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

Xinyi Xu<sup>†</sup>, Xiangjie Li<sup>†</sup>, Jingxiao Zhang<sup>†\*</sup>

<sup>†</sup>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  $\Delta^{-1}S_{\Gamma} = {\Delta^{-1}\mathbf{z} : \mathbf{z} \in S_{\Gamma}}$  under the CUPFC model.

The proof of Proposition 3.1:

*Proof.* Under the CUPFC model the full parameter space is  $(\boldsymbol{\mu}, \mathcal{S}_{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\Delta})$ . 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

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

<sup>\*</sup>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:

$$\mathbf{X}_y = \boldsymbol{\mu} + \Gamma \boldsymbol{eta}_{ ilde{y}} f_{y; ilde{y}} + oldsymbol{arepsilon}$$

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:

$$g(\mathbf{X}_{y}: y \in S) = (2\pi)^{-\frac{np}{2}} |\mathbf{\Delta}|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{y} \left(\mathbf{X}_{y} - \boldsymbol{\mu} - \boldsymbol{\Gamma} \boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\right)^{T} \cdot \boldsymbol{\Delta}^{-1} \left(\mathbf{X}_{y} - \boldsymbol{\mu} - \boldsymbol{\Gamma} \boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}}\right)\right\},$$

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

$$L_{\tilde{y}}(\boldsymbol{\mu}, \mathcal{S}_{\Gamma}, \boldsymbol{\beta}, \boldsymbol{\Delta}) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log|\boldsymbol{\Delta}| -\frac{1}{2} \sum_{y} \left( \mathbf{X}_{y} - \boldsymbol{\mu} - \boldsymbol{\Gamma} \boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}} \right)^{T} \boldsymbol{\Delta}^{-1} \left( \mathbf{X}_{y} - \boldsymbol{\mu} - \boldsymbol{\Gamma} \boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}} \right).$$
<sup>(1)</sup>

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

$$\sum_{y} \left( \mathbf{X}_{y} - \boldsymbol{\mu} - \boldsymbol{\Gamma} \boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}} \right)^{T} \boldsymbol{\Delta}^{-1} \left( \mathbf{X}_{y} - \boldsymbol{\mu} - \boldsymbol{\Gamma} \boldsymbol{\beta}_{\tilde{y}} f_{y;\tilde{y}} \right)$$
$$= \operatorname{trace} \left\{ \boldsymbol{\Delta}^{-1/2} \left( \mathbb{X} - \mathbb{F}_{\tilde{y}} \boldsymbol{\beta}_{\tilde{y}}^{T} \boldsymbol{\Gamma}^{T} \right)^{T} \left( \mathbb{X} - \mathbb{F}_{\tilde{y}} \boldsymbol{\beta}_{\tilde{y}}^{T} \boldsymbol{\Gamma}^{T} \right) \boldsymbol{\Delta}^{-1/2} \right\},$$

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 *k*th 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

$$L_{\tilde{y}}(\boldsymbol{\mu}, \mathcal{S}_{\boldsymbol{\Gamma}}, \boldsymbol{\beta}, \boldsymbol{\Delta}) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log|\boldsymbol{\Delta}| -\frac{1}{2} \operatorname{trace} \left\{ \boldsymbol{\Delta}^{-1/2} \left( \mathbb{X} - \mathbb{F}_{\tilde{y}} \boldsymbol{\beta}_{\tilde{y}}^{T} \boldsymbol{\Gamma}^{T} \right)^{T} \left( \mathbb{X} - \mathbb{F}_{\tilde{y}} \boldsymbol{\beta}_{\tilde{y}}^{T} \boldsymbol{\Gamma}^{T} \right) \boldsymbol{\Delta}^{-1/2} \right\}.$$
<sup>(2)</sup>

