \sigma^2 I_d$, $\Omega_0 = \sigma_0^2 I_{r-d}$, where σ and σ_0 are scalars, and that η_1, η_2 are in different reducing subspaces of Σ_X, which means $\eta_2^{\mathrm{T}}\Sigma_X\eta_1 = 0$ and

$\eta_1^{\mathrm{T}}\Sigma_X\eta_1 = I_d$. Substituting these conditions into V_0, avar$\{\sqrt{n}\mathrm{vec}(\hat{\beta})\}$ has the form

$$\sigma^2\Sigma_X^{-1}\otimes\Gamma_1\Gamma_1^{\mathrm{T}} + \sigma_0^2\{\Sigma_X^{-1}\otimes\Gamma_0 B B^{\mathrm{T}}\Gamma_0^{\mathrm{T}} + \eta_2(\eta_2^{\mathrm{T}}\Sigma_X\eta_2)^{-1}\eta_2^{\mathrm{T}}\otimes\Gamma_0 B_0 B_0^{\mathrm{T}}\Gamma_0^{\mathrm{T}}\}$$
$$+ k^{-1}(4\eta_2\eta_2^{\mathrm{T}}\otimes\Gamma_1\Gamma_1^{\mathrm{T}} + \eta_1\eta_1^{\mathrm{T}}\otimes\Gamma_0 B_0 B_0^{\mathrm{T}}\Gamma_0^{\mathrm{T}}) - c\eta_1\eta_1^{\mathrm{T}}\otimes\Gamma_0 B_0 B_0^{\mathrm{T}}\Gamma_0^{\mathrm{T}}, \tag{11}$$

where $k = \sigma^2/\sigma_0^2 + \sigma_0^2/\sigma^2 - 2$, $c = 1/(k^2\sigma_0^2 + k)$ and (B, B_0) is an orthogonal matrix.

When $d = p$, based on the discussion in § 3·1, the inner envelope model is an envelope model with $u = p$. Then we have $\eta_2 = 0$ and $B = 0$, $B_0 = I_{r-d}$, and (11) reduces to $\sigma^2\Sigma_X^{-1}\otimes\Gamma_1\Gamma_1^{\mathrm{T}} + (k^{-1} - c)\eta_1\eta_1^{\mathrm{T}}\otimes\Gamma_0\Gamma_0^{\mathrm{T}}$, which is exactly the same as avar$\{\sqrt{n}\mathrm{vec}(\hat{\beta})\}$ under the envelope model; see equation (5.7) in Cook et al. (2010). When $d = 0$, the inner envelope model reduces to the standard model. Then $\Gamma_1 = 0$, $\Gamma_0 = I_r$, $\eta_1 = 0$, $B_0 = 0$, $B = I_r$, and (11) has the form $\sigma_0^2\Sigma_X^{-1}\otimes I_r$, which is the same as the asymptotic variance under the standard model.

A comparison of the asymptotic variance given in (11) to the asymptotic variance of the standard estimator indicates that we might expect efficiency gains when d is close to p, so the dimension of the inner envelope is relatively large, or when r is large relative to d.

3·6. *Selection of d*

Many parameter selection methods can be used to select d, the dimension of $\mathcal{IE}_\Sigma(\mathcal{B})$. In this section, we describe methods which worked well in numerical experiments. The first methods are based on information criteria.

To use information criteria, we need the number of parameters in the model. For an inner envelope model with dimension d, there are $N(d) = p^2 + (p-d)(r-p) + r(r+1)/2$ parameters to be estimated. This is because we need pd parameters for η_1, $p(p-d)$ for η_2, $d(d+1)/2$, and $(r-d)(r-d+1)/2$ parameters for Ω and Ω_0 as they are symmetric matrices. We cannot estimate Γ_1 but only its span, so we are estimating span(Γ_1) on a $r\times d$ Grassmann manifold. Therefore $d(r-d)$ parameters are needed for Γ_1. It is the same with estimating B. If we fix an orthogonal basis (Γ_1, Γ_0), only the span of B can be estimated, so B is estimated on a $(r-d)\times(p-d)$ Grassmann manifold and $(p-d)(r-p)$ parameters are needed.

The maximized loglikelihood function $\hat{L}(d)$ under the inner envelope model with dimension d is

$$-\frac{nr}{2}\{1+\log(2\pi)\} - \frac{n}{2}\log|\hat{\Gamma}_1^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\hat{\Gamma}_1| - \frac{n}{2}\log|\hat{\Gamma}_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\hat{\Gamma}_0| - \frac{n}{2}\sum_{i=p-d+1}^{r-d}\log\{1+\tilde{\lambda}_i(\hat{\Gamma}_0)\}.$$

Then for a fixed d, Akaike's information criterion is $A(d) = -2\hat{L}(d) + 2N(d)$. We search d from 0 to p and choose d at the value that minimizes $A(d)$.

The Bayes information criterion for a fixed d is $B(d) = -2\hat{L}(d) + N(d)\log(n)$. Similarly, we select d by searching from 0 to p and choose d at the value that minimizes $B(d)$.

The dimension of the inner envelope model can also be determined by likelihood ratio testing. To test the hypothesis $d = d_0$ $(d_0 \leqslant p)$, the test statistic $\Lambda(d_0)$ can be constructed as $\Lambda(d_0) = 2\{\hat{L}(0) - \hat{L}(d_0)\}$. Here $\hat{L}(0)$ is the maximized value of the loglikelihood for the standard model and $\hat{L}(0) = -(nr/2)\{1+\log(2\pi)\} - (n/2)\log|\hat{\Sigma}_{\mathrm{res}}|$. Under the null hypothesis, $\Lambda(d_0)$ is asymptotically distributed as a chi-square random variable with $d_0(r-p)$ degrees of freedom, where $d_0(r-p)$ is the difference of the number of parameters between the full model and the inner envelope model with dimension d_0. The testing procedure can be started at $d = p$ with a common significance level, and we choose d at the first hypothesized value that is not rejected.

Our numerical experiments showed that the Bayes criterion tends to be preferred over Akaike's criterion and likelihood ratio testing, although all of them can perform well. For illustration we will use both Bayes and Akaike's criteria in our simulations and data analysis. The asymptotic behaviour of these criteria is similar to that in the multiple regression case discussed by Shao (1997). In the inner envelope model context, it can also be justified that the Bayes criterion will select the true model with probability tending to 1, and Akaike's will select a model that asymptotically contains the true model.

4. Simulation and data analysis results

4·1. *Simulation results*

In this section, we report results from simulation studies to provide insights into the behaviour of the inner envelope estimators. The computing of inner envelopes involves a Grassmann optimization of (9), and it can be performed numerically using MATLAB package sg_min 2.4.1 by Lippert which offers several optimization methods including Newton–Raphson iterations on a Grassmann manifold with an analytic first derivative and numerical second derivative of the objective function.

