SUFFICIENT DIMENSION REDUCTION FOR LONGITUDINAL DATA

Xuan Bi and Annie Qu

Department of Statistics, University of Illinois at Urbana-Champaign

Supplementary Material

S1 Proof of Theorem 1

Suppose $\mathcal{S}_\zeta$ is an arbitrary dimension-reduction subspace. Define $\mathbf{B}_\zeta$ as the basis matrix of $\mathcal{S}_\zeta$, and $\mathbf{P}_{B_\zeta} = \mathbf{B}_\zeta(\mathbf{B}_\zeta'\mathbf{B}_\zeta)^{-1}\mathbf{B}_\zeta'$ is the projection matrix. Note that the population version of QIF is $Q(\mathbf{b}) = (\mathrm{E}\mathbf{g})'\mathbf{W}^{-1}(\mathrm{E}\mathbf{g})$, where $\mathbf{g}$ is a (mp)-dimensional estimating function.

We first show that $Q(\mathbf{b}) \geq Q(\mathbf{P}_{B_\zeta}\mathbf{b})$ for any p-dimensional parameter $\mathbf{b} \in \mathbb{R}^p$. This implies that the minimizer of $Q(\mathbf{b})$, denoted as $\boldsymbol{\gamma}$, must lie in $\mathcal{S}_\zeta$, and thus lie in the central subspace $\mathcal{S}_{Y|\mathbf{X}} = \cap_\zeta \mathcal{S}_\zeta$. This is similar to the argument of Theorem 2.1 in Li and Duan (1989), and Proposition 8.1 in Cook (1998, p.144).

Since $Q(\mathbf{b}) = \{\mathrm{E}(\mathbf{W}^{-\frac{1}{2}}\mathbf{g})\}'\{\mathrm{E}(\mathbf{W}^{-\frac{1}{2}}\mathbf{g})\}$, we define $\mathbf{g}^* = \mathbf{W}^{-\frac{1}{2}}\mathbf{g}$. Then,

$$\mathrm{Var}(\mathbf{g}^*) = \mathbf{I}_{mp} = \mathrm{E}(\mathbf{g}^*\mathbf{g}^{*\prime}) - (\mathrm{E}\mathbf{g}^*)(\mathrm{E}\mathbf{g}^*)'.$$

Therefore,

$$\begin{aligned} Q(\mathbf{b}) &= (\mathrm{E}\mathbf{g}^*)'(\mathrm{E}\mathbf{g}^*) = \mathrm{tr}\{(\mathrm{E}\mathbf{g}^*)'(\mathrm{E}\mathbf{g}^*)\} \\ &= \mathrm{tr}\{(\mathrm{E}\mathbf{g}^*)(\mathrm{E}\mathbf{g}^*)'\} = \mathrm{tr}\{\mathrm{E}(\mathbf{g}^*\mathbf{g}^{*\prime}) - \mathbf{I}_{mp}\} \\ &= \mathrm{E}\{\mathrm{tr}(\mathbf{g}^*\mathbf{g}^{*\prime}) - mp\} = \mathrm{E}(\mathbf{g}^{*\prime}\mathbf{g}^*) - mp \\ &= \mathrm{E}[\mathrm{E}\{\mathbf{g}^{*\prime}(\mathbf{b}'\mathbf{X},\mathbf{Y})\mathbf{g}^*(\mathbf{b}'\mathbf{X},\mathbf{Y})\}|\mathbf{Y},\mathbf{B}_\zeta'\mathbf{X}] - mp. \end{aligned}$$

Note that $L(\mathbf{b}'\mathbf{X},\mathbf{Y}) = \mathbf{g}^{*\prime}(\mathbf{b}'\mathbf{X},\mathbf{Y})\mathbf{g}^*(\mathbf{b}'\mathbf{X},\mathbf{Y})$ is convex with respect to its first argument. Therefore,

$$\begin{aligned} Q(\mathbf{b}) &= \mathrm{E}[\mathrm{E}\{L(\mathbf{b}'\mathbf{X},\mathbf{Y})\}|\mathbf{Y},\mathbf{B}_\zeta'\mathbf{X}] - mp \\ &\geq \mathrm{E}\{L(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{Y},\mathbf{B}_\zeta'\mathbf{X}),\mathbf{Y})\} - mp \\ &= \mathrm{E}\{\mathbf{g}^{*\prime}(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{Y},\mathbf{B}_\zeta'\mathbf{X}),\mathbf{Y})\mathbf{g}^*(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{Y},\mathbf{B}_\zeta'\mathbf{X}),\mathbf{Y})\} - mp. \end{aligned}$$

