Supplement: Proofs and Technical Details for “Foundations for Envelope Models and Methods”

R. Dennis Cook and Xin Zhang

A Regression applications

A.1 Envelopes for weighted least squares

Consider a heteroscedastic linear model with data consisting of n independent copies of $(Y, \mathbf{X}, W)$, where $Y \in \mathbb{R}^1$, $\mathbf{X} \in \mathbb{R}^p$, and $W > 0$ is a weight with $E(W) = 1$:

$$Y = \mu + \boldsymbol{\beta}^T \mathbf{X} + \varepsilon / \sqrt{W}, \quad (\text{A.1})$$

where $\varepsilon \perp (\mathbf{X}, W)$ and $\text{var}(\varepsilon) = \sigma^2$. The constraint $E(W) = 1$ is without loss of generality and serves to normalize the weights so they have mean 1 and to make subsequent expressions a bit simpler. Fitting under this model is typically done by using weighted least squares (WLS):

$$(\hat{\mu}, \hat{\boldsymbol{\beta}}) = \arg \min n^{-1} \sum_{i=1}^n W_i (Y_i - \mu - \boldsymbol{\beta}^T \mathbf{X}_i)^2, \quad (\text{A.2})$$

where we normalize the sample weights so that $\bar{W} = 1$. Let $\boldsymbol{\Sigma}_{\mathbf{X}(W)} = E\{W(\mathbf{X} - E(W\mathbf{X}))(\mathbf{X} - E(W\mathbf{X}))^T\}$ be the weighted covariance matrix of $\mathbf{X}$, and let $\boldsymbol{\Sigma}_{\mathbf{X}Y(W)} = E[W\{\mathbf{X} - E(W\mathbf{X})\}\{Y - E(WY)\}]$ be the weighted covariance between $\mathbf{X}$ and Y . Then $\boldsymbol{\beta}_t = \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1} \boldsymbol{\Sigma}_{\mathbf{X}Y(W)}$ and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_t)$ converges to a normal random vector with mean 0 and variance $\mathbf{V}_{\boldsymbol{\beta}\boldsymbol{\beta}} = \sigma^2 \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1}$. According to Definition 3, we should now strive to estimate the $\sigma^2 \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1}$ -envelope of $\text{span}(\boldsymbol{\beta}_t)$, which can be done for a specified dimension u by using the 1D algorithm with $\widehat{\mathbf{M}} = s^2 \mathbf{S}_{\mathbf{X}(W)}^{-1}$ and $\widehat{\mathbf{U}} = \hat{\boldsymbol{\beta}} \hat{\boldsymbol{\beta}}^T$, where s^2 is the estimator of σ^2 from the WLS fit (A.2), and $\mathbf{S}_{\mathbf{X}(W)}$ is the sample version of $\boldsymbol{\Sigma}_{\mathbf{X}(W)}$. Once an estimated basis $\widehat{\boldsymbol{\Gamma}}$ for $\mathcal{E}_{\sigma^2 \boldsymbol{\Sigma}_{\mathbf{X}(W)}^{-1}}(\boldsymbol{\beta}_t)$ is obtained, we re-fit the WLS regression model, replacing $\mathbf{X}$ with $\widehat{\boldsymbol{\Gamma}}^T \mathbf{X}$, to estimate the coordinate vector $\boldsymbol{\eta}$,

$$(\hat{\mu}, \hat{\boldsymbol{\eta}}) = \arg \min n^{-1} \sum_{i=1}^n W_i (Y_i - \mu - \boldsymbol{\eta}^T \widehat{\boldsymbol{\Gamma}}^T \mathbf{X}_i)^2. \quad (\text{A.3})$$

The resulting envelope estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}}_{\text{env}} = \widehat{\boldsymbol{\Gamma}} \hat{\boldsymbol{\eta}} = \widehat{\boldsymbol{\Gamma}} (\widehat{\boldsymbol{\Gamma}}^T \mathbf{S}_{\mathbf{X}(W)} \widehat{\boldsymbol{\Gamma}})^{-1} \widehat{\boldsymbol{\Gamma}}^T \mathbf{S}_{\mathbf{X}Y(W)} \hat{\boldsymbol{\beta}} = \mathbf{P}_{\widehat{\boldsymbol{\Gamma}}(\mathbf{S}_{\mathbf{X}(W)})} \hat{\boldsymbol{\beta}}. \quad (\text{A.4})$$

Cross validation or a hold out sample could now be used straightforwardly as an aid to selecting u . When the weights are constant, this method has the same structure as partial least squares regression with the important exception that partial least squares uses the SIMPLS algorithm in place of the 1D algorithm that we recommend.

To take a likelihood approach with the envelope $\mathcal{E}_{\sigma^2 \Sigma_{\mathbf{X}(W)}^{-1}}(\beta_t)$, we follow Section 3.1 to decompose the full log-likelihood into the sum of the conditional log-likelihood from $Y|(\mathbf{X}, W)$, denoted by $C_n(\mu, \beta, \sigma^2) := C_n(\mu, \Gamma, \eta, \sigma^2)$, and the conditional log-likelihood from $\mathbf{X}|W$, denoted by $M_n(\psi_1, \Gamma)$, where ψ_1 denotes additional nuisance parameters. We condition on the observed values of W , assuming that it is ancillary. Then holding Γ fixed and partially maximizing $C_n(\mu, \Gamma, \eta, \sigma^2) + M_n(\psi_1, \Gamma)$, we get a likelihood-based objective function $L_n(\Gamma)$ that plays the same role as the moment-based objective function $J(\Gamma)$. Once an basis $\hat{\Gamma}$ is obtained by maximizing $L_n(\Gamma)$, the final estimators follow from (A.3) and (A.4).

Two kinds of assumptions on the distribution of $\mathbf{X}|W$ could be made depending on the sampling model. One is to assume that the predictors and the weights are independently sampled and that the predictor vector follows a multivariate normal distribution. By Proposition 2, we have $\mathcal{E}_{\sigma^2 \Sigma_{\mathbf{X}(W)}^{-1}}(\beta_t) = \mathcal{E}_{\Sigma_{\mathbf{X}}}(\beta_t)$ and the partially maximized log-likelihood up to a constant is (see derivation in the Supplement):

$$L_n(\Gamma) = -\frac{n}{2} \left\{ \log(s_{\Gamma}^2) + \log|\Gamma^T \mathbf{S}_{\mathbf{X}} \Gamma| + \log|\Gamma^T \mathbf{S}_{\mathbf{X}}^{-1} \Gamma| \right\}, \quad (\text{A.5})$$

where $s_{\Gamma}^2 = s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \Gamma (\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma)^{-1} \Gamma^T \mathbf{S}_{\mathbf{X}Y(W)}$ is the estimator of σ^2 from (A.3) with fixed Γ and $s_{Y(W)}^2 = \sum_{i=1}^n W_i (Y_i - \sum_{j=1}^n W_j Y_j / n) / n$ is the sample weighted variance for Y .

Alternatively, it might be reasonable to assume that $\mathbf{X}|W$ has a weighed variance similar to the error variance from $Y|(\mathbf{X}, W)$: $\mathbf{X}|W \sim N(\mu_{\mathbf{X}}, W^{-1} \Sigma)$. Then the partially maximized log-likelihood up to a constant is (see derivation in Section A.2):

$$\begin{aligned} L_n(\Gamma) &= -\frac{n}{2} \left\{ \log|\Gamma^T s^2 \mathbf{S}_{\mathbf{X}(W)}^{-1} \Gamma| + \log|\Gamma^T (s^2 \mathbf{S}_{\mathbf{X}(W)}^{-1} + \hat{\beta} \hat{\beta}^T)^{-1} \Gamma| \right\} \\ &= -\frac{n}{2} \left\{ \log|\Gamma^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \Gamma| + \log|\Gamma^T \mathbf{S}_{\mathbf{X}|Y(W)} \Gamma| \right\}, \end{aligned} \quad (\text{A.6})$$

where $\mathbf{S}_{\mathbf{X}|Y(W)} = \mathbf{S}_{\mathbf{X}(W)} - \mathbf{S}_{\mathbf{X}Y(W)} \mathbf{S}_{\mathbf{X}Y(W)}^T / s_{Y(W)}^2$ and the first equality shows equivalence between maximizing this likelihood-based objective function of Γ and minimizing the moment based objective function $J(\Gamma)$ with $\hat{\mathbf{M}} = s^2 \mathbf{S}_{\mathbf{X}(W)}^{-1}$ and $\hat{\mathbf{U}} = \hat{\beta} \hat{\beta}^T$.

816 The asymptotic results in Proposition 4 hold for $\hat{\beta}_{\text{env}}$ if (i) $\mathbf{X}$ is normal and independent of
817 W , and we use (A.5) to get $\hat{\Gamma}$; or (ii) $\mathbf{X}|W \sim N(\mu_{\mathbf{X}}, W^{-1}\Sigma)$ and we use (A.6) to get $\hat{\Gamma}$. Then
818 the asymptotic covariance of the envelope estimator will be asymptotically less than or equal
819 to that of the standard WLS estimator. Moreover, as long as the model (A.1) holds and both
820 ε and $\mathbf{X}|W$ follow a distribution with finite fourth moments, asymptotic normality of $\hat{\beta}_{\text{env}}$ is
821 guaranteed no matter which $L_n(\Gamma)$ we choose to maximize. Finally, if we fix $W = 1$, then both
822 objective functions in (A.5) and in (A.6) reduces to (2.1), which corresponds to the envelope
823 model for reducing $\mathbf{X}$ in the homoscedastic linear models (Cook et al. 2013).

