Supplement to "Sufficient dimension reduction and prediction through cumulative slicing PFC"

Xinyi Xu †, Xiangjie Li†,Jingxiao Zhang†*

†Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing 100872, China.

January 12, 2018

lemma 1. *Under the normal inverse model (1) in the article, let* $R(\mathbf{X}) = \mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{X}$. *Then* $R(\mathbf{X})$ *is the minimal sufficient linear reduction.*

The detailed proof of lemma 1 has been given by Cook and Forzani (2008)[2]. The goal consequently turns to estimate $\mathbf{\Delta}^{-1}\mathcal{S}_{\mathbf{\Gamma}} = \{\mathbf{\Delta}^{-1}\mathbf{z} : \mathbf{z} \in \mathcal{S}_{\mathbf{\Gamma}}\}$ under the CUPFC model.

The proof of Proposition 3.1:

Proof. Under the CUPFC model the full parameter space is $(\boldsymbol{\mu}, \mathcal{S}_{\mathbf{\Gamma}}, \boldsymbol{\beta}, \mathbf{\Delta})$. When we derive the MLE of these parameters we set d fixed and the selection of d deserves separate discussion.

Given a specific $\tilde{y} \in \mathbb{R}$ as the parameter in the model for $\mathbf{X}_y$, we have the conditional model

$$\mathbf{X}_y = \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\beta}_{\tilde{y}}\{I(y \le \tilde{y}) - Pr(Y \le \tilde{y})\} + \boldsymbol{\varepsilon}.$$

*Corresponding author. Email: zhjxiaoruc@163.com

Then for specific $y \in S_Y$, use the centered $f_{y;\tilde{y}}$ to stand for $I(y \leq \tilde{y}) - Pr(Y \leq \tilde{y})$ and $\mathbf{X}_y$ is presented as:

$$\mathbf{X}_y = \boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}} + \boldsymbol{\varepsilon}.$$

Given a group of observed response $S = (y_1, y_2, ..., y_n)$, the joint probability density function of $\mathbf{X}_y$, $y \in S$ is:

$$\begin{aligned} g(\mathbf{X}_y : y \in S) =& (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Delta}|^{-\frac{n}{2}} \exp\Big\{ -\frac{1}{2} \sum_y \Big(\mathbf{X}_y - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\Big)^T \\ & \cdot \boldsymbol{\Delta}^{-1} \Big(\mathbf{X}_y - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\Big)\Big\}, \end{aligned}$$

as $\mathbf{X}_y$ for different $y_1, ..., y_n$ are independent but not identically distributed.

The full log likelihood for $\mathbf{X}_y$ is

$$\begin{aligned} L_{\tilde{y}}(\boldsymbol{\mu}, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\beta}, \boldsymbol{\Delta}) =& -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Delta}| \\ & - \frac{1}{2}\sum_y \Big(\mathbf{X}_y - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\Big)^T \boldsymbol{\Delta}^{-1} \Big(\mathbf{X}_y - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\Big). \end{aligned} \tag{1}$$

For fixed $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$, equation (1) is maximized over $\boldsymbol{\mu}$ by $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$. Brought in $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$, note a conversion technique that

$$\begin{aligned} & \sum_y \Big(\mathbf{X}_y - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\Big)^T \boldsymbol{\Delta}^{-1} \Big(\mathbf{X}_y - \boldsymbol{\mu} - \boldsymbol{\Gamma}\boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\Big) \\ & = \mathrm{trace}\Big\{\boldsymbol{\Delta}^{-1/2}\Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\boldsymbol{\beta}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big)^T \Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\boldsymbol{\beta}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big)\boldsymbol{\Delta}^{-1/2}\Big\}, \end{aligned}$$

where $\mathbb{X}$ is the $n \times p$ matrix with rows $(\mathbf{X}_{y_i} - \bar{\mathbf{X}})^T$ which is $(\mathbf{X}_i - \bar{\mathbf{X}})^T$ actually, $\mathbb{F}_{\tilde{y}}$ is an $n \times 1$ matrix with the kth element $f_{y_k;\tilde{y}}$ $(k = 1, ..., n)$ and $\boldsymbol{\beta}_{\tilde{y}}$ is a $d \times 1$ matrix whose elements only depend on the specific $\tilde{y}$.

