Scaled Partial Envelope Model
in Multivariate Linear Regression

Jing Zhang[a,b], Zhensheng Huang[a] and Lixing Zhu[c,d]

[a] *School of Science, Nanjing University of Science and Technology*

[b] *School of Finance, Chuzhou University*

[c] *Center for Statistics and Data Science, Beijing Normal University at Zhuhai*

[d] *Department of Mathematics, Hong Kong Baptist University*

Supplementary Material

The Supplementary Material includes technical proofs of Propositions 1–3 in the main manuscript.

The following is proofs of the main results. The notations and definitions will be employed in our exposition. Let $\mathbb{R}^{m\times n}$ be the set of all real $m \times n$ matrices. The Grassmannian, which consists of the set of all u-dimensional subspaces of $\mathbb{R}^r(u \leq r)$, is denoted by $\mathcal{G}_{r,u}$. If $M \in \mathbb{R}^{m\times n}$, then $\text{span}(M) \subseteq \mathbb{R}^m$ is the subspace spanned by the columns of M. With $A \in \mathbb{R}^{a\times a}$ and a subspace $\mathcal{S} \subseteq \mathbb{R}^a, A\mathcal{S} = \{As : s \in \mathcal{S}\}$. If $\sqrt{n}(\widehat{\theta} - \theta)$ converges to a normal random vector with mean 0 and covariance matrix

V, we write its asymptotic covariance matrix as $\text{avar}(\sqrt{n}\widehat{\theta}) = V$. We use $P_{A(V)} = A(A^T V A)^{-1} A^T V$ to denote projection onto span(A) with the V inner product and use P_A to denote projection onto span(A) with the identity inner product. Let $Q_{A(V)} = I - P_{A(V)}$. We will use operators vec: $\mathbb{R}^{a\times b} \to \mathbb{R}^{ab}$, which vectorizes an arbitrary matrix by stacking its columns, and vech: $\mathbb{R}^{a\times a} \to \mathbb{R}^{a(a+1)/2}$, which vectorizes a symmetric matrix by extracting its columns of elements below or on the diagonal. The symbol bdiag$(\cdot)$ denotes a block diagonal matrix with the diagonal blocks as arguments. Let $A \otimes B$ denote the Kronecker product of matrices A and B, and let $A^\dagger$ denote the Moore–Penrose inverse of A. We employ $\widehat{\theta}_\xi$ to denote an estimator of θ with known true parameter value of ξ.

Maximum likelihood estimators

The maximum likelihood estimator of α is $\overline{Y}$. In that way, with the dimension of the $\Lambda^{-1}\Sigma\Lambda^{-1}$-partial envelope of $\Lambda^{-1}\mathcal{B}_1$, which is fixed at u_1, the loglikelihood function L_1 is

$$L_1 = -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\big\{(U - F_1\beta_1^T)\Sigma^{-1}(U - F_1\beta_1^T)^T\big\}, \tag{S0.1}$$

$$= -\frac{nr}{2}\log(2\pi) - \frac{n}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\left[\Sigma^{-1}\big\{n\widetilde{\Sigma}_{\mathrm{res}} + (\widetilde{\beta}_1 - \beta_1)F_1^T F_1(\widetilde{\beta}_1^T - \beta_1^T)\big\}\right], \tag{S0.2}$$

$$= -\frac{nr}{2}\log(2\pi) - n\log|\Lambda| - \frac{n}{2}\log|\Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T| - \frac{1}{2}\mathrm{tr}\big\{(U\Lambda^{-1} - F_1\eta^T\Gamma^T)(\Gamma\Omega\Gamma^T + \Gamma_0\Omega_0\Gamma_0^T)^{-1}(U\Lambda^{-1} - F_1\eta^T\Gamma^T)^T\big\}. \tag{S0.3}$$

There are three forms of the likelihood function: (S0.1), (S0.2), and (S0.3). (S0.1) is a common form, which has the observed data and parameters β_1 and Σ. (S0.2) substitutes sufficient statistics $\widetilde{\beta}_1$ and $\widetilde{\Sigma}_{\mathrm{res}}$ for the observed data in (S0.1). (S0.3) rewrites (S0.1) in light of the constituent parameters. (S0.3) has the identical form with the loglikelihood function from the partial envelope model, except we have the extra term $-n\log|\Lambda|$ and the response is $\Lambda^{-1}Y$. In addition, we maximize over all constituent parameters apart from Λ and Γ, and obtain the partially maximized form

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

Proof of Proposition 1.

Proposition 3.1 in Shapiro (1986) is employed to prove this proposition, and we will match our notations with Shapiro's proving process. We augment a subscript "s" in Shapiro's notation to distinguish better. The θ_s of Shapiro's setting is our $\phi = \left\{\lambda^T, \mathrm{vec}(\eta)^T, \mathrm{vec}(\Gamma)^T, \mathrm{vech}(\Omega)^T, \mathrm{vech}(\Omega_0)^T\right\}^T$. Shapiro's $\widehat{x}_s$ is equivalent to our $\left\{\mathrm{vec}(\widetilde{\beta}_1)^T, \mathrm{vech}(\widetilde{\Sigma}_{\mathrm{res}})^T\right\}^T$, and Shapiro's ξ_s is $\left\{\mathrm{vec}(\beta_1)^T, \mathrm{vech}(\Sigma)^T\right\}^T$ in our setting. The discrepancy function F_s is our loglikelihood function, except we delete a constant factor n.