824 A.2 Derivations for equation (A.6)

825 The normality assumption $Y|(\mathbf{X}, W) \sim N(\mu + \beta^T \mathbf{X}, \sigma^2/W)$ leads to the following conditional
826 log-likelihood:

$$C_n(\mu, \beta, \sigma^2) = -n/2 \log(\sigma^2) - \sum_{i=1}^n W_i(Y_i - \mu - \beta^T \mathbf{X}_i)/(2\sigma^2). \quad (\text{A.7})$$

827 By straightforward computation, the maximum likelihood estimators is obtained as:

$$\hat{\beta} = \mathbf{S}_{\mathbf{X}(W)}^{-1} \mathbf{S}_{\mathbf{X}Y(W)} \quad (\text{A.8})$$

$$\hat{\mu} = \sum_{i=1}^n W_i Y_i / n - \hat{\beta}^T \sum_{i=1}^n W_i \mathbf{X}_i / n := \bar{Y}_{(W)} - \hat{\beta}^T \bar{\mathbf{X}}_{(W)} \quad (\text{A.9})$$

$$\hat{\sigma}^2 = s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \mathbf{S}_{\mathbf{X}Y(W)} := s^2 \quad (\text{A.10})$$

828 Given Γ , which is a semi-orthogonal basis matrix for $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\beta_t)$, we then only need to maximize
829 $C_n(\mu, \beta, \sigma^2)$ under the constraint that $\beta = \Gamma \eta$. The resulting estimators are:

$$\hat{\eta}_{\Gamma} = (\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma)^{-1} \Gamma^T \mathbf{S}_{\mathbf{X}Y(W)} \quad (\text{A.11})$$

$$\hat{\mu}_{\Gamma} = \bar{Y}_{(W)} - \hat{\eta}_{\Gamma}^T \Gamma^T \bar{\mathbf{X}}_{(W)} \quad (\text{A.12})$$

$$\hat{\beta}_{\Gamma} = \Gamma \hat{\eta}_{\Gamma} = \mathbf{P}_{\Gamma(\mathbf{S}_{\mathbf{X}(W)})} \hat{\beta} \quad (\text{A.13})$$

$$\hat{\sigma}_{\Gamma}^2 = s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \Gamma (\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma)^{-1} \Gamma^T \mathbf{S}_{\mathbf{X}Y(W)} := s_{\Gamma}^2. \quad (\text{A.14})$$

830 If we replace Γ by its $\sqrt{n}$ -consistent estimator $\hat{\Gamma}$ from either a moment-based objective func-
831 tion or a likelihood-based objective function, then the above estimators will be our envelope
832 estimators.

833 To obtain a likelihood-based objective function for Γ , we need to compute the partially
 834 maximized conditional likelihood of $Y|(\mathbf{X}, W)$: $C_n(\Gamma) = C_n(\hat{\mu}_\Gamma, \hat{\beta}_\Gamma, \hat{\sigma}_\Gamma)$, plus the partially
 835 maximized marginal likelihood $M_n(\Gamma)$ of $\mathbf{X}|W$. First, we can get $C_n(\Gamma)$ from above deriva-
 836 tions as:

$$C_n(\Gamma) = -n/2 \log(s_\Gamma^2) - n/2. \quad (\text{A.15})$$

837 Assuming that $\mathbf{X}$ and W are independent and that $\mathbf{X} \sim N(\mu_\mathbf{X}, \Sigma_\mathbf{X})$, we have by Proposi-
 838 tion 2 that $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\beta_t) = \mathcal{E}_{\Sigma_\mathbf{X}}(\beta_t)$. The partially maximized log-likelihood $M_n(\Gamma)$ follows from
 839 Proposition 3:

$$M_n(\Gamma) = -\frac{n}{2} \{ \log |\Gamma^T \mathbf{S}_\mathbf{X} \Gamma| + \log |\Gamma^T \mathbf{S}_\mathbf{X}^{-1} \Gamma| + \log |\mathbf{S}_\mathbf{X}| \}. \quad (\text{A.16})$$

840 Therefore, the final objective function for Γ is

$$\begin{aligned} L_n(\Gamma) &= C_n(\Gamma) + M_n(\Gamma) \\ &= -\frac{n}{2} \{ 1 + \log(s_\Gamma^2) + \log |\Gamma^T \mathbf{S}_\mathbf{X} \Gamma| + \log |\Gamma^T \mathbf{S}_\mathbf{X}^{-1} \Gamma| + \log |\mathbf{S}_\mathbf{X}| \} \\ &= -\frac{n}{2} \log \{ s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \Gamma (\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma)^{-1} \Gamma^T \mathbf{S}_{\mathbf{X}Y(W)} \} \\ &\quad - \frac{n}{2} \{ \log |\Gamma^T \mathbf{S}_\mathbf{X} \Gamma| + \log |\Gamma^T \mathbf{S}_\mathbf{X}^{-1} \Gamma| \} + \text{constant}. \end{aligned} \quad (\text{A.17})$$

841 Alternatively, if we assume $\mathbf{X}|W \sim N(\mu_\mathbf{X}, W^{-1}\Sigma)$, $\Sigma > 0$ and use the fact that $E(W) =$
 842 1, then $\mu_\mathbf{X} = E(W\mathbf{X}) = E(\mathbf{X})$. Moreover, the covariance

$$\begin{aligned} \Sigma_{\mathbf{X}(W)} &= E [W(\mathbf{X} - \mu_\mathbf{X})(\mathbf{X} - \mu_\mathbf{X})^T] \\ &= E \{ E [W(\mathbf{X} - \mu_\mathbf{X})(\mathbf{X} - \mu_\mathbf{X})^T | W] \} \\ &= E \{ W \cdot \text{var}(\mathbf{X}|W) \} = E(W \cdot W^{-1}\Sigma) \\ &= \Sigma. \end{aligned}$$

The maximum likelihood estimators for $\mu_\mathbf{X}$ and Σ are then $\hat{\mu}_\mathbf{X} = \bar{\mathbf{X}}_{(W)}$ and $\hat{\Sigma}_\mathbf{X} = \mathbf{S}_{\mathbf{X}(W)}$.
 Follow the same calculation in the proof of Proposition 3, we have

$$M_n(\Gamma) = -\frac{n}{2} \left\{ \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma| + \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \Gamma| + \log |\mathbf{S}_{\mathbf{X}(W)}| \right\}.$$

Therefore, the final objective function for Γ up to a constant is

$$\begin{aligned}
L_n(\Gamma) &= C_n(\Gamma) + M_n(\Gamma) \\
&= -\frac{n}{2} \left\{ \log(s_\Gamma^2) + \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma| + \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \Gamma| \right\} \\
&= -\frac{n}{2} \log \left\{ s_{Y(W)}^2 - \mathbf{S}_{\mathbf{X}Y(W)}^T \Gamma (\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma)^{-1} \Gamma^T \mathbf{S}_{\mathbf{X}Y(W)} \right\} \\
&\quad -\frac{n}{2} \left\{ \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)} \Gamma| + \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \Gamma| \right\} \\
&= -\frac{n}{2} \left\{ \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \Gamma| + \log |\Gamma^T (\mathbf{S}_{\mathbf{X}(W)} - \mathbf{S}_{\mathbf{X}Y(W)} s_{Y(W)}^{-2} \mathbf{S}_{\mathbf{X}Y(W)}^T) \Gamma| \right\} \\
&:= -\frac{n}{2} \left\{ \log |\Gamma^T \mathbf{S}_{\mathbf{X}(W)}^{-1} \Gamma| + \log |\Gamma^T \mathbf{S}_{\mathbf{X}|Y(W)} \Gamma| \right\}.
\end{aligned} \tag{A.18}$$

A.3 Envelopes for Cox regression

Continuing from the discussion of Section A.3, the log partial likelihood for β is

$$C_n(\beta) = \sum_{i=1}^n \delta_i \left\{ \beta^T \mathbf{Z}_i - \log \left(\sum_{j=1}^n I(T_j \geq T_i) \exp(\beta^T \mathbf{Z}_j) \right) \right\}, \tag{A.19}$$

whose first-order derivative (score function) and second-order derivative (Hessian matrix) are

$$\begin{aligned}
C'_n(\beta) &= \sum_{i=1}^n \delta_i \left\{ \mathbf{Z}_i - \sum_{j=1}^n W_{ij} \mathbf{Z}_j \right\}, \\
C''_n(\beta) &= -\sum_{i=1}^n \delta_i \left\{ \left(\sum_{j=1}^n W_{ij} \mathbf{Z}_j \mathbf{Z}_j^T \right) - \left(\sum_{j=1}^n W_{ij} \mathbf{Z}_j \right) \left(\sum_{j=1}^n W_{ij} \mathbf{Z}_j^T \right) \right\},
\end{aligned}$$

where $W_{ij} = I(T_j \geq T_i) \exp(\beta^T \mathbf{Z}_j) / \left\{ \sum_{j=1}^n I(T_j \geq T_i) \exp(\beta^T \mathbf{Z}_j) \right\}$ is the weight with respect to subject i . The Fisher information matrix is $\mathbf{J}(\beta) = -\lim_{n \rightarrow \infty} n^{-1} C''_n(\beta)$. The implementation of envelope Cox model is similar to that of envelope GLM in Section 3.2 by adding the same $M_n(\Gamma)$ to the objective function. Theoretical properties of the envelope Cox model is included in Section 3.1.