Then we have

$$\begin{aligned} L_{\tilde{y}}(\boldsymbol{\mu}, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\beta}, \boldsymbol{\Delta}) =& -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Delta}| \\ & - \frac{1}{2}\mathrm{trace}\Big\{\boldsymbol{\Delta}^{-1/2}\Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\boldsymbol{\beta}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big)^T \Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\boldsymbol{\beta}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big)\boldsymbol{\Delta}^{-1/2}\Big\}. \end{aligned} \tag{2}$$

For fixed $\boldsymbol{\Delta}$ and $\boldsymbol{\Gamma}$, equation (2) is maximized over $\boldsymbol{\beta}_{\tilde{y}}$ by $\hat{\boldsymbol{\beta}}_{\tilde{y}} = \boldsymbol{\Gamma}^T \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Delta}^{-1})} \hat{\mathbf{B}}_{\tilde{y}}$, where $\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Delta}^{-1})} = \boldsymbol{\Gamma}(\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}^T\boldsymbol{\Delta}^{-1}$ is the projection onto $\mathcal{S}_{\boldsymbol{\Gamma}}$ in the $\boldsymbol{\Delta}^{-1}$ inner product and $\hat{\mathbf{B}}_{\tilde{y}} = \mathbb{X}^T\mathbb{F}_{\tilde{y}}(\mathbb{F}_{\tilde{y}}^T\mathbb{F}_{\tilde{y}})^{-1}$ with $\mathbb{F}_{\tilde{y}}$'s kth coordinate being $f_{y_k;\tilde{y}}$. $\hat{\mathbf{B}}_{\tilde{y}}$ is obviously the coefficient matrix from the multivariate OLS regression of $\mathbf{X}$ on $f_{\tilde{y}}$ (Cook & Forzani 2008)[2]. Pay attention that we usually set $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d}$ an orthonormal basis of $\mathcal{S}_{\boldsymbol{\Gamma}}$ without loss of generality, so the MLE $\boldsymbol{\Gamma}\hat{\boldsymbol{\beta}}_{\tilde{y}}$ will be $\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Delta}^{-1})}\hat{\mathbf{B}}_{\tilde{y}}$. We then substitute $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\beta}}_{\tilde{y}}$ into the log likelihood $L_{\tilde{y}}(\boldsymbol{\mu}, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\beta}, \boldsymbol{\Delta})$ to attain the MLE of $\mathcal{S}_{\boldsymbol{\Gamma}}$ and $\boldsymbol{\Delta}$.

Notice that

$$\Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\hat{\boldsymbol{\beta}}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big)^T\Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\hat{\boldsymbol{\beta}}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big) = \mathbb{X}^T\mathbb{X} - \mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Delta}^{-1})}\mathbb{X}^T\mathbb{F}_{\tilde{y}}(\mathbb{F}_{\tilde{y}}^T\mathbb{F}_{\tilde{y}})^{-1}\mathbb{F}_{\tilde{y}}^T\mathbb{X},$$

and

$$\boldsymbol{\Delta}^{-1/2}\mathbf{P}_{\boldsymbol{\Gamma}(\boldsymbol{\Delta}^{-1})} = \mathbf{P}_{\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}}\boldsymbol{\Delta}^{-1/2},$$

where we write $\mathbf{P}_{\mathbf{G}} = \mathbf{G}(\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T$ for a full rank matrix $\mathbf{G}$. We can easily obtain that

$$\begin{aligned}&\boldsymbol{\Delta}^{-1/2}\Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\hat{\boldsymbol{\beta}}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big)^T\Big(\mathbb{X} - \mathbb{F}_{\tilde{y}}\hat{\boldsymbol{\beta}}_{\tilde{y}}^T\boldsymbol{\Gamma}^T\Big)\boldsymbol{\Delta}^{-1/2}\\ =&n\Big(\boldsymbol{\Delta}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}^{-1/2} - \mathbf{P}_{\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}}\boldsymbol{\Delta}^{-1/2}\{\mathbb{X}^T\mathbf{P}_{\mathbb{F}_{\tilde{y}}}\mathbb{X}/n\}\boldsymbol{\Delta}^{-1/2}\Big),\end{aligned}$$