Because $\mathbf{B}_\zeta$ is the basis matrix of $\mathcal{S}_\zeta$, we have $\mathbf{X}|(\mathbf{Y},\mathbf{B}_\zeta'\mathbf{X}) \overset{\mathrm{d}}{=} \mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}$; and the linearity condition implies that $\mathrm{E}(\mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}) = \mathbf{P}_{B_\zeta}\mathbf{X}$. Hence,

$$\begin{aligned} Q(\mathbf{b}) &\geq \mathrm{E}\{\mathbf{g}^{*\prime}(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}), \mathbf{Y})\mathbf{g}^*(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}), \mathbf{Y})\} - mp \\ &= \mathrm{E}\{\mathbf{g}^{*\prime}((\mathbf{P}_{B_\zeta}\mathbf{b})'\mathbf{X}, \mathbf{Y})\mathbf{g}^*((\mathbf{P}_{B_\zeta}\mathbf{b})'\mathbf{X}, \mathbf{Y})\} - mp \\ &= Q(\mathbf{P}_{B_\zeta}\mathbf{b}). \end{aligned}$$

Next, we show that $\hat{\boldsymbol{\gamma}}$ is a strongly consistent estimator of $\boldsymbol{\gamma}$. This follows Theorem 5.1 of Li and Duan (1989), which states that the minimizer of the sample loss function converges to the minimizer of the risk function almost surely, if the objective loss function is convex with respect to its first arument. ■

S2 Proof of Corollary 1(transformation)

Following Cook (1998, p.115), if $\mathbf{h}$ is a function of $\mathbf{Y}$, then $\mathcal{S}_{\mathbf{h}(Y)|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$; and if $\mathbf{h}$ is one-to-one, then $\mathcal{S}_{\mathbf{h}(Y)|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$. Then Corollary 1 follows immediately from Theorem 1. ■

S3 Proof of Lemma 1

The first part of this proof shows that we gain more information and achieve higher efficiency by incorporating additional correlation information formulated by the moment condition $\mathbf{G}_2$.

We first orthogonalize $\mathbf{G}_2$ from $\mathbf{G}_1$ as

$$\mathbf{G}_2^* = \mathbf{G}_2 - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{G}_1,$$

where $\mathbf{C}_{21} = \mathrm{Cov}(\mathbf{G}_2, \mathbf{G}_1)$ and $\mathbf{C}_{11} = \mathrm{Var}(\mathbf{G}_1)$. After orthogonalization, $\mathrm{Cov}(\mathbf{G}_2^*, \mathbf{G}_1) = \mathbf{0}$. Let $\mathbf{G}^* = (\mathbf{G}_1', \mathbf{G}_2^{*\prime})'$, $\mathbf{C}^* = \mathrm{Var}(\mathbf{G}^*)$, and $\mathbf{C}_{22}^* = \mathrm{Var}(\mathbf{G}_2^*)$, then $\mathbf{C}_{22}^* = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$, where $\mathbf{C}_{22} = \mathrm{Var}(\mathbf{G}_2)$ and $\mathbf{C}_{12} = \mathrm{Cov}(\mathbf{G}_1, \mathbf{G}_2)$. Since $\dot{\mathbf{G}}_2^* = \dot{\mathbf{G}}_2 - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1$, the information matrix of the estimator by minimizing $\mathbf{G}'\mathbf{C}^{-1}\mathbf{G}$ is proportional to

$$\begin{aligned} \dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}} &= (\dot{\mathbf{G}}^*)'(\mathbf{C}^*)^{-1}(\dot{\mathbf{G}}^*) \\ &= \dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1 + (\dot{\mathbf{G}}_2^*)'(\mathbf{C}_{22}^*)^{-1}(\dot{\mathbf{G}}_2^*). \end{aligned}$$