We simulated the data from (8) with $r = 10$, $p = 8$ and $\alpha = 0$. The bases of the inner envelope and its complement (Γ_1, Γ_0), and B were constructed by orthogonalizing $r \times r$ and $(r - d) \times (p - d)$ matrices of independent uniform $(0, 1)$ random variables. The eigenvalues of Σ were chosen at 1, 5, 10, 50, 100, 500, 1000, 5000, 10 000 and 50 000, and d was fixed at 1, 4 and 7, with the inner envelope basis Γ_1 associated with the first d eigenvalues. The first d columns of β were Γ_1, and the other $p - d$ columns were outside the inner envelope, with η_1 and η_2 generated with independent standard normal variables. We expect notable efficiency gains from this construction since $\mathrm{var}(P_{\mathcal{I}\mathcal{E}} Y)$ is relatively small and thus we can estimate $P_{\mathcal{I}\mathcal{E}}\beta$ with relative greater precision. The elements in X took value 0 or 100 with probability 0·5 for each. Under this construction, $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$, and envelope models offer no gains in efficiency. We estimated the actual variance of $\hat{\beta}$ by computing the sample variance of $\hat{\beta}^{(j)}$, $j = 1, \ldots, 200$, from 200 replications for sample sizes 100, 200, 300, 500, 800 and 1200. The asymptotic variance of $\hat{\beta}$ is given by Theorem 3. We also estimated the actual variance of $\hat{\beta}$ by using 200 residual bootstrap samples.

The results for $d = 7$ are summarized in Fig. 1. In each panel, the vertical axis is the standard deviation for one element of $\hat{\beta}$ and the horizontal axis is the sample size. In Fig. 1(a), which is for an element of $\hat{\beta}$ inside the inner envelope, the two lines for the actual standard deviations of the standard and inner envelope estimators are well separated for all sample sizes and the actual standard deviation of the inner envelope estimator drops below the asymptotic standard deviation of the standard model estimator at a small sample size. The efficiency gains are predicted by Theorem 3. The bootstrap estimation of the actual standard deviations is quite reliable. In Fig. 1(b), which is for an element outside of the inner envelope, the performances of the standard model and inner envelope model are almost the same. The lines for the two models are entangled for both the asymptotic and actual standard deviations. The results for $d = 4$ were quite similar to those for $d = 7$, while the difference between the inner envelope and standard estimators of the element inside the inner envelope was notably less for $d = 1$, as expected.

We did another simulation with everything the same except $p = 2$. Then there is only one nontrivial choice for d, which is $d = 1$. Here the inner envelope model shows a greater advantage over the standard model than in Fig. 1, the actual and asymptotic standard deviations of inner envelope estimators being about 0·01 times the asymptotic standard deviation of the standard model estimator. This agrees with the discussion at the end of § 3·5: when the ratio of d/p is large and d/r is small, we expect to get larger gains.

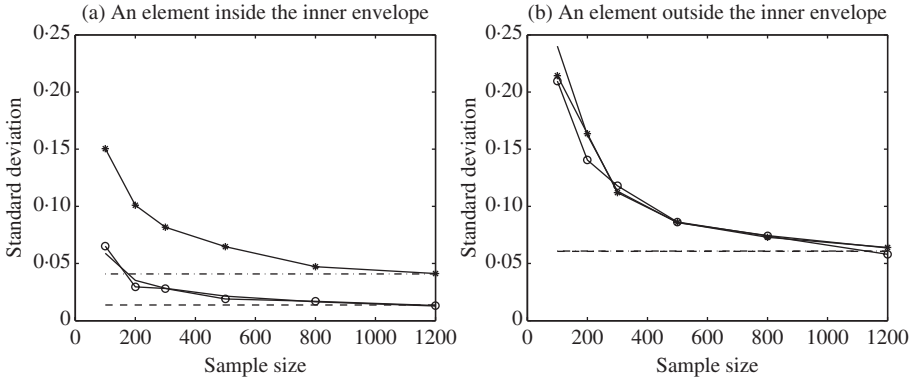

Fig. 1. Simulation results with $d = 7$, $p = 8$ on the asymptotic and actual standard deviations for elements of $\hat{\beta}$ inside and outside the inner envelope. Solid lines, the actual standard deviations of the inner envelope estimator; lines with circles, the actual standard deviations of the inner envelope estimator estimated by bootstrapping; lines with stars, the actual standard deviations of the standard model estimator; dashed lines, the asymptotic standard deviations of the inner envelope estimator; dot dash lines, the asymptotic standard deviations of the standard model estimator.

The next simulation tests the robustness of the inner envelope estimators with nonnormal errors. The scenario was the same as the simulation with $d = 7$ in Fig. 1, except that ε was generated as $\Sigma^{1/2}\epsilon$, where the elements of ϵ were independent and identically distributed standard normal, $t_6(3/2)^{1/2}$, $12^{1/2}\{U(0, 1) - 0 \cdot 5\}$, $(\chi^2_4 - 4)/\sqrt{8}$ random variables. Figure 2 confirms the consistency stated in Theorem 2, and also shows that a moderate departure from normality does not affect the estimator much even with small sample sizes.

4·2. *Data analysis*

In this section, we present three examples which demonstrate that inner envelopes can result in moderate to massive efficiency gains. We begin with the classical iris data which were analysed by Fisher (1936). Four characteristics, sepal length, sepal width, pedal length and pedal width, were measured for each specimen from three species of iris: *Iris setosa*, *Iris versicolor* and *Iris virginica*. Fisher analysed these data as an example for discriminant analysis; that is, to identify the species from the characteristics. We studied the relationship between the species and the characteristics from a different view, by fitting the data into the multivariate linear regression framework, taking the species as predictors and characteristics as responses to compare the species characteristics. Then each column of $\beta \in \mathbb{R}^{4\times 2}$ corresponds to the difference in the characteristic means for two species. The first column of β represents the group difference between *Iris setosa* and *Iris virginica*, while the second column represents the group difference between *Iris versicolor* and *Iris virginica*. We applied the envelope model to the data and $u = 4$ was inferred, which means that $\mathcal{E}_\Sigma(\mathcal{B}) = \mathbb{R}^r$ and that the envelope model offers no gains over the standard model. Then we fitted the inner envelope model to the data, and $d = 1$ was suggested by Bayes' criterion. Compared to the standard model, the standard deviations of the elements in $\hat{\beta}$ were reduced by 0·04% to 21·2%. Roughly speaking, to reduce a standard deviation by 21% in a standard analysis, the sample size should to be increased by about 61% and we expect that this gain would be worthwhile in many analyses.