and

$$\begin{aligned}L_{\tilde{y}}(\mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\Delta}) =& -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Delta}|\\ &- \frac{n}{2}\text{trace}\Big\{\boldsymbol{\Delta}^{-1/2}\hat{\boldsymbol{\Sigma}}\boldsymbol{\Delta}^{-1/2} - \mathbf{P}_{\boldsymbol{\Delta}^{-1/2}\boldsymbol{\Gamma}}\boldsymbol{\Delta}^{-1/2}\{\mathbb{X}^T\mathbf{P}_{\mathbb{F}_{\tilde{y}}}\mathbb{X}/n\}\boldsymbol{\Delta}^{-1/2}\Big\}.\end{aligned}$$

If $\tilde{y}$ takes value from $\{\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_m\}$, we can consider the integration of all the log likelihood functions $L_{\tilde{y}_i}(\mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\Delta})$, $i = 1, ..., m$, to maximize the weighted average

$$\bar{L}(\mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\Delta}) = \frac{1}{m}\sum_{i=1}^{m}\omega(\tilde{y}_i)L_{\tilde{y}_i}(\mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\Delta}),$$

where $\omega(\cdot)/m$ is a nonnegative weight function with respect to $\tilde{y}_i$. Then the goal is to maximize

$$\begin{aligned}\bar{L}(\mathcal{S}_{\mathbf{\Gamma}}, \mathbf{\Delta}) =& -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Delta}| \\ & - \frac{n}{2}\text{trace}\Big\{\mathbf{\Delta}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Delta}^{-1/2} - \mathbf{P}_{\mathbf{\Delta}^{-1/2}\mathbf{\Gamma}}\mathbf{\Delta}^{-1/2}\frac{1}{m}\sum_{i=1}^{m}\Big\{\omega(\tilde{y}_i)\mathbb{X}^T\mathbf{P}_{\mathbb{F}_{\tilde{y}_i}}\mathbb{X}/n\Big\}\mathbf{\Delta}^{-1/2}\Big\} \\ =& -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Delta}| \\ & - \frac{n}{2}\text{trace}\Big\{\mathbf{\Delta}^{-1/2}\hat{\mathbf{\Sigma}}\mathbf{\Delta}^{-1/2} - \mathbf{P}_{\mathbf{\Delta}^{-1/2}\mathbf{\Gamma}}\mathbf{\Delta}^{-1/2}\hat{\mathbf{\Sigma}}_{cu}\mathbf{\Delta}^{-1/2}\Big\}.\end{aligned}$$

Holding $\mathbf{\Delta}$ fixed, the log likelihood is maximized by choosing $\mathbf{P}_{\mathbf{\Delta}^{-1/2}\mathbf{\Gamma}}$ as the projection onto the space $\text{span}_d(\mathbf{\Delta}^{-1/2}\hat{\Sigma}_{cu}\mathbf{\Delta}^{-1/2})$, where $\text{span}_d(\mathbf{A})$ denotes the space spanned by the first d eigenvectors of $\mathbf{A}$. It means that the span of $\mathbf{\Delta}^{-1}\mathbf{\Gamma}$ is the span of $\mathbf{\Delta}^{-1/2}$ times the first d eigenvectors of $\mathbf{\Delta}^{-1/2}\hat{\Sigma}_{cu}\mathbf{\Delta}^{-1/2}$, which is $\mathcal{S}_d(\mathbf{\Delta}, \hat{\mathbf{\Sigma}}_{cu})$ exactly. The subspace $\mathcal{S}_d(\mathbf{\Delta}, \hat{\mathbf{\Sigma}}_{cu})$ can also be described as the span of $\mathbf{\Delta}^{-1}$ times the first d eigenvectors of $\hat{\mathbf{\Sigma}}_{cu}$ (Adragni & Cook 2009)[1].