Note that $\mathbf{C}_{22}^*$ is non-negative definite, so in the sense of Loewner ordering for matrices,

$$\dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}} \geq \dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1.$$

The following argument shows that if $\mathbf{G}_1$ contains all information about the parameter, adding additional moment conditions will not improve efficiency. That is, if $\mathbf{M}_1$ is proportional to $\mathbf{R}^{-1}$, then $\dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}} = \dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1$.

The detailed proof is provided as follows. Recall that $\mathbf{G}_l = \sum_{i=1}^n \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_l \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i)$, $l = 1, 2$. Assume $\mathbf{R}^{-1} = a_1 \mathbf{M}_1$, then

$$\dot{\mathbf{G}}_1 = -\frac{1}{a_1} \sum_{i=1}^n \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{R}^{-1} \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i + o_p(1), \text{and } \dot{\mathbf{G}}_2 = -\sum_{i=1}^n \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_2 \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i + o_p(1).$$

In addition,

$$\mathbf{C}_{11} = \frac{1}{a_1^2} \sum_{i=1}^n \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{R}^{-1} \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i, \text{and } \mathbf{C}_{21} = \frac{1}{a_1} \sum_{i=1}^n \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_2 \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i.$$

Thus, $\mathbf{C}_{11} = -\frac{1}{a_1} \dot{\mathbf{G}}_1 + o_p(1)$ and $\mathbf{C}_{21} = -\frac{1}{a_1} \dot{\mathbf{G}}_2 + o_p(1)$, and this results in $\dot{\mathbf{G}}_2^* = o_p(1)$. Therefore,

$$\dot{\mathbf{G}}' \mathbf{C}^{-1} \dot{\mathbf{G}} = \dot{\mathbf{G}}_1' \mathbf{C}_{11}^{-1} \dot{\mathbf{G}}_1 + o_p(1).$$

■

S4 Proof of Theorem 2

Theorem 18.11 of Kosorok (2008, p.341) shows that the marginal efficiency of two estimators leads to their joint efficiency on product spaces, given the condition that the two estimated parameters are differentiable with respect to their tangent space. The main goal of this proof is to verify this condition under the sufficient dimension reduction framework for longitudinal data, and thus the estimators by the proposed method have joint efficiency, leading to the efficiency of the central subspace. The definition of tangent space and differentiability with respect to the tangent space are provided in the following two paragraphs respectively.

Without loss of generosity, we assume $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d$ are linearly independent. Then $\boldsymbol{\gamma}_j \in \mathcal{S}_{Y|\mathbf{X}}$ implies $\text{Span}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d) = \mathcal{S}_{Y|\mathbf{X}}$, $j = 1, \ldots, d$. Set $s = d$ and $\mathbf{B}^* = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$. Suppose $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ is a $p \times d$ constant matrix in $\mathbb{R}^{p \times d}$, and let $\text{vec}(\mathbf{u}) = (\mathbf{u}_1', \ldots, \mathbf{u}_d')'$ denote the vectorization of $\mathbf{u}$. Suppose the transformed response $h_j(y_{it})$ is imposed for score function $\mathbf{S}_j$ such that the solution $\hat{\boldsymbol{\gamma}}_j$ of $\mathbf{S}_j = \mathbf{0}$ is an efficient estimator of $\boldsymbol{\gamma}_j, j = 1, \ldots, d$. Let $\mathbf{S} = (\mathbf{S}_1', \ldots, \mathbf{S}_d')'$. Define the tangent function to be $H = \mathbf{S}' \text{vec}(\mathbf{u})$. Then a tangent set is $\mathcal{T} = \{H = \mathbf{S}'\text{vec}(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^{p \times d}\}$. Since this tangent set is closed under linear combination, it is also a tangent space.