B Additional simulations: least squares regression

In this section, we report the results of numerical studies on the robustness of three estimators in linear regression: (1) the ordinary least squares (OLS) estimator, (2) the PLS estimator from the SIMPLS algorithm, and (3) the moment-based envelope estimator from the 1D algorithm in

t_6	$\Sigma_{\mathbf{x}}$			$0.01\Sigma_{\mathbf{x}}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	57	34	6	85	44	48
S.E.	1.6	0.08	0.003	0.6	1.2	2.7
$n = 200$	40	32	3	80	35	14
S.E.	1.8	0.8	0.09	1.0	0.8	0.6
$n = 800$	23	27	1	71	30	7
S.E.	1.4	1.0	0.5	1.5	0.8	0.2

$U(0, 1)$	$\Sigma_{\mathbf{x}}$			$0.01\Sigma_{\mathbf{x}}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	27	33	1.8	74	33	12.6
S.E.	2	1	0.05	2	1	0.5
$n = 200$	14	29	0.8	59	31	5.4
S.E.	1	1	0.03	2	1	0.2
$n = 800$	6.8	24	0.4	44	25	2.8
S.E.	0.4	1	0.01	2	1	0.1

χ_4^2	$\Sigma_{\mathbf{x}}$			$0.01\Sigma_{\mathbf{x}}$		
	Standard	SIMPLS	1D	Standard	SIMPLS	1D
$n = 50$	73	35	13	88	60	75
S.E.	1.4	1.0	0.8	0.3	2	2
$n = 200$	57	33	6.0	85	47	52
S.E.	2	1	0.2	1	1	2
$n = 800$	38	27	2.9	79	35	13
S.E.	2	1	0.1	1	1	1

Table 7: Simulations for heterogeneous covariance matrix $\Sigma_{\mathbf{x}} = \mathbf{\Gamma}\mathbf{\Omega}\mathbf{\Gamma}^T + \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T$. Averaged angles between the true parameter vector and three estimators for the least squares simulation. Standard error is included below each value.

t_6		Standard	SIMPLS	1D
$n = 50$	$\angle$	27.8 (0.7)	33.3 (0.7)	30.9 (0.9)
	LR	1.32	0.79	1.22
$n = 200$	$\angle$	13.1 (0.3)	18.1 (0.4)	14.0 (0.4)
	LR	1.06	0.91	1.05
$n = 800$	$\angle$	6.6 (0.2)	9.0 (0.2)	6.7 (0.2)
	LR	1.02	0.98	1.02

$U(0, 1)$		Standard	SIMPLS	1D
$n = 50$	$\angle$	6.8 (0.2)	22.2 (0.5)	6.9 (0.2)
	LR	1.02	0.71	1.01
$n = 200$	$\angle$	3.3 (0.08)	11.7 (0.3)	3.3 (0.08)
	LR	1.00	0.89	1.00
$n = 800$	$\angle$	1.7 (0.04)	5.9 (0.1)	1.7 (0.04)
	LR	1.00	0.97	1.00

χ_4^2		Standard	SIMPLS	1D
$n = 50$	$\angle$	51 (1.4)	50 (1.2)	53 (1.4)
	LR	1.42	1.21	1.39
$n = 200$	$\angle$	29.9 (0.8)	30.6 (0.7)	32.6 (0.9)
	LR	1.39	1.16	1.28
$n = 800$	$\angle$	15.0 (0.4)	16.2 (0.4)	15.9 (0.4)
	LR	1.08	1.04	1.07

Table 8: Simulations for isotropic covariance matrix $\Sigma_{\mathbf{X}} = 0.01\mathbf{I}_p$. Averaged angles and length ratios between the true parameter vector and various estimators. Standard errors of the averaged angles are included in parentheses.

856 Cook and Zhang (2014a). Same as in Section 5, for each of the simulation settings, we simu-
857 lated 100 datasets for each of the three sample sizes: $n = 50, 200$ and 800 . The true dimension
858 of the envelope was always used in estimation. We used the Matlab function *plsregress* to com-
859 pute the PLS estimator. From Proposition 4.1 in Cook et al. (2013), we know that the SIMPLS
860 algorithm produces $\sqrt{n}$ -consistent envelope estimators when the number of component in the
861 algorithm is chosen to be the dimension of the envelope. The univariate response was gener-
862 ated as $Y = \beta^T \mathbf{X} + \varepsilon$, where the elements of ε were $U(0, 1)$ variates. The covariance matrix
863 $\Sigma_{\mathbf{X}} = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T$, was generated as it was for logistic and the Poisson regression, ex-
864 cept the eigenvalues of Ω were 1 and 5, while the eigenvalues of Ω_0 ranged from 0.0002 to 20.
865 These choices introduce a degree of collinearity in the immaterial variation $\text{var}(\Gamma_0^T \mathbf{X})$, which
866 is the kind of regression that PLS estimators were designed to handle. We also used $0.01 \Sigma_{\mathbf{X}}$
867 instead of $\Sigma_{\mathbf{X}}$ to compare performance with a weaker signal.

868 The angles are shown in Table 7, the lengths being excluded because they produced no
869 comparative conclusion beyond those available from the results shown. The 1D algorithm did
870 the best in all situations, except when a relatively small sample sizes was combined with a weak
871 signal ($0.01 \Sigma_{\mathbf{X}}$) and t_6 or χ_4^2 errors. In those situations the SIMPLS might have a slight edge.
872 Because there was strong collinearity in $\mathbf{X}$, OLS did notably worse than the other two methods,
873 as expected. One exception to this was when $n = 800$ with uniformly distributed predictors.

874 We also repeated the simulations of Table 7 when $\Sigma_{\mathbf{X}} = 0.01 \mathbf{I}_p$, so envelope methods
875 reduce to the standard OLS estimator. The OLS and 1D estimators performed essentially the
876 same in all cases, while the SIMPLS algorithm did a bit worse with t_6 and χ_4^2 errors, and
877 notably worse with $U(0, 1)$ errors. Table 8 gives the results of the least squares simulation with
878 predictor covariance matrix $\Sigma_{\mathbf{X}} = 0.01 \mathbf{I}_p$.

879 **C Implementation details for envelope GLM in Section 3.2**

880 Let $\mathbb{X}_n$ be the $n \times p$ data matrix of $\mathbf{X}_i$'s and let $\mathbb{C}_n$ be the $n \times 1$ data matrix of $\mathcal{C}'(\vartheta_i)$, where
881 $\mathcal{C}'(\vartheta)$ is the first order derivative from Table 1. Then the $p \times u$ gradient matrix of $\partial C_n(\Gamma)/\partial \Gamma$
882 can be computed conveniently as follows.

Lemma 1.

$$\begin{aligned}
\frac{\partial C_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= \mathbb{X}_n^T \mathbb{C}_n \boldsymbol{\eta}^T + \mathbf{S}_{\mathbf{X}V(W)} \mathbb{C}_n^T \mathbb{X}_n \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \\
&\quad - \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbb{X}_n^T \mathbb{C}_n \mathbf{S}_{\mathbf{X}V(W)}^T \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \\
&\quad - \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}V(W)} \mathbb{C}_n^T \mathbb{X}_n \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}; \\
\frac{\partial M_n(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} &= -n \{ \mathbf{S}_{\mathbf{X}} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}} \mathbf{\Gamma})^{-1} + \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}}^{-1} \mathbf{\Gamma})^{-1} \}.
\end{aligned}$$

883 To check this Lemma, we compared the analytical form of the derivatives to the numerical
884 derivatives. The numerical derivatives of a function $f(\mathbf{\Gamma})$ is computed by varying each element
885 of $\mathbf{\Gamma}$ by a small number $\delta = 1.0 \times 10^{-6}$ and evaluate the function:

$$\left[\frac{\partial f(\mathbf{\Gamma})}{\partial \mathbf{\Gamma}} \right]_{ij} = \frac{f(\mathbf{\Gamma} + \delta \mathbf{G}_{ij}) - f(\mathbf{\Gamma} - \delta \mathbf{G}_{ij})}{2\delta}, \quad (\text{C.1})$$

886 where $\mathbf{G}_{ij} \in \mathbb{R}^{p \times u}$ is zero matrix except its ij -th entry is one. We found the difference between
887 our analytical derivative and the numerical derivative is around 10^{-7} for every element.

888 *Proof.* This proof of Lemma 1 mainly involves rather complicated matrix differentiation.

$$\begin{aligned}
\frac{\partial \vartheta_i}{\partial \text{vec}(\mathbf{\Gamma})} &= \frac{\partial}{\partial \text{vec}(\mathbf{\Gamma})} \{ \mathbf{X}_i^T \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}V(W)} \} \\
&= \frac{\partial}{\partial \text{vec}(\mathbf{\Gamma})} \{ (\mathbf{S}_{\mathbf{X}V(W)}^T \otimes \mathbf{X}_i^T) \text{vec}[\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T] \} \\
&= \left[\frac{\partial}{\partial \text{vec}(\mathbf{\Gamma})} \text{vec}(\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T) \right]^T (\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i). \quad (\text{C.2})
\end{aligned}$$

889 Apply the following chain rule:

$$\partial \text{vec}(\mathbf{ABC}) = (\mathbf{C}^T \mathbf{B}^T \otimes \mathbf{I}) \partial \text{vec}(\mathbf{A}) + (\mathbf{C}^T \otimes \mathbf{A}) \partial \text{vec}(\mathbf{B}) + (\mathbf{I} \otimes \mathbf{AB}) \partial \text{vec}(\mathbf{C}),$$

890 we get

$$\begin{aligned}
\partial \text{vec}(\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T) / \partial \text{vec}(\mathbf{\Gamma}) &= \mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3 \\
&= [\mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \otimes \mathbf{I}_p] \mathbf{I}_{pu} \\
&\quad + (\mathbf{\Gamma} \otimes \mathbf{\Gamma}) \frac{\partial \text{vec}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}}{\partial \text{vec}(\mathbf{\Gamma})} \\
&\quad + [\mathbf{I}_p \otimes \mathbf{\Gamma} (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}] \mathbf{K}_{pu}.
\end{aligned}$$

891 The second term after the last equal sign is computed by noticing that

$$\begin{aligned}
\partial \text{vec} \{ (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \} &= -\text{vec} \{ (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} [\partial (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})] (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \} \\
&= -((\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \otimes (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}) \partial \text{vec}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma}),
\end{aligned}$$

892 and that

$$\begin{aligned}\frac{\partial \text{vec}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})}{\partial \text{vec}(\mathbf{\Gamma})} &= (\mathbf{I}_u \otimes \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)}) \frac{\partial \text{vec}(\mathbf{\Gamma})}{\partial \text{vec}(\mathbf{\Gamma})} + (\mathbf{\Gamma}^T \otimes \mathbf{I}_u) \frac{\partial \text{vec}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)})}{\partial \text{vec}(\mathbf{\Gamma})} \\ &= (\mathbf{I}_u \otimes \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)}) + (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \otimes \mathbf{I}_u) \mathbf{K}_{pu}.\end{aligned}$$