This leads to the final maximized log likelihood for $\mathbf{\Delta}$

$$\bar{L}(\mathbf{\Delta}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\mathbf{\Delta}| - \frac{n}{2}\text{trace}\{\mathbf{\Delta}^{-1}\hat{\mathbf{\Sigma}}_{res}\} - \frac{n}{2}\sum_{i=d+1}^{p}\lambda_i(\mathbf{\Delta}^{-1}\hat{\mathbf{\Sigma}}_{cu}),$$

where $\hat{\mathbf{\Sigma}}_{res} = \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_{cu}$ and $\lambda_i(\mathbf{A})$ denotes the ith eigenvalue of $\mathbf{A}$.

Thus the MLEs of all the dimension reduction parameters are $\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$, $\hat{\mathbf{\Delta}}^{-1}\hat{\mathcal{S}}_{\mathbf{\Gamma}} = \mathcal{S}_d(\hat{\mathbf{\Delta}}, \hat{\mathbf{\Sigma}}_{cu})$, $\hat{\boldsymbol{\beta}}_{\tilde{y}} = (\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Delta}}^{-1}\hat{\mathbf{\Gamma}})^{-1}\hat{\mathbf{\Gamma}}^T\hat{\mathbf{\Delta}}^{-1}\hat{\mathbf{B}}_{\tilde{y}}$, where $\hat{\mathbf{\Gamma}}$ is any orthonormal basis for $\hat{\mathcal{S}}_{\mathbf{\Gamma}}$, and the $\hat{\mathbf{\Delta}}$ is obtained by maximizing $\bar{L}(\mathbf{\Delta})$. □

The detailed proof of Proposition 3.2 can be referred to Theorem 3.1 in Cook and Forzani (2008)[2]. Their conclusion can be directly utilized here since the demonstration process concerns only the form of $L_d(\mathbf{\Delta})$ but not the specific form of $\hat{\mathbf{\Sigma}}_{fit}$ or $\hat{\mathbf{\Sigma}}_{cu}$. The $\bar{L}(\mathbf{\Delta})$ in this article is as the same form as $L_d(\mathbf{\Delta})$ in Cook and Forzani (2008)[2].

The proof of Proposition 3.3:

Proof. From the development of Proposition 3.1, the MLE of $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$ is $\mathcal{S}_d(\hat{\boldsymbol{\Delta}}, \hat{\boldsymbol{\Sigma}}_{cu})$, which establishes the second form.

To deduce the third form from the second form we need a lemma.

lemma 2. *Let* $\tilde{\mathbf{V}} = \hat{\boldsymbol{\Sigma}}_{res}^{-1/2}\hat{\mathbf{V}}\mathbf{M}^{1/2}$, *where* $\mathbf{M} = (\mathbf{I}_p + \hat{\mathbf{K}})^{-1}$, *with* $\hat{\mathbf{V}}$ *and* $\hat{\mathbf{K}}$ *as in Proposition 3.2. Then* $\hat{\boldsymbol{\Delta}}^{1/2}\tilde{\mathbf{V}}$ *are the normalized eigenvectors of* $\hat{\boldsymbol{\Delta}}^{-1/2}\hat{\boldsymbol{\Sigma}}_{cu}\hat{\boldsymbol{\Delta}}^{-1/2}$.

The proof of Lemma 2 can be found in Cook and Forzani (2008)[2] which replaces $\hat{\boldsymbol{\Sigma}}_{cu}$ with $\hat{\boldsymbol{\Sigma}}_{fit}$ but makes no difference because it concerns only the form of $\hat{\boldsymbol{\Delta}}$ but not the specific form of $\hat{\boldsymbol{\Sigma}}_{fit}$ or $\hat{\boldsymbol{\Sigma}}_{cu}$ in the demonstration process. The form of $\hat{\boldsymbol{\Delta}}$ in this article is the same as in Cook and Forzani (2008)[2].