The second dataset, on wine recognition, comprises 178 samples of wines made from three different cultivars in Italy taken for a chemical analysis. We used six variables for the characteristics of the wines: alcohol, malic acid, ash, alkalinity of ash, magnesium and flavanoids. The regression of characteristics on cultivars was performed to study how different cultivars influence

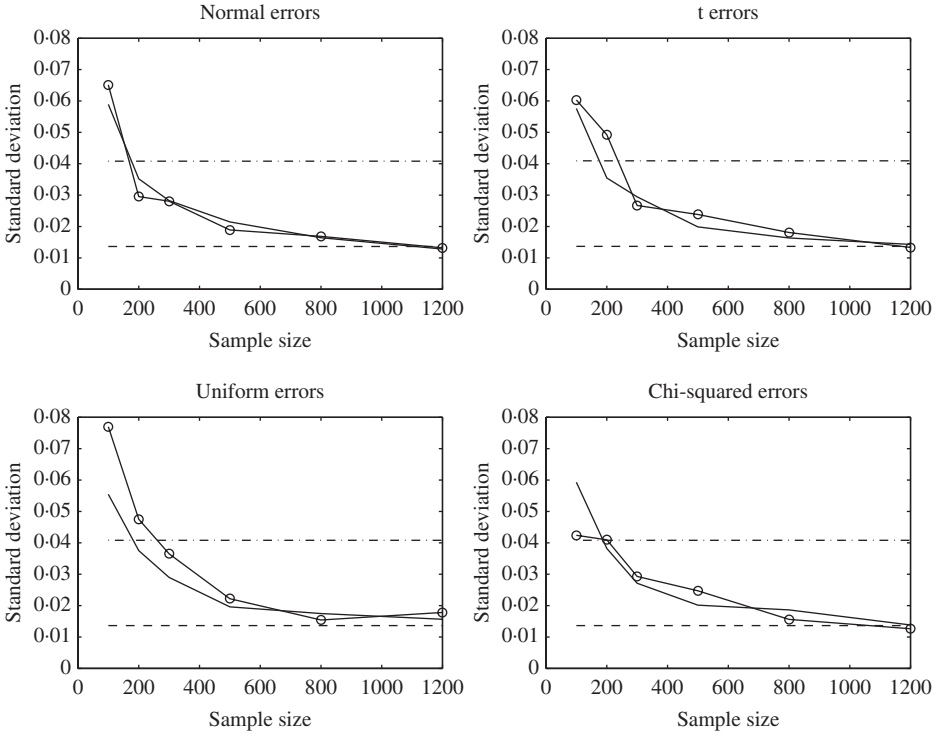

Fig. 2. Simulation results for different error types. The standard deviations are for an element inside the inner envelope. The line marks are the same as those in Fig. 1, but we omit the lines for the standard model estimator. The simulation scenario is the same as that for the $d = 7$ case in Fig. 1.

the aspects of the wines. Although not required by theory, for accuracy we did a log transformation to all the variables as the scatter-plot of the original variables showed an obvious departure from normality. The envelope model offered no gains as $u = 6$ was inferred. With $d = 1$, the inner envelope model reduced the standard deviations of elements in $\hat{\beta}$ by 0·2% to 30%. And to reduce the standard deviation by 30% in a standard analysis, we need to double the original sample size, which is an obvious gain.

The China climate data, obtained from the website of the National Center for Atmospheric Research, contain monthly measurements of average temperature and precipitation at 160 land stations in China from 1951 to 2000. Following Li et al. (2003), we took the monthly measurements of precipitation or temperature as responses, and the longitude, latitude of the stations and year as predictors, giving $r = 12$ and $p = 3$. First we applied the envelope model to the temperature data and $u = 12$ was inferred, indicating that the envelope model is equivalent to the standard model in this case. Then we fitted the data with the inner envelope model, and the Akaike criterion selected $d = 1$. The standard deviations of the elements in $\hat{\beta}$ were reduced by 0·02% to 65·9%. The sample size should be more than eight times the original sample size to gain a reduction of 65·9% in a standard analysis. With the precipitation data, results are similar and the standard deviations were reduced by 0·1% to 50·6%.

Acknowledgement

We are grateful to the editor and two referees for their insightful suggestions and comments that helped us improve the paper. The wine data were obtained from http://www.ailab.si/orange/datasets.psp. Research for this article was supported in part by a grant from the U.S. National Science Foundation.

APPENDIX

Maximum likelihood estimators

Given the data $Y_1, \ldots, Y_n$, the likelihood function L has the form

$$L = \{\det(\Sigma)\}^{-n/2} \times \text{etr}\{-2^{-1}(U - F\beta^{\mathrm{T}})\Sigma^{-1}(U - F\beta^{\mathrm{T}})^{\mathrm{T}}\},$$

where $\text{etr}(\cdot)$ denotes the composite function $\exp \circ \text{tr}(\cdot)$. As $\beta = \Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}}$, we follow Lemma 4.2 in Cook et al. (2010) and write $L = L_1L_2$, where

$$L_1 = \{\det(\Omega_1)\}^{-n/2} \times \text{etr}\{-2^{-1}(U\Gamma_1 - F\eta_1)\Omega_1^{-1}(U\Gamma_1 - F\eta_1)^{\mathrm{T}}\},$$
$$L_2 = \{\det(\Omega_0)\}^{-n/2} \times \text{etr}\{-2^{-1}(U\Gamma_0 - F\eta_2 B^{\mathrm{T}})\Omega_0^{-1}(U\Gamma_0 - F\eta_2 B^{\mathrm{T}})^{\mathrm{T}}\}.$$

Maximizing L_1 first while keeping Γ_1 fixed, the maximum likelihood estimator of η_1^{T} is $\hat{\eta}_1^{\mathrm{T}} = (U\Gamma_1)^{\mathrm{T}}F(F^{\mathrm{T}}F)^{-1}$. We use a hat on a parameter to denote both its intermediate estimator with unknown quantities and its final estimators. Substitute $\hat{\eta}_1^{\mathrm{T}}$ back to L_1, then

$$L_{11} = \{\det(\Omega_1)\}^{-n/2} \times \text{etr}(-2^{-1}Q_F U\Gamma_1\Omega_1^{-1}\Gamma_1^{\mathrm{T}}U^{\mathrm{T}}Q_F).$$

To maximize on Ω_1, we have $\hat{\Omega}_1 = \Gamma_1^{\mathrm{T}}\hat{\Sigma}_{\text{res}}\Gamma_1$. According to Lemma 4.3 in Cook et al. (2010), the maximum value of $\log(L_1)$ is a constant plus $-2^{-1}\log|\Gamma_1^{\mathrm{T}}\hat{\Sigma}_{\text{res}}\Gamma_1|$.