893 Hence the second term equals to

$$\begin{aligned}\mathbf{T}_2 &= -(\mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)}) \\ &\quad - (\mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}) \mathbf{K}_{pu}.\end{aligned}$$

894 Finally, substitute the three terms into (C.2), we have the three terms from

$$\begin{aligned}\frac{\partial \vartheta_i}{\partial \text{vec}(\mathbf{\Gamma})} &= (\mathbf{T}_1 + \mathbf{T}_2 + \mathbf{T}_3)^T (\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i) \\ \mathbf{T}_1^T (\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i) &= [(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i] \\ &= [\boldsymbol{\eta} \otimes \mathbf{X}_i] \\ \mathbf{T}_3^T (\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i) &= \{[\mathbf{I}_p \otimes \mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1}] \mathbf{K}_{pu}\}^T (\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i) \\ &= \mathbf{K}_{pu}^T [\mathbf{S}_{\mathbf{X}V(W)} \otimes (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i]\end{aligned}$$

895

$$\begin{aligned}\mathbf{T}_2^T (\mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{X}_i) &= -[(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}V(W)} \otimes \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i] \\ &\quad - \mathbf{K}_{pu}^T [\mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}V(W)} \otimes (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i] \\ &= -[\boldsymbol{\eta} \otimes \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i] \\ &\quad - \mathbf{K}_{pu}^T [\mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma} \boldsymbol{\eta} \otimes (\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i]\end{aligned}$$

896 Notice that each of the above terms is a $pu \times 1$ vector, and then the $\mathbf{K}_{pu}^T$ serves only to change

897 $\text{vec}(\mathbf{A}) \rightarrow \text{vec}(\mathbf{A}^T)$, therefore

$$\begin{aligned}\frac{\partial \vartheta_i}{\partial \text{vec}(\mathbf{\Gamma})} &= \text{vec}(\mathbf{M}_1 + \mathbf{M}_3^T) + \text{vec}(\mathbf{M}_2 + \mathbf{M}_4^T) \\ \mathbf{M}_1 &= [\mathbf{X}_i \cdot \boldsymbol{\eta}^T] \\ \mathbf{M}_3 &= [(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i \cdot \mathbf{S}_{\mathbf{X}V(W)}^T] \\ \mathbf{M}_2 &= -[\mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma}(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i \cdot \boldsymbol{\eta}^T] \\ \mathbf{M}_4 &= -[(\mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)} \mathbf{\Gamma})^{-1} \mathbf{\Gamma}^T \mathbf{X}_i \cdot \boldsymbol{\eta}^T \mathbf{\Gamma}^T \mathbf{S}_{\mathbf{X}(W)}],\end{aligned}$$

898 where $\mathbf{M}_1, \mathbf{M}_2$ and $\mathbf{M}_3^T, \mathbf{M}_4^T$ are all $p \times u$ matrices.

To implement, we have the gradient matrix as

$$\frac{\partial C_n(\Gamma)}{\partial \Gamma} + \frac{\partial M_n(\Gamma)}{\partial \Gamma} = \frac{\partial C_n(\Gamma)}{\partial \Gamma} - n\{\mathbf{S}_X \Gamma (\Gamma^T \mathbf{S}_X \Gamma)^{-1} + \mathbf{S}_X^{-1} \Gamma (\Gamma^T \mathbf{S}_X^{-1} \Gamma)^{-1}\}, \quad (\text{C.3})$$

where $\frac{\partial C_n(\Gamma)}{\partial \Gamma}$ is re-arranged as a $p \times u$ matrix

$$\begin{aligned} \frac{\partial C_n(\Gamma)}{\partial \Gamma} &= \left\{ \sum_{i=1}^n C'(\vartheta_i) \frac{\partial \vartheta_i}{\partial \text{vec}(\Gamma)} \right\}_{p \times u} \\ &= \sum_{i=1}^n C'(\vartheta_i) \{ \mathbf{X}_i \cdot \boldsymbol{\eta}^T + \mathbf{S}_{XV(W)} \cdot \mathbf{X}_i^T \Gamma (\Gamma^T \mathbf{S}_{X(W)} \Gamma)^{-1} \} \\ &\quad - \sum_{i=1}^n C'(\vartheta_i) [\mathbf{S}_{X(W)} \Gamma (\Gamma^T \mathbf{S}_{X(W)} \Gamma)^{-1} \Gamma^T \mathbf{X}_i \cdot \boldsymbol{\eta}^T] \\ &\quad - \sum_{i=1}^n C'(\vartheta_i) [(\Gamma^T \mathbf{S}_{X(W)} \Gamma)^{-1} \Gamma^T \mathbf{X}_i \cdot \boldsymbol{\eta}^T \Gamma^T \mathbf{S}_{X(W)}]^T \\ &= \mathbb{X}_n^T \mathbb{C}_n \boldsymbol{\eta}^T + \mathbf{S}_{XV(W)} \mathbb{C}_n^T \mathbb{X}_n \Gamma (\Gamma^T \mathbf{S}_{X(W)} \Gamma)^{-1} \\ &\quad - \mathbf{S}_{X(W)} \Gamma (\Gamma^T \mathbf{S}_{X(W)} \Gamma)^{-1} \Gamma^T \mathbb{X}_n^T \mathbb{C}_n \boldsymbol{\eta}^T \\ &\quad - \mathbf{S}_{X(W)} \Gamma \boldsymbol{\eta} \mathbb{C}_n^T \mathbb{X}_n \Gamma (\Gamma^T \mathbf{S}_{X(W)} \Gamma)^{-1}. \end{aligned}$$

□

D Enveloping a matrix-valued parameter

Definition 3 is sufficiently general to cover a matrix-valued parameter $\phi \in \mathbb{R}^{r \times c}$ by considering the $\text{avar}(\sqrt{n} \text{vec}(\hat{\phi}))$ -envelope of $\text{vec}(\phi)$. However, matrix-valued parameters come with additional structure that is often desirable to maintain during estimation. For instance, in the multivariate linear model reviewed in Section 1.3, envelope construction was constrained to reflect separate row and column reduction of the matrix parameter β . The advantages of row and column reductions have been discussed by Li et al. (2010) and Hung et al. (2012). Although our primary focus is on vector-valued parameters, in this section we indicate how to adapt for a matrix-valued parameter.

Suppose that $\sqrt{n}(\hat{\phi} - \phi_t)$ converges to a matrix normal distribution with mean 0, column variance $\Delta_L \in \mathbb{S}^{r \times r}$ and row variance $\Delta_R \in \mathbb{S}^{c \times c}$. (See Dawid (1981) for background on the matrix normal distribution.) Then

$$\sqrt{n}(\text{vec}(\hat{\phi}) - \text{vec}(\phi_t)) \rightarrow N(0, \Delta_R \otimes \Delta_L), \quad (\text{D.1})$$

and direct application of Definition 3 yields the envelope $\mathcal{E}_{\Delta_R \otimes \Delta_L}(\text{vec}(\phi_t))$. However, this envelope may not preserve the intrinsic row-column structure of ϕ_t . In the following definition we introduce a restricted class of envelopes that maintain the matrix structure of ϕ_t . We define the Kronecker product of two subspaces as $\mathcal{A} \otimes \mathcal{B} = \{\mathbf{a} \otimes \mathbf{b} | \mathbf{a} \in \mathcal{A}; \mathbf{b} \in \mathcal{B}\}$.

Definition 4. Assume that $\hat{\phi}$ is asymptotically matrix normal as give in (D.1). Then the tensor envelope for ϕ_t , denoted by $\mathcal{K}_{\Delta_R \otimes \Delta_L}(\phi_t)$, is defined as the intersection of all reducing subspaces $\mathcal{E}$ of $\Delta_R \otimes \Delta_L$ that contain $\text{span}(\text{vec}(\phi_t))$ and can be written as $\mathcal{E} = \mathcal{E}_R \otimes \mathcal{E}_L$ with $\mathcal{E}_R \subseteq \mathbb{R}^c$ and $\mathcal{E}_L \subseteq \mathbb{R}^r$.

We see from this definition that $\mathcal{K}_{\Delta_R \otimes \Delta_L}(\phi_t)$ always exist and is the smallest subspace with the required properties. Let $\mathbf{L} \in \mathbb{R}^{r \times d_c}$, $d_c < r$, and $\mathbf{R} \in \mathbb{R}^{c \times d_r}$, $d_r < c$ be semi-orthogonal matrices such that $\mathcal{K}_{\Delta_R \otimes \Delta_L}(\phi_t) = \text{span}(\mathbf{R} \otimes \mathbf{L})$. By definition, $\text{span}(\phi_t) \subseteq \text{span}(\mathbf{L})$ and $\text{span}(\phi_t^T) \subseteq \text{span}(\mathbf{R})$. Hence we have $\phi_t = \mathbf{L}\boldsymbol{\eta}\mathbf{R}^T$ and $\text{vec}(\phi_t) = (\mathbf{R} \otimes \mathbf{L})\text{vec}(\boldsymbol{\eta})$ for some $\boldsymbol{\eta} \in \mathbb{R}^{d_c \times d_r}$.

The next proposition shows how to factor $\mathcal{K}_{\Delta_R \otimes \Delta_L}(\phi_t) \in \mathbb{R}^{rc}$ into the tensor product of envelopes $\mathcal{E}_{\Delta_R}(\phi_t^T) \in \mathbb{R}^r$ and $\mathcal{E}_{\Delta_L}(\phi_t) \in \mathbb{R}^c$ for the row and column spaces of ϕ_t . These tensor factors are envelopes in smaller spaces that preserve the row and column structure and can facilitate analysis and interpretation.

Proposition 7. $\mathcal{K}_{\Delta_R \otimes \Delta_L}(\phi_t) = \mathcal{E}_{\Delta_R}(\phi_t^T) \otimes \mathcal{E}_{\Delta_L}(\phi_t)$.