Now, from the second form and Lemma 2, span of the first d columns of $\hat{\boldsymbol{\Delta}}^{-1/2}\hat{\boldsymbol{\Delta}}^{1/2}\tilde{\mathbf{V}} = \tilde{\mathbf{V}}$ is the MLE of $\boldsymbol{\Delta}^{-1}\mathcal{S}_{\boldsymbol{\Gamma}}$. Since $\tilde{\mathbf{V}} = \hat{\boldsymbol{\Sigma}}_{res}^{-1/2}\hat{\mathbf{V}}\mathbf{M}^{1/2}$ and $\mathbf{M}$ is diagonal full rank with the first d elements equal to 1, the span of the first d columns of $\tilde{\mathbf{V}}$ is the same of the first d columns of $\hat{\boldsymbol{\Sigma}}_{res}^{-1/2}\hat{\mathbf{V}}$. $\hat{\mathbf{V}}$ are the eigenvectors of $\hat{\boldsymbol{\Sigma}}_{res}^{-1/2}\hat{\boldsymbol{\Sigma}}_{cu}\hat{\boldsymbol{\Sigma}}_{res}^{-1/2}$, so the span of the first d columns of $\hat{\boldsymbol{\Sigma}}_{res}^{-1/2}\hat{\mathbf{V}}$ is $\mathcal{S}_d(\hat{\boldsymbol{\Sigma}}_{res}, \hat{\boldsymbol{\Sigma}}_{cu})$, which proves the third form.

The proof of the fourth form follows from the third form and the fact that the eigenvectors of $\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{cu}$ and $\hat{\boldsymbol{\Sigma}}_{res}^{-1}\hat{\boldsymbol{\Sigma}}_{cu}$ are identical, with corresponding eigenvalues $\hat{\lambda}_i/(1+\hat{\lambda}_i)$ and $\hat{\lambda}_i$, $i = 1, ..., p$.

Note that for symmetric matrices $\mathbf{A}$ and $\mathbf{B}$, the eigenvalues and eigenvectors of $\mathbf{AB}$ and $\mathbf{A}^{1/2}\mathbf{B}\mathbf{A}^{1/2}$ are identical. Thus with $\hat{\mathbf{v}}_i$ and $\hat{\lambda}_i$ as in Proposition 3.2 we have

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{res}^{-1}\hat{\boldsymbol{\Sigma}}_{cu}\hat{\mathbf{v}}_i = \hat{\lambda}_i\hat{\mathbf{v}}_i &\Leftrightarrow \hat{\lambda}_i^{-1}\hat{\boldsymbol{\Sigma}}_{cu}\hat{\mathbf{v}}_i = \hat{\boldsymbol{\Sigma}}_{res}\hat{\mathbf{v}}_i \\ &\Leftrightarrow (\hat{\lambda}_i^{-1}+1)\hat{\boldsymbol{\Sigma}}_{cu}\hat{\mathbf{v}}_i = (\hat{\boldsymbol{\Sigma}}_{res}+\hat{\boldsymbol{\Sigma}}_{cu})\hat{\mathbf{v}}_i = \hat{\boldsymbol{\Sigma}}\hat{\mathbf{v}}_i \\ &\Leftrightarrow \hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\Sigma}}_{cu}\hat{\mathbf{v}}_i = \hat{\lambda}_i/(1+\hat{\lambda}_i)\hat{\mathbf{v}}_i.\end{aligned}$$

The conclusion follows because $\hat{\mathbf{\Sigma}}_{res} = \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_{cu} > 0$ and $\hat{\lambda}_i/(1+\hat{\lambda}_i)$ is a strictly monotonic function of $\hat{\lambda}_i$.

The proof of the first form follows from the third form and the fact that the eigenvectors of $\hat{\mathbf{\Sigma}}_{res}^{-1}\hat{\mathbf{\Sigma}}$ and $\hat{\mathbf{\Sigma}}_{res}^{-1}\hat{\mathbf{\Sigma}}_{cu}$ are identical, with corresponding eigenvalues $(1+\hat{\lambda}_i)$ and $\hat{\lambda}_i$, $i = 1, ..., p$.

$$\hat{\mathbf{\Sigma}}_{res}^{-1}\hat{\mathbf{\Sigma}}_{cu}\hat{\mathbf{v}}_i = \hat{\lambda}_i\hat{\mathbf{v}}_i \Leftrightarrow \hat{\mathbf{\Sigma}}_{res}^{-1}\hat{\mathbf{\Sigma}}\hat{\mathbf{v}}_i = (\mathbf{I}_p + \hat{\mathbf{\Sigma}}_{res}^{-1}\hat{\mathbf{\Sigma}}_{cu})\hat{\mathbf{v}}_i = (1+\hat{\lambda}_i)\hat{\mathbf{v}}_i$$

The conclusion follows because $\hat{\mathbf{\Sigma}}_{res} = \hat{\mathbf{\Sigma}} - \hat{\mathbf{\Sigma}}_{cu} > 0$ and $(1+\hat{\lambda}_i)$ is a strictly monotonic function of $\hat{\lambda}_i$.