Fixing Γ_0, the maximization of $-L_2$ over $\text{span}(B)$, η_2 and Ω_0 follows Theorem 3.1 in Cook & Forzani (2008): $\hat{\Omega}_0 = \Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\text{res}}\Gamma_0 + (\Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\text{res}}\Gamma_0)^{1/2}\tilde{V}(\Gamma_0)\tilde{K}(\Gamma_0)\tilde{V}(\Gamma_0)^{\mathrm{T}}(\Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\text{res}}\Gamma_0)^{1/2}$, $\hat{B}$ is the orthogonal basis of $\hat{\mathcal{S}}_B = \hat{\Omega}_0\mathcal{S}_{p-d}(\hat{\Omega}_0, \Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\text{fit}}\Gamma_0)$, and $\hat{\eta}_2^{\mathrm{T}} = (\hat{B}^{\mathrm{T}}\hat{\Omega}_0^{-1}\hat{B})^{-1}\hat{B}^{\mathrm{T}}\hat{\Omega}_0^{-1}\Gamma_0^{\mathrm{T}}U^{\mathrm{T}}F(F^{\mathrm{T}}F)^{-1}$. The maximum value of L_2 is then a constant plus

$$-\frac{n}{2}\log|\Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\text{res}}\Gamma_0| - \frac{n}{2}\sum_{i=p-d+1}^{r-d}\log\{1 + \tilde{\lambda}_i(\Gamma_0)\} - \frac{n(r-d)}{2},$$

where $\hat{\Sigma}_{\text{res}} = U^{\mathrm{T}}Q_F U/n$ and $\hat{\Sigma}_{\text{fit}} = U^{\mathrm{T}}P_F U/n$. Using the fact that, for a nonsingular matrix A, $|\Gamma_0^{\mathrm{T}}A\Gamma_0| = |A||\Gamma^{\mathrm{T}}A^{-1}\Gamma|$, and adding the partially maximized $\log(L_1)$ and $\log(L_2)$, the objective function to maximize is

$$\log|G^{\mathrm{T}}\hat{\Sigma}_{\text{res}}G| + \log|G^{\mathrm{T}}\hat{\Sigma}_{\text{res}}^{-1}G| + \sum_{i=p-d+1}^{r-d}\log\{1 + \tilde{\lambda}_i(G_0)\},$$

where the optimization is taken over the Grassmann manifold $\mathbb{G}^{r\times d}$. A value of $\mathbb{G}$ that maximizes the function is the maximum likelihood estimator of Γ_1.

Proofs

Here the limits of all stochastic quantities refer to either convergence in probability or convergence in distribution. The type of convergence should be clear from context.

Proof of Proposition 1. Because $\mathcal{IE}_M(\mathcal{V})$ itself is a reducing subspace of M that is contained in $\mathcal{V}$, we have $\mathcal{IE}_M(\mathcal{V}) \subseteq \sum_i \mathcal{V}_i$.

Now we need only to show $\mathcal{IE}_M(\mathcal{V}) \supseteq \sum_i \mathcal{V}_i$. If there is an element v in $\sum_i \mathcal{V}_i$ but not in $\mathcal{IE}_M(\mathcal{V})$, then there exists a $\mathcal{V}_{i_0}$ so that $v \in \mathcal{V}_{i_0}$. Let $\mathcal{T} = \mathcal{V}_{i_0} + \mathcal{IE}_M(\mathcal{V})$. Then $\mathcal{T}$ is a reducing subspace in $\mathcal{V}$ that has a bigger dimension than $\mathcal{IE}_M(\mathcal{V})$, which is a contradiction since $\mathcal{IE}_M(\mathcal{V})$ has maximal dimension. □

Proof of Proposition 2. Let $\mathcal{R} = \mathcal{E}_M(\mathcal{S}^{\perp})$. For the equality to hold, we need to show that (a) $\mathcal{R}^{\perp}$ is a reducing subspace of M, (b) $\mathcal{R}^{\perp}$ is contained in $\mathcal{S}$, and (c) $\mathcal{R}^{\perp}$ is the space with maximum dimension that satisfies (a) and (b).

For (a), since $\mathcal{R}$ is a reducing subspace of M, we have $M\mathcal{R} \subseteq \mathcal{R}$ and $M\mathcal{R}^{\perp} \subseteq \mathcal{R}^{\perp}$; this indicates that $\mathcal{R}^{\perp}$ is also a reducing subspace of M. For (b), as $\mathcal{R} \supseteq \mathcal{S}^{\perp}$, we have $\mathcal{R}^{\perp} \subseteq \mathcal{S}$. For (c), if $\mathcal{R}^{\perp}$ does not have maximum dimension, then we can find $\mathcal{R}_0 \supset \mathcal{R}^{\perp}$, and $\mathcal{R}_0$ satisfies (a) and (b). Then $\mathcal{R}_0^{\perp}$ will be a reducing

subspace of M and $\mathcal{R}_0^\perp \supseteq \mathcal{S}^\perp$; also $\mathcal{R}_0^\perp$ has a smaller dimension than $\mathcal{R}$, which contradicts that $\mathcal{R}$ is the smallest reducing subspace of M that contains $\mathcal{S}^\perp$. So (c) is also satisfied. □

Proof of Proposition 3. Since the errors are independent, $\hat{\Sigma}_Y \to \Sigma_Y$. We have

$$\begin{aligned}\Sigma_Y &= \mathrm{var}(Y) = \mathrm{E}\{\mathrm{var}(Y \mid X)\} + \mathrm{var}\{\mathrm{E}(Y \mid X)\} \\ &= \Gamma_1\Omega_1\Gamma_1^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}} + (\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})\Sigma_X(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}},\end{aligned}$$

$\hat{\Sigma}_{\mathrm{fit}} = U^{\mathrm{T}}P_F U/n$, $U = F(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}} + e - 1_n\bar{\varepsilon}^{\mathrm{T}}$, where 1_n is an $n \times 1$ vector of 1s and the ith row of $e \in \mathbb{R}^{n\times r}$ is $\varepsilon_i^{\mathrm{T}}$, $i = 1, \ldots, n$. Then

$$\begin{aligned}\hat{\Sigma}_{\mathrm{fit}} &= U^{\mathrm{T}}P_F U/n \\ &= (\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})F^{\mathrm{T}}F(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}}/n + (e - 1_n\bar{\varepsilon}^{\mathrm{T}})^{\mathrm{T}}F(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}}/n \\ &\quad + (\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})F^{\mathrm{T}}(e - 1_n\bar{\varepsilon}^{\mathrm{T}})/n + (e - 1_n\bar{\varepsilon}^{\mathrm{T}})^{\mathrm{T}}P_F(e - 1_n\bar{\varepsilon}^{\mathrm{T}})/n.\end{aligned}$$

As $F^{\mathrm{T}}F/n \to \Sigma_X$, the first term converges in probability to $(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})\Sigma_X(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}}$. Since the errors are independent of the predictors, $e^{\mathrm{T}}F/n \to 0$. As $1_n^{\mathrm{T}}F = 0$, $(e - 1_n\bar{\varepsilon}^{\mathrm{T}})^{\mathrm{T}}F = 0$. By Slutsky's theorem,

$$(e - 1_n\bar{\varepsilon}^{\mathrm{T}})^{\mathrm{T}}P_F(e - 1_n\bar{\varepsilon}^{\mathrm{T}})/n = \{(e - 1_n\bar{\varepsilon}^{\mathrm{T}})F/n\}(F^{\mathrm{T}}F/n)^{\dagger}\{F^{\mathrm{T}}(e - 1_n\bar{\varepsilon}^{\mathrm{T}})/n\} \to 0.$$