For example, in reference to model (1.1), the distribution of $\hat{\beta} = \mathbf{S}_{\mathbf{X}\mathbf{Y}}^T \mathbf{S}_{\mathbf{X}}^{-1}$ satisfies (D.1) with $\Delta_R = \Sigma_{\mathbf{X}}^{-1}$ and $\Delta_L = \Sigma_{\mathbf{Y}|\mathbf{X}}$. The tensor envelope is then

$$\mathcal{K}_{\Sigma_{\mathbf{X}}^{-1} \otimes \Sigma_{\mathbf{Y}|\mathbf{X}}}(\beta) = \mathcal{E}_{\Sigma_{\mathbf{X}}^{-1}}(\beta^T) \otimes \mathcal{E}_{\Sigma_{\mathbf{Y}|\mathbf{X}}}(\beta).$$

If we are interested in reducing only the column space of β , which corresponds to response reduction, we would use $\mathbb{R}^p \otimes \mathcal{E}_{\Sigma_{\mathbf{Y}|\mathbf{X}}}(\beta)$ for constructing an envelope estimator of β , and then $\beta = \mathbf{L}\boldsymbol{\eta}\mathbf{I}_p = \mathbf{L}\boldsymbol{\eta}$ where $\mathbf{L}$ is a semi-orthogonal basis for $\mathcal{E}_{\Sigma_{\mathbf{Y}|\mathbf{X}}}(\beta)$, which reproduces the envelope construction in Cook et al. (2010). Similarly, if we are interested in only predictor reduction, we would take $\mathcal{E}_{\Sigma_{\mathbf{X}}}(\beta^T) \otimes \mathbb{R}^r$, which reproduces the envelope construction in Cook et al. (2013). More generally, the definition of the tensor envelope and Proposition 7 straightforwardly connects and combines the envelope models in the predictor space and in the response space, leading to the simultaneous envelope method in Cook and Zhang (2014b).

E Proofs

E.1 Proposition 5 and Corollary 2

Proof. The asymptotic normality and the asymptotic covariance of $\hat{\phi}_n$ can be see by definition.

We write the Fisher information $\mathbf{F}(\boldsymbol{\theta}_t)$ as its natural block matrix structure

$$\mathbf{F}(\boldsymbol{\theta}_t) = \mathbf{F}(\boldsymbol{\psi}_t, \boldsymbol{\phi}_t) = \begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{C} \end{pmatrix},$$

where $\mathbf{A} = -\mathbf{E}_{\boldsymbol{\theta}_t} \{ (\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi}) / \partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T) \}$, $\mathbf{B} = -\mathbf{E}_{\boldsymbol{\theta}_t} \{ (\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi}) / \partial \boldsymbol{\psi} \partial \boldsymbol{\phi}^T) \}$ and $\mathbf{C} = -\mathbf{E}_{\boldsymbol{\theta}_t} \{ (\partial^2 L(\boldsymbol{\psi}, \boldsymbol{\phi}) / \partial \boldsymbol{\phi} \partial \boldsymbol{\phi}^T) \}$. Then $\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t) = \mathbf{F}_{\phi\phi}^{-1}(\boldsymbol{\theta}_t) = (\mathbf{C} - \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B})^{-1}$ by block matrix inversion.

The Hessian matrix for the constrained optimization (4.2)-(4.4) is

$$\mathbf{F}(\boldsymbol{\psi}_t, \boldsymbol{\eta}_t) = \begin{pmatrix} \mathbf{A} & \mathbf{B}\boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^T \mathbf{B}^T & \boldsymbol{\Gamma}^T \mathbf{C} \boldsymbol{\Gamma} \end{pmatrix}, \quad (\text{E.1})$$

since $\partial / \partial \boldsymbol{\eta} = \boldsymbol{\Gamma}^T \cdot \partial / \partial \boldsymbol{\phi}$ under the constraint that $\boldsymbol{\phi} = \boldsymbol{\Gamma} \boldsymbol{\eta}$. Therefore the asymptotic variance for $\hat{\phi}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma} \hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}$ is

$$\begin{aligned} \text{avar}(\sqrt{n} \hat{\phi}_{\boldsymbol{\Gamma}}) &= \boldsymbol{\Gamma} \text{avar}(\sqrt{n} \hat{\boldsymbol{\eta}}_{\boldsymbol{\Gamma}}) \boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \mathbf{C} \boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T \mathbf{B}^T \mathbf{A}^{-1} \mathbf{B} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \\ &= \boldsymbol{\Gamma} \{ \boldsymbol{\Gamma}^T \mathbf{V}_{\phi\phi}^{-1}(\boldsymbol{\theta}_t) \boldsymbol{\Gamma} \}^{-1} \boldsymbol{\Gamma}^T \\ &= \mathbf{P}_{\boldsymbol{\Gamma}} \mathbf{V} \mathbf{P}_{\boldsymbol{\Gamma}} - \mathbf{P}_{\boldsymbol{\Gamma}} \mathbf{V} \boldsymbol{\Gamma}_0 (\boldsymbol{\Gamma}_0^T \mathbf{V} \boldsymbol{\Gamma}_0)^{-1} \boldsymbol{\Gamma}_0^T \mathbf{V} \mathbf{P}_{\boldsymbol{\Gamma}}, \end{aligned}$$

where the last equality is obtained by applying the equality (Cook and Forzani 2009),

$$(\boldsymbol{\Gamma}^T \mathbf{V}^{-1} \boldsymbol{\Gamma})^{-1} = \boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma} - \boldsymbol{\Gamma}^T \mathbf{V} \boldsymbol{\Gamma}_0 (\boldsymbol{\Gamma}_0^T \mathbf{V} \boldsymbol{\Gamma}_0)^{-1} \boldsymbol{\Gamma}_0^T \mathbf{V} \boldsymbol{\Gamma}. \quad (\text{E.2})$$

Then by requiring that $\text{span}(\boldsymbol{\Gamma})$ reduces $\mathbf{V}_{\phi\phi}(\boldsymbol{\theta}_t)$, we have $\text{avar}(\sqrt{n} \hat{\phi}_{\boldsymbol{\Gamma}}) = \mathbf{P}_{\boldsymbol{\Gamma}} \mathbf{V} \mathbf{P}_{\boldsymbol{\Gamma}}$.

Similarly for the asymptotic variance of $\hat{\boldsymbol{\psi}}_n$ and $\hat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}$ we have the following results from block matrix inversion and (E.2).

$$\begin{aligned} \text{avar}(\sqrt{n} \hat{\boldsymbol{\psi}}_{\boldsymbol{\Gamma}}) &= (\mathbf{A} - \mathbf{B} \boldsymbol{\Gamma} (\boldsymbol{\Gamma}^T \mathbf{C} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^T \mathbf{B}^T)^{-1} \\ &\leq (\mathbf{A} - \mathbf{B} \mathbf{C}^{-1} \mathbf{B}^T)^{-1} = \text{avar}(\sqrt{n} \hat{\boldsymbol{\psi}}_n). \end{aligned}$$

□

955 E.2 Proposition 6

956 From Cook and Zhang (2014a; Proposition 6), we know that $\widehat{\Gamma}\widehat{\Gamma}^T$ is $\sqrt{n}$ -consistent for the pro-
 957 jection onto the envelope $\mathcal{E}_{\mathbf{V}_{\phi\phi}}(\phi_t)$. Then the proof follows from the proof of $\sqrt{n}$ -consistency
 958 of the k -th direction in the 1D algorithm, where we replace the sample objective function $J_k(\mathbf{g})$
 959 with $L_n(\boldsymbol{\psi}, \widehat{\Gamma}\boldsymbol{\eta})$.

960 E.3 Proposition 2

961 *Proof.* The envelope of interest is $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\beta_t)$. The Fisher information per observation is then
 962 obtained by taking the expectation of the second derivative of $\log f(y|\boldsymbol{\alpha}, \beta^T \mathbf{X}) + \log g(x|\boldsymbol{\psi})$
 963 with respect to $\boldsymbol{\theta}$. Clearly, the information matrix is block diagonal in $(\boldsymbol{\alpha}, \beta)$ and $\boldsymbol{\psi}$ so we need
 964 only the second derivatives of $\log f(y|\boldsymbol{\alpha}, \beta^T \mathbf{X})$. Then the Fisher information matrix for $(\boldsymbol{\alpha}, \beta)$
 965 can be written unambiguously in the form

$$\mathbf{F}(\boldsymbol{\alpha}_t, \beta_t) = \mathbb{E} \begin{pmatrix} \mathbf{A}(Y, \beta_t^T \mathbf{X}, \boldsymbol{\alpha}_t) & \mathbf{a}(Y, \beta_t^T \mathbf{X}, \boldsymbol{\alpha}_t) \mathbf{X}^T \\ \mathbf{X} \mathbf{a}^T(Y, \beta_t^T \mathbf{X}, \boldsymbol{\alpha}_t) & c(Y, \beta_t^T \mathbf{X}, \boldsymbol{\alpha}_t) \mathbf{X} \mathbf{X}^T \end{pmatrix},$$

966 where $\mathbf{A}$ is the second derivative matrix with respect to $\boldsymbol{\alpha}$, the off diagonal blocks are the cross
 967 derivatives $\partial^2 / \partial \boldsymbol{\alpha} \partial \beta^T$ with all factors collected into $\mathbf{a}$ except $\mathbf{X}$, and $c(Y, \beta_t^T \mathbf{X}, \boldsymbol{\alpha}_t) \mathbf{X} \mathbf{X}^T$ is
 968 the second derivative with respect to β . The asymptotic covariance matrix for $\widehat{\beta}_n$ is then

$$\mathbf{V}_{\beta\beta} = \{\mathbb{E}(c \mathbf{X} \mathbf{X}^T) - \mathbb{E}(\mathbf{X} \mathbf{a}^T) \mathbb{E}^{-1}(\mathbf{A}) \mathbb{E}(\mathbf{a} \mathbf{X}^T)\}^{-1}.$$