□

The proof of Theorem 3.4:

Proof. We study consistency of the estimator $\mathcal{S}_d(\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}}_{cu})$ under the inverse model (1) no matter what the real form of $\mathbf{f}_y$ or the nature of $\boldsymbol{\varepsilon}$ is. Since $\mathcal{S}_d(\hat{\mathbf{\Sigma}}, \hat{\mathbf{\Sigma}}_{cu})$ is the span of $\hat{\mathbf{\Sigma}}^{-1}$ times the first d eigenvectors of $\hat{\mathbf{\Sigma}}_{cu}$, which equals the span of the first d eigenvectors of $\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}_{cu}$, it is sufficient to consider the property of $\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}_{cu}$.

Under the inverse model $\mathbf{X}_y = \boldsymbol{\mu} + \mathbf{\Gamma}\boldsymbol{\nu}_y + \boldsymbol{\varepsilon}$, the covariance matrix of $\mathbf{X} \in \mathbb{R}^p$ is $\mathbf{\Sigma} = \mathbf{\Gamma}\mathbf{V}\mathbf{\Gamma}^T + \mathbf{\Delta}$, where $\mathbf{V} = \text{var}(\boldsymbol{\nu}_Y)$ is positive definite. Given pre-specified $\tilde{y}_1, \tilde{y}_2, ..., \tilde{y}_m$, Define that

$$\mathbf{\Sigma}_{cu} = \frac{1}{m}\mathbf{\Sigma}^{1/2}\Big(\text{Corr}(\mathbf{X}, f_{Y;\tilde{y}_1}), ..., \text{Corr}(\mathbf{X}, f_{Y;\tilde{y}_m})\Big)\Big(\text{Corr}(\mathbf{X}, f_{Y;\tilde{y}_1}), ..., \text{Corr}(\mathbf{X}, f_{Y;\tilde{y}_m})\Big)^T\mathbf{\Sigma}^{1/2},$$

where $\mathbf{X} \in \mathbb{R}^p$ and $f_{Y;\tilde{y}_i} = I(Y \le \tilde{y}_i) - Pr(Y \le \tilde{y}_i)$.

It is known that the sample covariance matrix $\hat{\mathbf{\Sigma}} = \mathbb{X}^T\mathbb{X}/n$ is a $\sqrt{n}$-consistent estimator of $\mathbf{\Sigma}$. Hence $\hat{\mathbf{\Sigma}}^{-1}$ is a $\sqrt{n}$-consistent estimator of $\mathbf{\Sigma}^{-1}$ (Cook & Forzani

2008)[2]. Without loss of generality we assume that $\omega(\cdot) = 1$, then

$$\begin{aligned}\hat{\boldsymbol{\Sigma}}_{cu} =& \frac{1}{m}\sum_{i=1}^{m}\{\mathbb{X}^T\mathbf{P}_{\mathbb{F}_{\tilde{y}_i}}\mathbb{X}/n\} = \frac{1}{m}\sum_{i=1}^{m}\{\mathbb{X}^T\mathbb{F}_{\tilde{y}_i}(\mathbb{F}_{\tilde{y}_i}^T\mathbb{F}_{\tilde{y}_i})^{-1}\mathbb{F}_{\tilde{y}_i}^T\mathbb{X}/n\} \\ =& \frac{1}{m}\sum_{i=1}^{m}\left\{\frac{\mathbb{X}^T\mathbb{F}_{\tilde{y}_i}}{n}\Big(\frac{\mathbb{F}_{\tilde{y}_i}^T\mathbb{F}_{\tilde{y}_i}}{n}\Big)^{-1}\frac{\mathbb{F}_{\tilde{y}_i}^T\mathbb{X}}{n}\right\}.\end{aligned}$$