So $\hat{\Sigma}_{\mathrm{fit}} \to \Sigma_{\mathrm{fit}} = (\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})\Sigma_X(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}}$. Since $\hat{\Sigma}_Y = \hat{\Sigma}_{\mathrm{fit}} + \hat{\Sigma}_{\mathrm{res}}$, $\hat{\Sigma}_{\mathrm{res}} \to \Sigma_{\mathrm{res}} = \Gamma_1\Omega_1\Gamma_1^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$. □

Proof of Proposition 4. As $\breve{\lambda}_i(G_0)$ is the ith eigenvalue of the positive semi-definite matrix $A(G_0) \equiv (G_0^{\mathrm{T}}\Sigma_{\mathrm{res}}G_0)^{-1/2}(G_0^{\mathrm{T}}\Sigma_{\mathrm{fit}}G_0)(G_0^{\mathrm{T}}\Sigma_{\mathrm{res}}G_0)^{-1/2}$, we have $\breve{\lambda}_i(G_0) \geqslant 0$. When $G_0 = \Gamma_0$,

$$\begin{aligned}A(\Gamma_0) &= (\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{res}}\Gamma_0)^{-1/2}\{\Gamma_0^{\mathrm{T}}(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})\Sigma_X(\Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}})^{\mathrm{T}}\Gamma_0\}(\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{res}}\Gamma_0)^{-1/2} \\ &= \Omega_0^{-1/2}B\eta_2^{\mathrm{T}}\Sigma_X\eta_2 B^{\mathrm{T}}\Omega_0^{-1/2}.\end{aligned}$$

Because B is an $(r-d) \times (p-d)$ matrix, the rank of $A(\Gamma_0)$ is at most $p-d$, then the $(p-d+1)$th to the $(r-d)$th eigenvalues are all 0, and $\breve{\lambda}_i(\Gamma_0) = 0$ for $i = p-d+1, \ldots, r-d$. So the term $\sum_{i=p-d+1}^{r-d}\log\{1 + \breve{\lambda}_i(G_0)\}$ is minimized at $G_0 = \Gamma_0$.

By Appendix A.6 of Cook (2007), for any semi-orthogonal matrix $G_1 \in \mathbb{R}^{r\times d}$, and its completion G_0, $\log|G_1^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}G_1| + \log|G_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}G_0| \geqslant \log|\Gamma_1^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\Gamma_1| + \log|\Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\Gamma_0|$ for all G. So we have $\Gamma_1 = \mathrm{argmin}_{G_1}\tilde{L}_{\mathrm{inner}}(G_1)$. □

Proof of Theorem 1. In the proof below, we use the notation R_A to represent a possible value for a parameter A. Objective functions for estimation of A are then to be optimized over R_A.

Since $\beta = \Gamma_1\eta_1^{\mathrm{T}} + \Gamma_0 B\eta_2^{\mathrm{T}}$ and $\Sigma = \Gamma_1\Omega_1\Gamma_1^{\mathrm{T}} + \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$, the Fisher consistency of β and Σ follows if the estimators of all parameters Γ_1, B, η_1, η_2, Ω_1 and Ω_0 are Fisher consistent. The Fisher consistency of Γ_1 is given in Proposition 4.

Estimators of η_1 and Ω_1 are Fisher consistent because the sample objective function to minimize is $\log|R_{\Omega_1}| + \mathrm{tr}\{(U\Gamma_1 - FR_{\eta_1})R_{\Omega_1}^{-1}(U\Gamma_1 - FR_{\eta_1})^{\mathrm{T}}/n\}$. It is known that $R_{\eta_1} = \eta_1$ and $R_{\Omega_1} = \Omega_1$ minimize the population version of this objective function, and therefore the estimators of η_1 and Ω_1 obtained in § 3 are Fisher consistent.

The estimator of Ω_0 is also Fisher consistent as the sample version of its objective function to minimize is $\log|R_{\Omega_0}| + \mathrm{tr}(R_{\Omega_0}^{-1}\Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{res}}\Gamma_0) + \sum_{i=p-d+1}^{r-d}\tilde{\lambda}_i$, where $\tilde{\lambda}_i$ is the ith eigenvalue of the matrix $R_{\Omega_0}\Gamma_0^{\mathrm{T}}\hat{\Sigma}_{\mathrm{fit}}\Gamma_0$. The population version of this objective function is $\log|R_{\Omega_0}| + \mathrm{tr}(R_{\Omega_0}^{-1}\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{res}}\Gamma_0) + \sum_{i=p-d+1}^{r-d}\breve{\lambda}_i$, where $\breve{\lambda}_i$ is the ith eigenvalue of the matrix $R_{\Omega_0}\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{fit}}\Gamma_0$. Since $R_{\Omega_0}\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{fit}}\Gamma_0 = R_{\Omega_0}B\eta_2^{\mathrm{T}}\Sigma_X\eta_2 B^{\mathrm{T}}$, η_2 is a $p \times (p-d)$ matrix and thus the rank of $R_{\Omega_0}\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{fit}}\Gamma_0$ is at most $p-d$, so we have $\sum_{i=p-d+1}^{r-d}\breve{\lambda}_i = 0$. As $\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{res}}\Gamma_0 = \Omega_0$, let $W = \Omega_0^{1/2}R_{\Omega_0}^{-1}\Omega_0^{1/2}$, it is then equivalent to maximize $\log|W| - \mathrm{tr}(W)$. Denoting the eigenvalues of W as $a_1, a_2, \ldots, a_{r-d}$, then we need to maximize $\log(a_1) + \cdots + \log(a_{r-d}) - (a_1 + \cdots + a_{r-d})$. The maximum can be obtained at $a_i = 1$, for $i = 1, \ldots, r-d$. As W is positive definite, by the spectral theorem W can be only I_{r-d}. So $R_{\Omega_0} = \Omega_0$ maximizes the objective function.

The estimator of B is Fisher consistent since the population version of the objective function to maximize is $\mathrm{tr}(P_{\Omega_0^{-1/2}R_B}\Omega_0^{-1/2}\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{fit}}\Gamma_0\Omega_0^{-1/2})$. This is maximized when $\mathrm{span}(\Omega_0^{-1/2}R_B)$ equals the span of the first $p-d$ eigenvectors of $\Omega_0^{-1/2}\Gamma_0^{\mathrm{T}}\Sigma_{\mathrm{fit}}\Gamma_0\Omega_0^{-1/2} = \Omega_0^{-1/2}B\eta_2^{\mathrm{T}}\Sigma_X\eta_2 B\Omega_0^{-1/2}$. Then the objective function is maximized at $R_B = B$.