969 Since $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\beta_t) = \mathcal{E}_{\mathbf{V}_{\beta\beta}^{-1}}(\beta_t)$, it is sufficient to consider, making the additional assumption that
 970 $\mathbb{E}(\mathbf{X}|\boldsymbol{\Gamma}^T \mathbf{X})$ is a linear function of $\mathbf{X}$ and requiring that $\text{span}(\beta_t) \subseteq \text{span}(\boldsymbol{\Gamma})$,

$$\begin{aligned} \mathbf{P}_{\boldsymbol{\Gamma}} \mathbf{V}_{\beta\beta}^{-1}(\beta_t) \mathbf{Q}_{\boldsymbol{\Gamma}} &= \mathbf{P}_{\boldsymbol{\Gamma}} \{\mathbb{E}(c \mathbf{X} \mathbf{X}^T) - \mathbb{E}(\mathbf{X} \mathbf{a}^T) \mathbb{E}^{-1}(\mathbf{A}) \mathbb{E}(\mathbf{a} \mathbf{X}^T)\} \mathbf{Q}_{\boldsymbol{\Gamma}} \\ &= \mathbf{P}_{\boldsymbol{\Gamma}} \{\mathbb{E}(c \boldsymbol{\Gamma} \boldsymbol{\Gamma}^T \mathbf{X} \cdot \mathbf{X}^T) - \mathbb{E}(\mathbf{X} \mathbf{a}^T) \mathbb{E}^{-1}(\mathbf{A}) \mathbb{E}(\mathbf{a} \mathbf{X}^T)\} \mathbf{Q}_{\boldsymbol{\Gamma}} \\ &= \mathbf{P}_{\boldsymbol{\Gamma}} \{\mathbb{E}(c \mathbf{X} \mathbb{E}(\mathbf{X}^T | Y, \boldsymbol{\Gamma}^T \mathbf{X})) - \mathbb{E}(\mathbf{X} \mathbf{a}^T) \mathbb{E}^{-1}(\mathbf{A}) \mathbb{E}(\mathbf{a} \mathbb{E}(\mathbf{X}^T | Y, \boldsymbol{\Gamma}^T \mathbf{X}))\} \mathbf{Q}_{\boldsymbol{\Gamma}}. \end{aligned}$$

971 This follows since c and $\mathbf{a}$ are functions of only Y and $\beta^T \mathbf{X}$ and $\text{span}(\beta_t) \subseteq \text{span}(\Gamma)$. Next,
 972 using the fact that we are in a regression and thus $Y \perp \mathbf{X} | \beta^T \mathbf{X}$ implies $Y \perp \mathbf{X} | \Gamma^T \mathbf{X}$, we have

$$\begin{aligned}
 \mathbf{P}_\Gamma \mathbf{V}_{\beta\beta}^{-1} \mathbf{Q}_\Gamma &= \mathbf{P}_\Gamma \{E(c\mathbf{X}E(\mathbf{X}^T | \Gamma^T \mathbf{X})) - E(\mathbf{X}\mathbf{a}^T)E^{-1}(\mathbf{A})E(\mathbf{a}E(\mathbf{X}^T | \Gamma^T \mathbf{X}))\} \mathbf{Q}_\Gamma \\
 &= \mathbf{P}_\Gamma \{E(c\mathbf{X}\mathbf{X}^T \mathbf{P}_{\Gamma(\Sigma_{\mathbf{X}})}) - E(\mathbf{X}\mathbf{a}^T)E^{-1}(\mathbf{A})E(\mathbf{a}\mathbf{X}^T \mathbf{P}_{\Gamma(\Sigma_{\mathbf{X}})})\} \mathbf{Q}_\Gamma \\
 &= \mathbf{P}_\Gamma \{E(c\mathbf{X}\mathbf{X}^T) - E(\mathbf{X}\mathbf{a}^T)E^{-1}(\mathbf{A})E(\mathbf{a}\mathbf{X}^T)\} \mathbf{P}_{\Gamma(\Sigma_{\mathbf{X}})} \mathbf{Q}_\Gamma \\
 &= \mathbf{P}_\Gamma \mathbf{V}_{\beta\beta}^{-1} \mathbf{P}_{\Gamma(\Sigma_{\mathbf{X}})} \mathbf{Q}_\Gamma \\
 &= \Gamma \{\Gamma^T \mathbf{V}_{\beta\beta}^{-1} \Gamma\} \{\Gamma^T \Sigma_{\mathbf{X}} \Gamma\}^{-1} \Gamma^T \Sigma_{\mathbf{X}} \mathbf{Q}_\Gamma.
 \end{aligned}$$

973 The second equality above uses the linearity condition and the rest is algebra. Since Γ has
 974 full column rank and the next two matrix factors in $\{\cdot\}$ are positive definite, we see that
 975 $\mathbf{P}_\Gamma \mathbf{V}_{\beta\beta}^{-1} \mathbf{Q}_\Gamma = 0$ if and only if $\Gamma^T \Sigma_{\mathbf{X}} \mathbf{Q}_\Gamma = 0$; that is, if and only if $\text{span}(\Gamma)$ reduces $\Sigma_{\mathbf{X}}$.
 976 Consequently, $\text{span}(\Gamma)$ reduces $\mathbf{V}_{\beta\beta}$ if and only if $\text{span}(\Gamma)$ reduces $\Sigma_{\mathbf{X}}$. Note that in this
 977 proof we did not require α and β to be orthogonal parameters, nor did we require an expo-
 978 nential family regression. The only requirements are the regression structure and the linearity
 979 condition. $\square$

980 E.4 Proposition 3

981 *Proof.* The multivariate normal log-likelihood up to a constant is

$$M_n(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{X}}) = -\frac{n}{2} \{ \log |\Sigma_{\mathbf{X}}| + \text{trace}[\Sigma_{\mathbf{X}}^{-1} \sum_{i=1}^n (\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X}_i - \boldsymbol{\mu}_{\mathbf{X}})^T / n] \}, \quad (\text{E.3})$$

982 Since $\text{span}(\Gamma) = \mathcal{E}_{\Sigma_{\mathbf{X}}}(\beta_t)$ reduces $\Sigma_{\mathbf{X}}$, we need to maximize $M_n(\boldsymbol{\mu}_{\mathbf{X}}, \Sigma_{\mathbf{X}})$ under the con-
 983 straint that $\text{span}(\Gamma)$ reduces $\Sigma_{\mathbf{X}}$. This can be done by substitute $\Sigma_{\mathbf{X}} = \Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T$
 984 into the objective function, where Ω and Ω_0 are two symmetric positive definite matrices. The
 985 resulting objective function $M_n(\boldsymbol{\mu}_{\mathbf{X}}, \Omega, \Omega_0, \Gamma)$ can be first maximized over $\boldsymbol{\mu}_{\mathbf{X}}$ to get $\hat{\boldsymbol{\mu}}_{\mathbf{X}} = \bar{\mathbf{X}}$.
 986 Then the partially maximized objective function can be expressed as

$$\begin{aligned}
 -\frac{2}{n} M_n(\Omega, \Omega_0, \Gamma) &= \log |\Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T| + \text{trace}[(\Gamma \Omega \Gamma^T + \Gamma_0 \Omega_0 \Gamma_0^T)^{-1} \mathbf{S}_{\mathbf{X}}] \\
 &= \log |\Omega| + \log |\Omega_0| + \text{trace}\{(\Gamma \Omega^{-1} \Gamma^T + \Gamma_0 \Omega_0^{-1} \Gamma_0^T) \mathbf{S}_{\mathbf{X}}\} \\
 &= \log |\Omega| + \text{trace}(\Gamma \Omega^{-1} \Gamma^T \mathbf{S}_{\mathbf{X}}) + \log |\Omega_0| + \text{trace}\{(\Gamma_0 \Omega_0^{-1} \Gamma_0^T \mathbf{S}_{\mathbf{X}})\} \\
 &= \log |\Omega| + \text{trace}(\Omega^{-1} \Gamma^T \mathbf{S}_{\mathbf{X}} \Gamma) + \log |\Omega_0| + \text{trace}(\Omega_0^{-1} \Gamma_0^T \mathbf{S}_{\mathbf{X}} \Gamma_0).
 \end{aligned}$$

987 For any symmetric positive definite matrix $\mathbf{M}$, we apply the following result to optimize over
 988 $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ in $M_n(\boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\Gamma})$,

$$\arg \min_{\mathbf{A}} \{\log |\mathbf{A}| + \text{trace}(\mathbf{A}^{-1}\mathbf{M})\} = \mathbf{M}, \quad (\text{E.4})$$

989 where the minimization is over all positive definite symmetric matrices. Therefore we have
 990 $\widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}$, $\widehat{\boldsymbol{\Omega}}_{0,\boldsymbol{\Gamma}} = \boldsymbol{\Gamma}_0^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}_0$ and

$$\widehat{\boldsymbol{\Sigma}}_{\mathbf{X},\boldsymbol{\Gamma}} = \boldsymbol{\Gamma} \widehat{\boldsymbol{\Omega}}_{\boldsymbol{\Gamma}} \boldsymbol{\Gamma}^T + \boldsymbol{\Gamma}_0 \widehat{\boldsymbol{\Omega}}_{0,\boldsymbol{\Gamma}} \boldsymbol{\Gamma}_0^T = \mathbf{P}_{\boldsymbol{\Gamma}} \mathbf{S}_{\mathbf{X}} \mathbf{P}_{\boldsymbol{\Gamma}} + \mathbf{Q}_{\boldsymbol{\Gamma}} \mathbf{S}_{\mathbf{X}} \mathbf{Q}_{\boldsymbol{\Gamma}}. \quad (\text{E.5})$$

991 The remaining partially maximized form is $M_n(\boldsymbol{\Gamma}) = -n/2 \{\log |\boldsymbol{\Gamma}^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}| + \log |\boldsymbol{\Gamma}_0^T \mathbf{S}_{\mathbf{X}} \boldsymbol{\Gamma}_0|\}$
 992 where we have omitted a constant p from $\text{trace}\{\mathbf{I}_u\} + \text{trace}\{\mathbf{I}_{p-u}\}$. The rest of the proof is a
 993 direct consequence of the following implication from Cook et al. (2013): Suppose $\mathbf{A} \in \mathbb{S}^{t \times t}$ is
 994 non-singular and that the column partition $(\mathbf{B}, \mathbf{B}_0) \in \mathbb{R}^{t \times t}$ is orthogonal. Then (1) $|\mathbf{B}^T \mathbf{A} \mathbf{B}| =$
 995 $|\mathbf{A}| \cdot |\mathbf{B}_0^T \mathbf{A}^{-1} \mathbf{B}_0|$ and (2) $|\mathbf{A}| \leq |\mathbf{B}^T \mathbf{A} \mathbf{B}| \cdot |\mathbf{B}_0^T \mathbf{A} \mathbf{B}_0|$ if and only if $\text{span}(\mathbf{B})$ reduces $\mathbf{A}$. $\square$

996 E.5 Proposition 4

997 *Proof.* Let $\mathbf{E}_p \in \mathbb{R}^{p^2 \times p(p+1)/2}$ denote the expansion operator and let $\mathbf{C}_p \in \mathbb{R}^{p(p+1)/2 \times p^2}$ denote
 998 the contraction operator such that they connect vec and vech operators as $\text{vec}(\mathbf{A}) = \mathbf{E}_p \text{vech}(\mathbf{A})$
 999 and $\text{vech}(\mathbf{A}) = \mathbf{C}_p \text{vec}(\mathbf{A})$ for any $\mathbf{A} \in \mathbb{S}^{p \times p}$.