As $\mathbb{X}^T\mathbb{F}_{\tilde{y}_i}/n$ is a $\sqrt{n}$-consistent estimator of $\mathrm{cov}(\mathbf{X}, f_{Y;\tilde{y}_i})$ and $(\mathbb{F}_{\tilde{y}_i}^T\mathbb{F}_{\tilde{y}_i}/n)^{-1}$ is a $\sqrt{n}$-consistent estimator of $\mathrm{var}(f_{Y;\tilde{y}_i})$, then $(\mathbb{X}^T\mathbb{F}_{\tilde{y}_i}/n)(\mathbb{F}_{\tilde{y}_i}^T\mathbb{F}_{\tilde{y}_i}/n)^{-1}(\mathbb{F}_{\tilde{y}_i}^T\mathbb{X}/n)$ converges at $\sqrt{n}$ rate to $\boldsymbol{\Sigma}^{1/2}\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})^T\boldsymbol{\Sigma}^{1/2}$ when n approaches ∞ (Cook & Forzani 2008)[2].

Next we consider the convergence of $\sum_{i=1}^{m}(\mathbb{X}^T\mathbb{F}_{\tilde{y}_i}/n)(\mathbb{F}_{\tilde{y}_i}^T\mathbb{F}_{\tilde{y}_i}/n)^{-1}(\mathbb{F}_{\tilde{y}_i}^T\mathbb{X}/n) = \sum_{i=1}^{m}\hat{\boldsymbol{\Sigma}}_{\tilde{y}_i}$. As $\forall\epsilon > 0$, $\forall i \in \{1, ..., m\}$,

$$P\Big(\Big|\hat{\boldsymbol{\Sigma}}_{\tilde{y}_i} - \boldsymbol{\Sigma}^{1/2}\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})^T\boldsymbol{\Sigma}^{1/2}\Big| < \epsilon\Big) \to 1,$$

we can conclude that

$$\begin{aligned}&P\Big(\frac{1}{m}\Big|\sum_{i=1}^{m}\hat{\boldsymbol{\Sigma}}_{\tilde{y}_i} - \sum_{i=1}^{m}\boldsymbol{\Sigma}^{1/2}\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})^T\boldsymbol{\Sigma}^{1/2}\Big| < \epsilon\Big) \\ \geq& P\Big(\frac{1}{m}\sum_{i=1}^{m}\Big|\hat{\boldsymbol{\Sigma}}_{\tilde{y}_i} - \boldsymbol{\Sigma}^{1/2}\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})^T\boldsymbol{\Sigma}^{1/2}\Big| < \epsilon\Big) \\ \geq& P\Big(\forall i \in \{1, ..., m\}, \Big|\hat{\boldsymbol{\Sigma}}_{\tilde{y}_i} - \boldsymbol{\Sigma}^{1/2}\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})^T\boldsymbol{\Sigma}^{1/2}\Big| < \epsilon\Big) \\ =& P\Big(\max_i\Big|\hat{\boldsymbol{\Sigma}}_{\tilde{y}_i} - \boldsymbol{\Sigma}^{1/2}\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})\mathrm{Corr}(\mathbf{X}, f_{Y;\tilde{y}_i})^T\boldsymbol{\Sigma}^{1/2}\Big| < \epsilon\Big) \to 1,\end{aligned}$$

as $n \to \infty$. Thus $\hat{\boldsymbol{\Sigma}}_{cu}$ converges to $\boldsymbol{\Sigma}_{cu}$ at rate not less than $\sqrt{n}$ since $\boldsymbol{\Sigma}_{\tilde{y}_i}$ is $\sqrt{n}$-consistent.