The estimator of η_2 is Fisher consistent since the population version of the objective function to minimize is $g(R_{\eta_2}) = \mathrm{tr}\{\Sigma_2^{\dagger}\mathrm{var}(Y - \Gamma_0 B R_{\eta_2}^{\mathrm{T}}X)\}$, where $\Sigma_2 = \Gamma_0\Omega_0\Gamma_0^{\mathrm{T}}$. This is because the part of the likelihood function related to η_2 is $\mathrm{tr}\{(U - F\eta_2 B^{\mathrm{T}}\Gamma_0^{\mathrm{T}})\Sigma_2^{\dagger}(U - F\eta_2 B^{\mathrm{T}}\Gamma_0^{\mathrm{T}})\}^{\mathrm{T}}$, we can rewrite this part as $\mathrm{tr}\{\Sigma_2^{\dagger}\mathrm{var}(Y - \Gamma_0 B\eta_2^{\mathrm{T}}X)\}$, $g(\eta_2) - g(R_{\eta_2}) = \mathrm{tr}\{\Sigma_2^{\dagger}\mathrm{var}(Y - \Gamma_0 B\eta_2^{\mathrm{T}}X) - \Sigma_2^{\dagger}\mathrm{var}(Y - \Gamma_0 B R_{\eta_2}^{\mathrm{T}}X)\} = -\mathrm{tr}[\Sigma_2^{\dagger}\mathrm{var}\{\Gamma_0 B(\eta_2 - R_{\eta_2})^{\mathrm{T}}X\}] \leqslant 0$. Hence $g(R_{\eta_2})$ is minimized at η_2. □

Proof of Theorem 2. Since the inner envelope model is overparameterized, we will apply Proposition 4.1 of Shapiro (1986) to prove Theorem 2. To apply the proposition, we will check the assumptions first. Along the discussion, we will match Shapiro's notations in our context.

Shapiro's x in our context has the form $x^{\mathrm{T}} = \{\mathrm{vec}^{\mathrm{T}}(\hat{\beta}_{\mathrm{sm}}), \mathrm{vech}^{\mathrm{T}}(\hat{\Sigma}_{\mathrm{sm}})\}$, where $\hat{\beta}_{\mathrm{sm}}$ and $\hat{\Sigma}_{\mathrm{sm}}$ denote the estimators of β and Σ using the standard model. We need to show first that x is asymptotically normally distributed when the errors have finite fourth moments. Although the proof of asymptotic normality of $\hat{\beta}_{\mathrm{sm}}$ is known, we were unable to find in the literature a proof of the asymptotic normality of x.

Define $R = U - F\hat{\beta}_{\mathrm{sm}}^{T}$. Then $\hat{\Sigma}_{\mathrm{sm}} = n^{-1}(U - F\hat{\beta}_{\mathrm{sm}}^{T})^{\mathrm{T}}(U - F\hat{\beta}_{\mathrm{sm}}^{T}) = n^{-1}R^{\mathrm{T}}R$. Since $R = U - P_F U = Q_F U \equiv Q_F(F\beta^{\mathrm{T}} + e - 1_n\bar{\varepsilon}^{\mathrm{T}}) = Q_F e - 1_n\bar{\varepsilon}^{\mathrm{T}}$, we have $R^{\mathrm{T}}R = e^{\mathrm{T}}e - n\bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}} - e^{\mathrm{T}}P_F e$. Let e_{ij} and σ_{ij} denote the elements of $e \in \mathbb{R}^{n\times r}$ and $\Sigma \in \mathbb{R}^{r\times r}$, and let $e_{k\cdot}$ and $e_{\cdot k}$ denote the kth row and column of e. Then, noting that $\sqrt{n}\bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}}$ converges to zero in probability,

$$\begin{aligned}\sqrt{n}\{\mathrm{vech}(\hat{\Sigma}_{\mathrm{sm}}) - \mathrm{vech}(\Sigma)\} &= \sqrt{n}\mathrm{vech}\left(n^{-1}e^{\mathrm{T}}e - \bar{\varepsilon}\bar{\varepsilon}^{\mathrm{T}} - \Sigma - n^{-1}e^{\mathrm{T}}P_F e\right)\\ &\equiv n^{-1/2}\sum_{i=1}^{n} B_i - n^{-1/2}\mathrm{vech}(e^{\mathrm{T}}P_F e) + o_p(1),\end{aligned}$$

where $\sum_{i=1}^{n} B_i = \mathrm{vech}(e_{i\cdot}^{\mathrm{T}}e_{i\cdot} - \Sigma)$ and the B_is are independent and identically distributed random vectors with mean zero, length $r(r+1)/2$ and typical element $e_{ij}e_{ik} - \sigma_{jk}$ $(k = 1, \ldots, r;\ j = k, \ldots, r)$. The (i, j)th element in $e^{\mathrm{T}}P_F e/\sqrt{n}$ is $e_{\cdot i}^{\mathrm{T}}P_F e_{\cdot j}/\sqrt{n}$ and we denote the elements of $P_F \in \mathbb{R}^{n\times n}$ by p_{ij}. Let $B_n = n^{-1/2}\mathrm{E}(e_{\cdot i}^{\mathrm{T}}P_F e_{\cdot j})$. Then

$$B_n = n^{-1/2}\sum_{a=1}^{n}\sum_{b=1}^{n}\mathrm{E}(e_{ai}p_{ab}e_{bj}) = n^{-1/2}\sum_{a=1}^{n}p_{aa}\mathrm{E}(e_{ai}e_{aj}) = \frac{p}{\sqrt{n}}\mathrm{E}(e_{1i}e_{1j}) \to 0.$$

The penultimate equality holds since the observations from different samples are independent. For each (i, j) the $e_{ai}e_{aj}$s are independent and identically distributed random variables with common finite variance. The final equality holds since P_F is a rank p projection matrix, $\sum_{a=1}^{n}p_{aa} = \mathrm{tr}(P_F) = p$.