1000 Because the standard estimator $(\widehat{\boldsymbol{\alpha}}_n, \widehat{\boldsymbol{\beta}}_n)$ and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}} = \mathbf{S}_{\mathbf{X}}$ were obtained separately from
 1001 $C_n(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and the marginal multivariate normal likelihood of $\mathbf{X}$, and because that the pa-
 1002 rameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are orthogonal, the asymptotic variance of $\widehat{\mathbf{h}}_n$ is block diagonal as
 1003 $\text{avar}(\sqrt{n}\widehat{\boldsymbol{\alpha}}_n) = \mathbf{V}(\boldsymbol{\alpha}_t) = \mathbf{F}^{-1}(\boldsymbol{\alpha}_t)$, $\text{avar}(\sqrt{n}\widehat{\boldsymbol{\beta}}_n) = \mathbf{V}(\boldsymbol{\beta}_t) = \mathbf{F}^{-1}(\boldsymbol{\beta}_t)$ and $\text{avar}[\sqrt{n}\text{vech}(\widehat{\boldsymbol{\Sigma}}_{\mathbf{X}})] =$
 1004 $\mathbf{F}_{\boldsymbol{\Sigma}_{\mathbf{X}}}^{-1}$, where the Fisher information matrix $\mathbf{F}_{\boldsymbol{\Sigma}_{\mathbf{X}}} = \mathbf{E}_p^T (\boldsymbol{\Sigma}_{\mathbf{X}}^{-1} \otimes \boldsymbol{\Sigma}_{\mathbf{X}}^{-1}) \mathbf{E}_p / 2$. (Cook et al. 2013).
 1005 Hence,

$$\begin{aligned} \mathbf{F}(\mathbf{h}_t) &= \mathbf{E} \left\{ \left(\frac{\partial L_n(\mathbf{h})}{\partial \mathbf{h}} \right)_{\mathbf{h}=\mathbf{h}_t} \left(\frac{\partial L_n(\mathbf{h})}{\partial \mathbf{h}^T} \right)_{\mathbf{h}=\mathbf{h}_t} \right\} = -\mathbf{E} \left\{ \left(\frac{\partial^2 L_n(\mathbf{h})}{\partial \mathbf{h} \partial \mathbf{h}^T} \right)_{\mathbf{h}=\mathbf{h}_t} \right\} \\ &= \text{diag}(\mathbf{F}(\boldsymbol{\alpha}_t), \mathbf{F}(\boldsymbol{\beta}_t), \mathbf{F}_{\boldsymbol{\Sigma}_{\mathbf{X}}}). \end{aligned}$$

1006 Following Cook et al. (2013), the asymptotic variance for $\widehat{\mathbf{h}}_{\text{env}} = \mathbf{h}(\widehat{\boldsymbol{\phi}}_{\text{env}})$ can be expressed
 1007 as $\text{avar}(\widehat{\mathbf{h}}_{\text{env}}) = \mathbf{H}(\mathbf{H}^T \mathbf{F}(\mathbf{h}_t) \mathbf{H})^\dagger \mathbf{H}^T$, where $\mathbf{H} = \partial \mathbf{h} / \partial \boldsymbol{\xi} = (\partial \mathbf{h}_i / \partial \boldsymbol{\xi}_j)_{i=1, \dots, 3, j=1, \dots, 5}$ is the
 1008 gradient matrix and $\dagger$ denotes the Moore-Penrose generalized inverse. By direct computation,

we found

$$\mathbf{H} = \begin{pmatrix} \mathbf{I}_{m-p} & 0 & 0 & 0 & 0 \\ 0 & \mathbf{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{I}_p & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_p(\mathbf{\Gamma}\mathbf{\Omega} \otimes \mathbf{I}_p - \mathbf{\Gamma} \otimes \mathbf{\Gamma}_0\mathbf{\Omega}_0\mathbf{\Gamma}_0^T) & \mathbf{C}_p(\mathbf{\Gamma} \otimes \mathbf{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\mathbf{\Gamma}_0 \otimes \mathbf{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}. \quad (\text{E.6})$$

Since $\text{avar}(\hat{\mathbf{h}}_{\text{env}}) = \mathbf{H}(\mathbf{H}^T \mathbf{F}(\mathbf{h}_t) \mathbf{H})^\dagger \mathbf{H}^T$ depends only on $\text{span}(\mathbf{H})$, we transform $\mathbf{H} \rightarrow \mathbf{H}_1$ similar to Cook et al. (2013) so that $\text{span}(\mathbf{H}) = \text{span}(\mathbf{H}_1)$ and

$$\text{avar}(\sqrt{n}\hat{\mathbf{h}}_{\text{env}}) = \mathbf{H}_1(\mathbf{H}_1^T \mathbf{F}(\mathbf{h}_t) \mathbf{H}_1)^\dagger \mathbf{H}_1^T = \sum_{j=1}^5 \mathbf{H}_{1j}(\mathbf{H}_{1j}^T \mathbf{F}(\mathbf{h}_t) \mathbf{H}_{1j})^\dagger \mathbf{H}_{1j}^T, \quad (\text{E.7})$$

where $\mathbf{H}_1 = (\mathbf{H}_{11}, \dots, \mathbf{H}_{15})$ has the blockwise form of

$$\begin{aligned} \mathbf{H}_1 &= \begin{pmatrix} \mathbf{H}_{11} & \mathbf{H}_{12} & \mathbf{H}_{13} & \mathbf{H}_{14} & \mathbf{H}_{15} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{I}_{m-p} & 0 & 0 & 0 & 0 \\ 0 & \mathbf{\Gamma} & \boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_0 & 0 & 0 \\ 0 & 0 & 2\mathbf{C}_p(\mathbf{\Gamma}\mathbf{\Omega} \otimes \mathbf{\Gamma}_0 - \mathbf{\Gamma} \otimes \mathbf{\Gamma}_0\mathbf{\Omega}_0) & \mathbf{C}_p(\mathbf{\Gamma} \otimes \mathbf{\Gamma})\mathbf{E}_u & \mathbf{C}_p(\mathbf{\Gamma}_0 \otimes \mathbf{\Gamma}_0)\mathbf{E}_{p-u} \end{pmatrix}. \end{aligned}$$

Because of the zeros blocks in $\mathbf{H}_{11}$, $\mathbf{H}_{14}$ and $\mathbf{H}_{15}$, they have no contribution to the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\text{env}}$, which is the middle $p \times p$ block of $\text{avar}(\sqrt{n}\hat{\mathbf{h}}_{\text{env}})$:

$$\text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_{\text{env}}) = \mathbf{\Gamma}(\mathbf{H}_{12}^T \mathbf{F}(\mathbf{h}_t) \mathbf{H}_{12})^\dagger \mathbf{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_0)(\mathbf{H}_{13}^T \mathbf{F}(\mathbf{h}_t) \mathbf{H}_{13})^\dagger (\boldsymbol{\eta} \otimes \mathbf{\Gamma}_0^T). \quad (\text{E.8})$$

Then by straightforward calculations similar to Cook et al. (2010), we have

$$\text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_{\text{env}}) = \mathbf{\Gamma} \{ \mathbf{\Gamma}^T \mathbf{F}(\boldsymbol{\beta}_t) \mathbf{\Gamma} \}^{-1} \mathbf{\Gamma}^T + (\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_0) \mathbf{M}^{-1} (\boldsymbol{\eta} \otimes \mathbf{\Gamma}_0^T),$$

where $\mathbf{F}(\boldsymbol{\beta}_t) = \{ \text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_n) \}^{-1}$ and $\mathbf{M} = (\boldsymbol{\eta} \otimes \mathbf{\Gamma}_0^T) \mathbf{F}(\boldsymbol{\beta}_t) (\boldsymbol{\eta}^T \otimes \mathbf{\Gamma}_0) + \mathbf{\Omega} \otimes \mathbf{\Omega}_0^{-1} + \mathbf{\Omega}^{-1} \otimes \mathbf{\Omega}_0 - 2\mathbf{I}_{u(p-u)}$.

From Proposition 5, we recognize the first term in $\text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_{\text{env}})$ is $\text{avar}(\sqrt{n}\hat{\boldsymbol{\beta}}_{\mathbf{\Gamma}})$ which equals to $\text{avar}(\sqrt{n}\mathbf{\Gamma}\hat{\boldsymbol{\eta}}_{\mathbf{\Gamma}})$. To further interpret the second term, we follow the same calculation as in Cook et al. (2010) by making the following substitutes: $\boldsymbol{\Sigma} \rightarrow \boldsymbol{\Sigma}_{\mathbf{X}}$, $\boldsymbol{\Sigma}_{\mathbf{X}} \otimes \boldsymbol{\Sigma}^{-1} \rightarrow \mathbf{F}(\boldsymbol{\beta}_t)$ and the dimension $r \rightarrow p$. Then the conclusion follows.