Combined with model (2) in the article we have

$$\begin{aligned}\mathrm{corr}(\mathbf{X}, f_{Y;\tilde{y}_i}) =& \boldsymbol{\Sigma}^{-1/2}\mathrm{Cov}(\boldsymbol{\mu} + \boldsymbol{\Gamma}\boldsymbol{\nu}_Y + \boldsymbol{\varepsilon}, f_{Y;\tilde{y}_i})\mathrm{var}(f_{Y;\tilde{y}_i})^{-1/2} \\ =& \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}\mathrm{Cov}(\boldsymbol{\nu}_Y, f_{Y;\tilde{y}_i})\mathrm{var}(f_{Y;\tilde{y}_i})^{-1/2} \\ =& \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\Gamma}\mathbf{V}^{1/2}\mathrm{Corr}(\boldsymbol{\nu}_Y, f_{Y;\tilde{y}_i}).\end{aligned}$$

$\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}_{cu}$ therefore converges to

$$\begin{aligned}\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{cu} =& \frac{1}{m}\mathbf{\Sigma}^{-1}\mathbf{\Gamma}\mathbf{V}^{1/2}\mathbf{C}\mathbf{C}^T\mathbf{V}^{1/2}\mathbf{\Gamma}^T \\ =& \frac{1}{m}(\mathbf{\Gamma}\mathbf{V}\mathbf{\Gamma}+\mathbf{\Delta})^{-1}\mathbf{\Gamma}\mathbf{V}^{1/2}\mathbf{C}\mathbf{C}^T\mathbf{V}^{1/2}\mathbf{\Gamma}^T.\end{aligned}$$

at not-less-than $\sqrt{n}$ rate, and as a result the first d eigenvectors of $\hat{\mathbf{\Sigma}}^{-1}\hat{\mathbf{\Sigma}}_{cu}$ converge at not-less-than $\sqrt{n}$ rate to the corresponding eigenvectors of $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{cu}$.

Now we focus on the relationship between $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{cu}$ and $\mathbf{\Delta}^{-1}\mathcal{S}_{\mathbf{\Gamma}}$. Based on

$$(\mathbf{\Gamma}\mathbf{V}\mathbf{\Gamma}^T+\mathbf{\Delta})^{-1} = \mathbf{\Delta}^{-1} - \mathbf{\Delta}^{-1}\mathbf{\Gamma}(\mathbf{V}^{-1}+\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{\Gamma})^{-1}\mathbf{\Gamma}^T\mathbf{\Delta}^{-1},$$

we simplify $\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{cu}$ as

$$\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{cu} = \frac{1}{m}\mathbf{\Delta}^{-1}\mathbf{\Gamma}\mathbf{K}\mathbf{V}^{1/2}\mathbf{C}\mathbf{C}^T\mathbf{V}^{1/2}\mathbf{\Gamma}^T,$$

where $\mathbf{K} = (\mathbf{V}^{-1}+\mathbf{\Gamma}^T\mathbf{\Delta}^{-1}\mathbf{\Gamma})^{-1}\mathbf{V}^{-1}$ is a full rank $d\times d$ matrix. Clearly $\text{span}(\mathbf{\Sigma}^{-1}\mathbf{\Sigma}_{cu}) \subseteq \mathbf{\Delta}^{-1}\mathcal{S}_{\mathbf{\Gamma}}$ with equality if and only if the rank of $\mathbf{\Gamma}\mathbf{K}\mathbf{V}^{1/2}\mathbf{C}\mathbf{C}^T\mathbf{V}^{1/2}\mathbf{\Gamma}^T$ is equal to d. Since $\mathbf{\Gamma} \in \mathbb{R}^{p\times d}$ has full column rank and both $\mathbf{K}$ and $\mathbf{V}$ is a full rank matrix, the rank of $\mathbf{\Gamma}\mathbf{K}\mathbf{V}^{1/2}\mathbf{C}\mathbf{C}^T\mathbf{V}^{1/2}\mathbf{\Gamma}^T$ is equal to d if and only if the rank of $\mathbf{C}\mathbf{C}^T$ is equal to d, which requires that $\mathbf{C}$ has rank d. □

References

[1] Adragni, K. P. and Cook, R. D. Sufficient Dimension Reduction and Prediction in Regression. *Philosophical Transaction of the Royal Society A*, 2009, 367(1906):4385–4405.

[2] Cook, R. D. and Forzani, L. Principal fitted components for dimension reduction in regression. *Statistical Science*, 2008, 23(4):485–501.