Let $A_n = \mathrm{var}\left(n^{-1/2}e_{\cdot i}^{\mathrm{T}}P_F e_{\cdot j}\right) = n^{-1}\mathrm{var}\left(\sum_{a=1}^{n}\sum_{b=1}^{n}e_{ai}p_{ab}e_{bj}\right)$. Then

$$\begin{aligned}A_n &= n^{-1}\sum_{a=1}^{n}\sum_{b=1}^{n}p_{ab}^2\mathrm{var}(e_{ai})\mathrm{var}(e_{bj}) + n^{-1}\sum_{a=1}^{n}\sum_{b=1}^{n}p_{ab}^2\mathrm{cov}(e_{ai}, e_{aj})\mathrm{cov}(e_{bi}, e_{bj})\\ &\quad + n^{-1}\sum_{a=1}^{n}p_{aa}^2\{\mathrm{var}(e_{ai}e_{aj}) - \mathrm{var}(e_{ai})\mathrm{var}(e_{aj}) - \mathrm{cov}(e_{ai}, e_{aj})^2\}\\ &= n^{-1}\{\mathrm{var}(e_{1i})\mathrm{var}(e_{1j}) + \mathrm{cov}(e_{1i}, e_{1j})\mathrm{cov}(e_{1i}, e_{1j})\}\sum_{a=1}^{n}\sum_{b=1}^{n}p_{ab}^2\\ &\quad + n^{-1}\{\mathrm{var}(e_{1i}e_{1j}) - \mathrm{var}(e_{1i})\mathrm{var}(e_{1j}) - \mathrm{cov}(e_{1i}, e_{1j})^2\}\sum_{a=1}^{n}p_{aa}^2.\end{aligned}$$

As P_F is a projection matrix, $\sum_{a=1}^n \sum_{b=1}^n p_{ab}^2 = p$, and the first summand will go to zero as $n \to 0$. Because all p_{aa}s are nonnegative, $\sum_{a=1}^n p_{aa}^2 \leqslant \left(\sum_{a=1}^n p_{aa}\right)^2 = p^2$, the second summand will also go to zero as $n \to 0$. So $n^{-1/2}\mathrm{var}\left(e_{\cdot i}^{\mathrm{T}} P_F e_{\cdot j}\right) \to 0$. Then by the Chebyshev inequality, $n^{-1/2} e_{\cdot i}^{\mathrm{T}} P_F e_{\cdot j}$ converges in probability to zero. Since i and j are arbitrary, each element in $e^{\mathrm{T}} P_F e$ converges to zero in probability. Consequently, $\sqrt{n}\{\mathrm{vech}(\hat{\Sigma}_{\mathrm{sm}}) - \mathrm{vech}(\Sigma)\} = n^{-1/2} \sum_{i=1}^n B_i + o_p(1)$.

Now apply Theorem B on page 30 of Serfling (1980) to prove the asymptotic normality for $\{\mathrm{vec}(\hat{\beta}_{fm})^{\mathrm{T}}, \mathrm{vech}(\hat{\Sigma}_{fm})\}^{\mathrm{T}}$. For simplicity, we study the asymptotic normality for $\{\mathrm{vec}(\hat{\beta}_{fm})^{\mathrm{T}}, \mathrm{vec}(\hat{\Sigma}_{fm})\}^{\mathrm{T}} \equiv \sum_{i=1}^n V_i/n$ instead. Let f_i^{T} denote the ith row of $F(F^{\mathrm{T}}F/n)^{-1}$ $(i = 1, \ldots, n)$. Then $\mathrm{vec}(\hat{\beta}_{fm} - \beta) = n^{-1} \sum_{i=1}^n f_i \otimes \varepsilon_i \equiv n^{-1} \sum_{i=1}^n A_i$. And V_i can be written as

$$V_i = \begin{pmatrix} A_i \\ B_i^* \end{pmatrix} = \begin{pmatrix} f_i \otimes \varepsilon_i \\ \varepsilon_i \otimes \varepsilon_i - \mathrm{vec}(\Sigma) \end{pmatrix}.$$

Let v_i^{T} be a $1 \times n$ row vector whose ith element is 1 and 0 otherwise, then $f_i^{\mathrm{T}} = v_i^{\mathrm{T}} F(F^{\mathrm{T}}F/n)^{-1}$, and we have $\|V_i\|^2 = \|A_i\|^2 + \|B_i^*\|^2 = \|\varepsilon_i\|^2 \|f_i\|^2 + \|B_i^*\|^2 = \|\varepsilon_i\|^2 v_i^{\mathrm{T}} F(F^{\mathrm{T}}F/n)^{-2} F^{\mathrm{T}} v_i + \|B_i^*\|^2$.

Let $a_i = \|f_i\|^2/n = v_i^{\mathrm{T}} F(F^{\mathrm{T}}F/n)^{-2} F^{\mathrm{T}} v_i/n$, and notice that

$$\begin{aligned} \sum_{i=1}^n a_i &= \sum_{i=1}^n \mathrm{tr}\{(F^{\mathrm{T}}F/n)^{-1} F^{\mathrm{T}} v_i v_i^{\mathrm{T}} F (F^{\mathrm{T}}F)^{-1}\} \\ &= \mathrm{tr}\{(F^{\mathrm{T}}F/n)^{-1}\} \to \mathrm{tr}(\Sigma_X^{-1}), \quad n \to \infty. \end{aligned}$$

Then for any $\epsilon > 0$,

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \mathrm{E}\{\|V_i\|^2 I(\|V_i\| > \epsilon\sqrt{n})\} \\ &\quad = \sum_{i=1}^n \mathrm{E}[(\|\varepsilon_i\|^2 a_i + \|B_i^*\|^2/n) I\{(\|\varepsilon_i\|^2 a_i + \|B_i^*\|^2/n) \geqslant \epsilon^2\}] \\ &\quad \leqslant \mathrm{E}[\{\|\varepsilon_i\|^2 \,\mathrm{tr}(F^{\mathrm{T}}F/n)^{-1} + \|B_i^*\|\} I\{(\|\varepsilon_i\|^2 \max_{i \leqslant n} a_i + \|B_i^*\|^2/n) \geqslant \epsilon^2\}] \to 0. \end{aligned}$$

The convergence in the last step is because the error has finite fourth moments, $\mathrm{E}\{\|\varepsilon_i\|^2 \,\mathrm{tr}(F^{\mathrm{T}}F/n)^{-1} + \|B_i^*\|\} < \infty$, and $\max_{i \leqslant n} p_{ii} \to 0$ leads to $\max_{i \leqslant n} a_i \to 0$, making the indicator function converge to zero. The justification for $\max_{i \leqslant n} a_i \to 0$ is as follows. For any matrix $A \in \mathbb{R}^{q \times p}$, symmetric matrix $B \in \mathbb{R}^{q \times q}$, the maximum and minimum eigenvalue of B, $\lambda_{\max}(B)$, $\lambda_{\min}(B)$, we have $A^{\mathrm{T}} A \lambda_{\min}(B) \leqslant A^{\mathrm{T}} B A \leqslant A^{\mathrm{T}} A \lambda_{\max}(B)$. In our case

$$\begin{aligned} \max_{i \leqslant n} a_i &= \max_{i \leqslant n} v_i^{\mathrm{T}} F (F^{\mathrm{T}}F/n)^{-2} F^{\mathrm{T}} v_i/n \\ &= \max_{i \leqslant n} v_i^{\mathrm{T}} F (F^{\mathrm{T}}F/n)^{-1/2} (F^{\mathrm{T}}F/n)^{-1} (F^{\mathrm{T}}F/n)^{-1/2} F^{\mathrm{T}} v_i \\ &\leqslant \max_{i \leqslant n} p_{ii} \lambda_{\max}\{(F^{\mathrm{T}}F/n)^{-1}\} \to 0. \end{aligned}$$