□

E.6 Corollary 1

Proof. From Proposition 2, we can see $\mathcal{E}_{\mathbf{V}_{\beta\beta}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\boldsymbol{\Sigma}_{\mathbf{X}}}(\boldsymbol{\beta}_t) = \mathcal{E}_{\sigma^2\mathbf{I}_p}(\boldsymbol{\beta}_t) = \text{span}(\boldsymbol{\beta}_t)$. Therefore, the envelope is one-dimensional and $\boldsymbol{\beta}_t = \mathbf{\Gamma}\boldsymbol{\eta}$ with $\mathbf{\Gamma} = \boldsymbol{\beta}_t / \|\boldsymbol{\beta}_t\| \in \mathbb{R}^{p \times 1}$ and

1026 $\eta = \|\beta_t\| \in \mathbb{R}^1$. The covariances $\Omega = \sigma^2$ and $\Omega_0 = \sigma^2 \mathbf{I}_{p-1}$. The matrix $\mathbf{M}$ in Proposition 4
 1027 becomes $\mathbf{M} = (\eta \Gamma_0^T) \mathbf{V}_{\beta\beta}^{-1} (\eta \Gamma_0)$ because $\Omega^{-1} \otimes \Omega_0 = \Omega \otimes \Omega_0^{-1} = \mathbf{I}_{p-1}$ cancel with the last
 1028 term. Hence,

$$\begin{aligned}
 \text{avar}(\sqrt{n} \widehat{\beta}_{\text{env}}) &= \mathbf{P}_\Gamma \mathbf{V}_{\beta\beta} \mathbf{P}_\Gamma + (\eta \Gamma_0) \mathbf{M}^{-1} (\eta \Gamma_0^T) \\
 &= \mathbf{P}_\Gamma \mathbf{V}_{\beta\beta} \mathbf{P}_\Gamma + \eta^2 \Gamma_0 (\eta^2 \Gamma_0^T \mathbf{V}_{\beta\beta}^{-1} \Gamma_0)^{-1} \Gamma_0^T \\
 &= \mathbf{P}_\Gamma \mathbf{V}_{\beta\beta} \mathbf{P}_\Gamma + \mathbf{Q}_\Gamma \mathbf{V}_{\beta\beta} \mathbf{Q}_\Gamma \\
 &= \mathbf{V}_{\beta\beta} = \text{avar}(\sqrt{n} \widehat{\beta}),
 \end{aligned}$$

1029 where the last two equalities have used the fact that Γ reduces $\mathbf{V}_{\beta\beta}$. □

1030 E.7 Proposition 7

1031 *Proof.* Let $(\mathbf{R} \otimes \mathbf{L})_0 := (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})$ is a $pr \times (pr - d_X d_Y)$ semi-orthogonal
 1032 matrix so that $\mathbf{O} := (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L}, \mathbf{R} \otimes \mathbf{L})$ is an orthogonal basis for $\mathbb{R}^{pr}$, where
 1033 $(\mathbf{R}, \mathbf{R}_0)$ and $(\mathbf{L}, \mathbf{L}_0)$ are orthogonal bases for $\mathbb{R}^p$ and $\mathbb{R}^r$.

1034 We first prove that the following statements are equivalent.

- 1035 1. $\text{span}(\mathbf{R} \otimes \mathbf{L})$ reduce $\Delta_R \otimes \Delta_L$.
- 1036 2. $\text{span}(\mathbf{R})$ reduces Δ_R and $\text{span}(\mathbf{L})$ reduces Δ_L .
- 1037 3. $\mathbf{O}(\Delta_R \otimes \Delta_L) \mathbf{O}^T = \text{diag}\{\mathbf{M}_1, \mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$, where $\mathbf{M}_1 = \mathbf{R}^T \Delta_R \mathbf{R} \otimes \mathbf{L}^T \Delta_L \mathbf{L}$, $\mathbf{M}_2 =$
 1038 $\mathbf{R}_0^T \Delta_R \mathbf{R}_0^T \otimes \mathbf{L}_0 \Delta_L \mathbf{L}_0^T$, $\mathbf{M}_3 = \mathbf{R}^T \Delta_R \mathbf{R}^T \otimes \mathbf{L}_0 \Delta_L \mathbf{L}_0^T$ and $\mathbf{M}_4 = \mathbf{R}_0^T \Delta_R \mathbf{R}_0^T \otimes \mathbf{L} \Delta_L \mathbf{L}^T$.
- 1039 4. $\mathbf{O}(\Delta_R \otimes \Delta_L) \mathbf{O}^T = \text{diag}\{\mathbf{M}_1, \mathbf{M}_0\}$, where $\mathbf{M}_0 = (\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})(\Delta_R \otimes$
 1040 $\Delta_L)(\mathbf{R}_0 \otimes \mathbf{L}_0, \mathbf{R} \otimes \mathbf{L}_0, \mathbf{R}_0 \otimes \mathbf{L})^T$.

1041 By definition, we know that Statement 1 is equivalent to 4 and that 2 is equivalent to 3.
 1042 Also, 3 implies 4. We then only need to show that 4 implies 3. From 4, we know that the
 1043 off-diagonal blocks of $\mathbf{O}(\Delta_R \otimes \Delta_L) \mathbf{O}^T$ are all zero matrices. More explicitly,

$$\begin{aligned}
 0 &= \mathbf{R}^T \Delta_R \mathbf{R} \otimes \mathbf{L}^T \Delta_L \mathbf{L}_0 = \mathbf{R}^T \Delta_R \mathbf{R} \otimes \mathbf{L}_0^T \Delta_L \mathbf{L} \\
 &= \mathbf{R}^T \Delta_R \mathbf{R}_0 \otimes \mathbf{L}^T \Delta_L \mathbf{L} = \mathbf{R}_0^T \Delta_R \mathbf{R} \otimes \mathbf{L}^T \Delta_L \mathbf{L} \\
 &= \mathbf{R}^T \Delta_R \mathbf{R}_0 \otimes \mathbf{L}^T \Delta_L \mathbf{L}_0 = \mathbf{R}_0^T \Delta_R \mathbf{R} \otimes \mathbf{L}^T \Delta_L \mathbf{L}_0.
 \end{aligned} \tag{E.9}$$

Without loss of generality, we assume that $\mathbf{R}^T \Delta_R \mathbf{R}$, $\mathbf{L}^T \Delta_L \mathbf{L}$, $\mathbf{R}_0^T \Delta_R \mathbf{R}_0$ and $\mathbf{L}_0^T \Delta_L \mathbf{L}_0$ are all nonzero. Otherwise, the problem degenerates to the case of reducing either Δ_R or Δ_L and the proof would be trivial. Because $\mathbf{R}^T \Delta_R \mathbf{R}$, $\mathbf{L}^T \Delta_L \mathbf{L}$, $\mathbf{R}_0^T \Delta_R \mathbf{R}_0$ and $\mathbf{L}_0^T \Delta_L \mathbf{L}_0$ are all nonzero, we must have from (E.9) that

$$0 = \mathbf{L}^T \Delta_L \mathbf{L}_0 = \mathbf{L}_0^T \Delta_L \mathbf{L} = \mathbf{R}_0^T \Delta_R \mathbf{R} = \mathbf{R}^T \Delta_R \mathbf{R}_0,$$

which implies the following blocks within $\mathbf{M}_0$ are all zero matrices:

$$\begin{aligned} 0 &= \mathbf{R}_0^T \Delta_R \mathbf{R}_0 \otimes \mathbf{L}^T \Delta_L \mathbf{L}_0 = \mathbf{R}_0^T \Delta_R \mathbf{R}_0 \otimes \mathbf{L}_0^T \Delta_L \mathbf{L} \\ &= \mathbf{R}^T \Delta_R \mathbf{R}_0 \otimes \mathbf{L}_0^T \Delta_L \mathbf{L}_0 = \mathbf{R}_0^T \Delta_R \mathbf{R} \otimes \mathbf{L}_0^T \Delta_L \mathbf{L}_0 \\ &= \mathbf{R}_0^T \Delta_R \mathbf{R} \otimes \mathbf{L}^T \Delta_L \mathbf{L}_0 = \mathbf{R}^T \Delta_R \mathbf{R}_0 \otimes \mathbf{L}^T \Delta_L \mathbf{L}_0. \end{aligned}$$

Hence, $\mathbf{M}_0 = \text{diag}\{\mathbf{M}_2, \mathbf{M}_3, \mathbf{M}_4\}$, which leads to the equivalence of the four statements.

From Definition 3, Definition 4 and the equivalence between statement 1 and 2, we reach the conclusion of Proposition 7.

□

References

- [1] COOK, R. D. AND FORZANI, L. (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association*, **104**, 197–208.
- [2] COX, D.R. (1972), Regression models and life-tables (with Discussion), *J. R. Statist. Soc. B*, **34**, 187–220.
- [3] COX, D.R. (1975), Partial likelihood, *Biometrika*, **62**, 269–276.
- [4] DAWID, A.P. (1981), Some matrix-variate distribution theory: Notational considerations and a Bayesian application, *Biometrika*, **68**, 265–274.
- [5] EATON, M.L. (1986), A characterization of spherical distributions, *Journal of Multivariate Analysis*, **20**, 272–276.
- [6] HUNG, H., WU, P. S., TU, I. P. AND HUNG, S. Y. (2012). On multilinear principal component analysis of order-two tensors. *Biometrika*, **99**, 569–583.

- 1065 [7] LI, B., KIM, M.K. AND ALTMAN, M. (2010). On dimension folding of matrix
1066 or array valued statistical objects. *Annals of Statistics*, **38**, 1097–1121.
- 1067 [8] SEBER, G.A.F. (2008), A matrix handbook for statisticians, *Wiley-Interscience*.