$$
\begin{aligned}
F_s =& L_1/n, \\
=& -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}\mathrm{tr}\big\{(U - F_1\beta_1^T)\Sigma^{-1}(U - F_1\beta_1^T)^T/n\big\}, \\
=& -\frac{r}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| \\
& - \frac{1}{2}\mathrm{tr}\Big[\Sigma^{-1}\big\{n\widetilde{\Sigma}_{\mathrm{res}} + (\widetilde{\beta}_1 - \beta_1)(F_1^T F_1/n)(\widetilde{\beta}_1^T - \beta_1^T)\big\}\Big].
\end{aligned}
$$

When we build F_s under a normal likelihood function, it meets the conditions 1–4 in ξ 3 of Shapiro (1986). Shapiro's Δ_s is the gradient matrix $\partial\xi_s/\partial\theta_s$, and it is identical to H in our setting. Let $e = U - F_1\beta_1^T$, $V_s = \mathrm{bdiag}\big\{(F_1^T F_1/n) \otimes \Sigma^{-1}, E_r^T(\Sigma^{-1} \otimes \Sigma^{-1})E_r/2\big\}$ of Shapiro's context is 1/2 times the Hessian matrix $\partial^2 F_s/\partial\xi_s\partial\xi_s^T$, which is evaluated at (ξ_s, ξ_s). When we suppose $\sum_{i=1}^{n}(R_{1|2})_i(R_{1|2})_i^T/n > 0$, V_s is full rank and $\mathrm{rank}(\Delta_s^T V_s \Delta_s) = \mathrm{rank}(\Delta_s)$. Hence, all conditions in Proposition 1 are satisfied, and the maximizers $\widehat{\beta}_1$ and $\widehat{\Sigma}$ are uniquely defined.

□

Proof of Proposition 2.

We begin the proof from the asymptotic covariance matrix $\Delta_s(\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T V_s \Gamma_s V_s \Delta_s(\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T$, which is provided at the end of Proposition 3. Shaprio's $\Gamma_s = V_s^{-1}$ is under the additional assumption of normality. Hence, the asymptotic covariance matrix has the form $\Delta_s(\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T$, which is

$V = H(H^TJH)^\dagger H^T$. We employ merely our notation, which contains simplifying V.

We compute straightly

$$H = \partial\big\{\text{vec}(\beta_1)^T, \text{vech}(\Sigma)^T\big\}^T/\partial\phi^T,$$
$$= \big\{D_\Lambda h_o(I_{p_1} \otimes \Lambda^{-1})L, D_\Lambda G_o\big\},$$
$$= (H_1, H_2),$$

where H_1 and H_2 are defined to simplify subsequent formulas. Because V is unchanging under full rank linear conversions of the columns of H, below we convert the columns of H by the nonsingular matrix

$$T = \begin{pmatrix} I_{r-1} & 0 \\ -(H_2^TJH_2)^\dagger H_2^TJH_1 & I_{r(r+1)/2} \end{pmatrix}.$$

Then $HT = \big(Q_{H_2(J)}H_1, H_2\big)$ and $T^TH^TJHT = \text{bdiag}\big(H_1^TQ_{H_2(J)}^TJQ_{H_2(J)}H_1, G_o^TJ_oG_o\big)$. Therefore, we have

$$V = HT(T^TH^TJHT)^\dagger T^TH^T,$$
$$= J^{-1/2}PJ^{-1/2} + D_\Lambda G_o(G_o^TJ_oG_o)^\dagger G_o^TD_\Lambda^T,$$

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 equation is identical to V_2. The first term can be represented as V_1 by employing the identities

$$\begin{aligned} Q_{H_2(J)}H_1 &= D_\Lambda Q_{G_o(J_o)} D_\Lambda^{-1} H_1, \\ &= D_\Lambda Q_{G_o(J_o)} h_o L \Lambda_1^{-1}, \\ &= D_\Lambda A_o \Lambda_1^{-1}, \end{aligned}$$

where $\Lambda_1 = \mathrm{diag}(\lambda_2, \ldots, \lambda_r)$.

□

Proof of Proposition 3.

Proposition 2 is a particular case of Proposition 3. When Γ is overly parameterized, we use Proposition 4.1 in Shapiro (1986) to build the proof. The conditions for Proposition 4.1 are identical to Proposition 3.1 in Shapiro, except with an additional assumption that $n^{1/2}(\widehat{x}_s - \xi_s)$ is asymptotically normal. When we discussed the proof of our Proposition 1, we have demonstrated that all the conditions in Shapiro's Proposition 3.1 are satisfied. Then, the condition on $p_{1,ii}$ ensures that the asymptotic distribution of $n^{1/2}\left[\left\{\mathrm{vec}(\widetilde{\beta}_1)^T, \mathrm{vech}(\widetilde{\Sigma}_{\mathrm{res}})^T\right\}^T - \left\{\mathrm{vec}(\beta_1)^T, \mathrm{vech}(\Sigma)^T\right\}^T\right]$ is multivariate normal, so the additional assumption is also satisfied. Hence, from Proposition 4.1 of Shapiro (1986) and employing Shapiro's notation, the asymptotic variance has the form $\Delta_s(\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T V_s \Gamma_s V_s \Delta_s (\Delta_s^T V_s \Delta_s)^\dagger \Delta_s^T$, where Shapiro's Γ_s is the asymptotic variance of $\left\{\mathrm{vec}(\widetilde{\beta}_1)^T, \mathrm{vech}(\widetilde{\Sigma}_{\mathrm{res}})^T\right\}^T$.

□

Bibliography

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association 81*, pp. 142–149.