Now the covariance matrix for V_i has the form

$$W_i = \begin{pmatrix} f_i f_i^{\mathrm{T}} \otimes \Sigma & \mathrm{E}(f_i \varepsilon_i^{\mathrm{T}} \otimes \varepsilon_i \varepsilon_i^{\mathrm{T}}) \\ \mathrm{E}(f_i \varepsilon_i^{\mathrm{T}} \otimes \varepsilon_i \varepsilon_i^{\mathrm{T}})^{\mathrm{T}} & \mathrm{E}\{\varepsilon_i \otimes \varepsilon_i - \mathrm{vec}(\Sigma)\}^2 \end{pmatrix}.$$

Since $\sum_{i=1}^n f_i/n = \sum_{i=1}^n v_i^{\mathrm{T}} F(F^{\mathrm{T}}F/n)^{-1}/n = 1_n^{\mathrm{T}} F(F^{\mathrm{T}}F)^{-1}/n = 0$, and

$$\begin{aligned} \sum_{i=1}^n f_i f_i^{\mathrm{T}}/n &= \sum_{i=1}^n (F^{\mathrm{T}}F/n)^{-1} F^{\mathrm{T}} v_i v_i^{\mathrm{T}} F (F^{\mathrm{T}}F/n)^{-1}/n \\ &= (F^{\mathrm{T}}F/n)^{-1} \to \Sigma_X^{-1}, \quad n \to \infty, \end{aligned}$$

$(W_1 + \cdots + W_n)/n$ will converge to

$$\begin{pmatrix} \Sigma_X^{-1} \otimes \Sigma & 0 \\ 0 & \mathrm{E}\{\varepsilon_i \otimes \varepsilon_i - \mathrm{vec}(\Sigma)\}^2 \end{pmatrix}.$$

By Serfling (1980), $\sum_{i=1}^n V_i/n$ is asymptotically normal. As B_i is part of B_i^*, we have

$$\frac{1}{\sqrt{n}} \begin{pmatrix} A_i \\ B_i \end{pmatrix} \to N_{rp+r(r+1)/2}(0, C_2),$$

where C_2 is some constant matrix. In other words, $\sqrt{n}(x - \xi) \to N_{rp+r(r+1)/2}(0, C_2)$.

Shapiro's ξ in our context is $\xi^{\mathrm{T}} = \{\mathrm{vec}^{\mathrm{T}}(\beta), \mathrm{vech}^{\mathrm{T}}(\Sigma)\}$. Now we give the minimum discrepancy function f_{MDF}. Given $Y_1, Y_2, \ldots, Y_n$, the likelihood function is $L = \{\det(\Sigma)\}^{-n/2} \times \mathrm{etr}\{-2^{-1}(U - F\beta^{\mathrm{T}})\Sigma^{-1}(U - F\beta^{\mathrm{T}})^{\mathrm{T}}\}$. We have

$$\begin{aligned} &\mathrm{tr}\{(U - F\beta^{\mathrm{T}})\Sigma^{-1}(U - F\beta^{\mathrm{T}})^{\mathrm{T}}\} \\ &= \mathrm{tr}\{(U - F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} + F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})\Sigma^{-1}(U - F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} + F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})^{\mathrm{T}}\} \\ &= \mathrm{tr}\{\Sigma^{-1}(U - F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} + F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})^{\mathrm{T}}(U - F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} + F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})\} \\ &= \mathrm{tr}[\Sigma^{-1}\{n\hat{\Sigma}_{\mathrm{sm}} + (F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})^{\mathrm{T}}(F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})\}]. \end{aligned}$$

The last equality follows because $Q_F F = 0$ and that makes the cross product terms 0. Now the likelihood function is $L = \{\det(\Sigma)\}^{-n/2} \times \mathrm{etr}[-2^{-1}\Sigma^{-1}\{n\hat{\Sigma}_{\mathrm{sm}} + (F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})^{\mathrm{T}}(F\hat{\beta}_{\mathrm{sm}}^{\mathrm{T}} - F\beta^{\mathrm{T}})\}]$. The maximum value of L, denoted as $L_{\max}$, is reached when $x = \xi$, and $L_{\max} = \{\det(\hat{\Sigma}_{\mathrm{sm}})\}^{-n/2} \times \exp(-2^{-1}nr)$. Then f_{MDF} is formed as $f_{\mathrm{MDF}} = L_{\max} - L$.

Although f_{MDF} is written in terms of β, Σ, β_{sm} and Σ_{sm}, there must be one-to-one functions f_1 from the product space of β and Σ to ξ, f_2 from the product space of $\hat{\beta}_{\mathrm{sm}}$ and $\hat{\Sigma}_{\mathrm{sm}}$ to x so that $\xi = f_1(\beta, \Sigma)$ and $x = f_2(\hat{\beta}_{\mathrm{sm}}, \hat{\Sigma}_{\mathrm{sm}})$. As f_{MDF} is constructed under normal likelihood function, it satisfies the four conditions for f_{MDF} in § 3 in Shapiro (1986). Also, the function g defined by Shapiro in (2.1) is twice continuous differentiable, where ξ is defined before and θ are parameters in the inner envelope model $\{\mathrm{vec}(\eta_1)^{\mathrm{T}}, \mathrm{vec}(\eta_2)^{\mathrm{T}}, \mathrm{vec}(B)^{\mathrm{T}}, \mathrm{vec}(\Gamma_1)^{\mathrm{T}}, \mathrm{vech}(\Omega_1)^{\mathrm{T}}, \mathrm{vech}(\Omega_0)^{\mathrm{T}}\}^{\mathrm{T}}$. Therefore all the assumptions of Shapiro's Proposition 4.1 are satisfied, and $\hat{\beta}$ and $\hat{\Sigma}$, obtained by minimizing f_{MDF}, are consistent estimators of β and Σ, with the rate of $\sqrt{n}$. □

REFERENCES

COOK, R. D. (2007). Fisher lecture: dimension reduction in regression. *Statist. Sci.* **22**, 1–26.

COOK, R. D. & FORZANI, L. (2008). Principal fitted components for dimension reduction in regression. *Statist. Sci.* **23**, 485–501.

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

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* **7**, 179–88.

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.

LI, K. C., ARAGON, Y., SHEDDEN, K. & THOMAS-AGNAN, C. (2003). Dimension reduction for multivariate response data. *J. Am. Statist. Assoc.* **98**, 99–109.

SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

SHAO, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statist. Sinica* **7**, 221–64.

SHAPIRO, A. (1986). Asymptotic theory of overparameterized structural models. *J. Am. Statist. Assoc.* **81**, 142–9.

SU, Z. & COOK, R. D. (2011). Partial envelopes for efficient estimation in multivariate linear regression. *Biometrika* **98**, 133–46.

[*Received March* 2011. *Revised March* 2012]