For an arbitrarily small $\delta \geq 0$ and fixed $\text{vec}(\mathbf{B}) = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d')'$, suppose the model has a true parameter $\text{vec}(\mathbf{B}) + \delta \text{vec}(\mathbf{u})$. A parameter $\boldsymbol{\gamma}$ is differentiable with respect to the tangent space $\mathcal{T}$, if $d\boldsymbol{\gamma}/d\delta|_{\delta=0} = \dot{\boldsymbol{\psi}}(H)$, where $\dot{\boldsymbol{\psi}}(\cdot)$ is a bounded linear operator.

Since $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ and $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$, there exists a $d \times d$ matrix $\mathbf{D}$, such that $\mathbf{B}^* = \mathbf{B}\mathbf{D}$. Since the $pd \times pd$ information matrix of $\text{vec}(\mathbf{B}) = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d')'$ is bounded, the information matrix $\dot{\mathbf{S}}' \widetilde{\mathbf{C}}^{-1} \dot{\mathbf{S}}$ of $(\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_d')'$ is also bounded, where $\widetilde{\mathbf{C}} = \text{Var}(\mathbf{S})$.

For any direction $\mathbf{u} \in \mathbb{R}^{p\times d}$ and an arbitrarily small $\delta \geq 0$,

$$\begin{aligned}
\frac{d\boldsymbol{\gamma}_j}{d\delta} &= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}+\delta\mathbf{u})}\text{vec}(\mathbf{u}) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{BD}+\delta\mathbf{uD})}\frac{d\text{vec}(\mathbf{BD}+\delta\mathbf{uD})}{d\text{vec}(\mathbf{B}+\delta\mathbf{u})}\text{vec}(\mathbf{u}) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{BD}+\delta\mathbf{uD})}(\mathbf{D}'\otimes\mathbf{I}_p)\text{vec}(\mathbf{u}) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)}(\mathbf{D}'\otimes\mathbf{I}_p)\text{vec}(\mathbf{u}),
\end{aligned}$$

where $\mathbf{u}^* = \mathbf{uD}$, $j = 1,\ldots,d$ and $\otimes$ denotes the Kronecker product.

Similar to Lemma 1, we can show that $\widetilde{\mathbf{C}} = -\dot{\mathbf{S}} + o_p(1)$. And $\text{E}(\mathbf{S}) = \mathbf{0}$ implies $-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}} = \text{E}(\mathbf{SS}') + o_p(1)$. Therefore,

$$\begin{aligned}
\frac{d\boldsymbol{\gamma}_j}{d\delta} &= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)}(\mathbf{D}'\otimes\mathbf{I}_p)(-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}})^{-1}\{\text{E}(\mathbf{SS}')\}\text{vec}(\mathbf{u}) + o_p(1) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)}(\mathbf{D}'\otimes\mathbf{I}_p)(-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}})^{-1}\{\text{E}(\mathbf{S}H)\} + o_p(1).
\end{aligned}$$

Define $\dot{\boldsymbol{\psi}}_j(H) = d\boldsymbol{\gamma}_j/d\delta|_{\delta=0}$ for any tangent function $H \in \mathcal{T}$. Since $\boldsymbol{\gamma}_j$ is the j-th column of $\mathbf{B}^*$, $d\boldsymbol{\gamma}_j/d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)|_{\delta=0}$ is bounded. Because $\mathbf{D}$ is a bounded linear transformation and $(-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}})$ is also bounded, it follows that $\dot{\boldsymbol{\psi}}_j(\cdot)$ is a bounded linear operator. Therefore, $\boldsymbol{\gamma}_j$ is differentiable with respect to the tangent space $\mathcal{T}$, $j = 1,\ldots,d$.

Following Theorem 18.11 of Kosorok (2008, p.341), we conclude that $(\hat{\boldsymbol{\gamma}}_1,\ldots,\hat{\boldsymbol{\gamma}}_d)$ is an asymptotic efficient estimator of $(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_d)$. ■

References

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics.* Wiley, New York.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference.* Springer, New York.

Li, K. C. and Duan, N. (1989). Regression analysis under link violation. *Ann. Statist.* **17**, 1009-1052.