For fixed  $\Delta$  and  $\Gamma$ , equation (2) is maximized over  $\beta_{\tilde{y}}$  by  $\hat{\beta}_{\tilde{y}} = \Gamma^T \mathbf{P}_{\Gamma(\Delta^{-1})} \hat{\mathbf{B}}_{\tilde{y}}$ , where  $\mathbf{P}_{\Gamma(\Delta^{-1})} = \Gamma(\Gamma^T \Delta^{-1} \Gamma)^{-1} \Gamma^T \Delta^{-1}$  is the projection onto  $\mathcal{S}_{\Gamma}$  in the  $\Delta^{-1}$  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 *k*th 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  $\Gamma \in \mathbb{R}^{p \times d}$  an orthonormal basis of  $\mathcal{S}_{\Gamma}$  without loss of generality, so the MLE  $\Gamma \hat{\beta}_{\tilde{y}}$  will be  $\mathbf{P}_{\Gamma(\Delta^{-1})} \hat{\mathbf{B}}_{\tilde{y}}$ . We then substitute  $\hat{\mu}$  and  $\hat{\beta}_{\tilde{y}}$  into the log likelihood  $L_{\tilde{y}}(\mu, \mathcal{S}_{\Gamma}, \beta, \Delta)$  to attain the MLE of  $\mathcal{S}_{\Gamma}$ and  $\Delta$ .

Notice that

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

and

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

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

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

and

$$L_{\tilde{y}}(\mathcal{S}_{\Gamma}, \mathbf{\Delta}) = -\frac{np}{2} \log(2\pi) - \frac{n}{2} \log|\mathbf{\Delta}| - \frac{n}{2} \operatorname{trace} \left\{ \mathbf{\Delta}^{-1/2} \hat{\mathbf{\Sigma}} \mathbf{\Delta}^{-1/2} - \mathbf{P}_{\mathbf{\Delta}^{-1/2}\Gamma} \mathbf{\Delta}^{-1/2} \{ \mathbb{X}^T \mathbf{P}_{\mathbb{F}_{\tilde{y}}} \mathbb{X}/n \} \mathbf{\Delta}^{-1/2} \right\}.$$

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  $L_{\tilde{y}_i}(\mathcal{S}_{\Gamma}, \Delta)$ , i = 1, ..., m, to maximize the weighted average

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

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

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

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

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

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

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

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

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  $L_d(\Delta)$  but not the specific form of  $\hat{\Sigma}_{fit}$  or  $\hat{\Sigma}_{cu}$ . The  $\bar{L}(\Delta)$  in this article is as the same form as  $L_d(\Delta)$  in Cook and Forzani (2008)[2].

The proof of Proposition 3.3:

*Proof.* From the development of Proposition 3.1, the MLE of  $\Delta^{-1}S_{\Gamma}$  is  $S_d(\hat{\Delta}, \hat{\Sigma}_{cu})$ , 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  $\hat{\Sigma}_{cu}$  with  $\hat{\Sigma}_{fit}$  but makes no difference because it concerns only the form of  $\hat{\Delta}$  but not the specific form of  $\hat{\Sigma}_{fit}$  or  $\hat{\Sigma}_{cu}$  in the demonstration process. The form of  $\hat{\Delta}$  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  $\hat{\Delta}^{-1/2}\hat{\Delta}^{1/2}\tilde{\mathbf{V}} = \tilde{\mathbf{V}}$  is the MLE of  $\Delta^{-1}\mathcal{S}_{\Gamma}$ . Since  $\tilde{\mathbf{V}} = \hat{\Sigma}_{res}^{-1/2}\hat{\mathbf{V}}\mathbf{M}^{1/2}$  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  $\hat{\Sigma}_{res}^{-1/2}\hat{\mathbf{V}}$ .  $\hat{\mathbf{V}}$  are the eigenvectors of  $\hat{\Sigma}_{res}^{-1/2}\hat{\Sigma}_{cu}\hat{\Sigma}_{res}^{-1/2}$ , so the span of the first d columns of  $\hat{\Sigma}_{res}^{-1/2}\hat{\mathbf{V}}$  is  $\mathcal{S}_d(\hat{\Sigma}_{res},\hat{\Sigma}_{cu})$ , which proves the third form.

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

Note that for symmetric matrices **A** and **B**, the eigenvalues and eigenvectors of **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

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

The conclusion follows because  $\hat{\Sigma}_{res} = \hat{\Sigma} - \hat{\Sigma}_{cu} > 0$  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  $\hat{\Sigma}_{res}^{-1}\hat{\Sigma}$  and  $\hat{\Sigma}_{res}^{-1}\hat{\Sigma}_{cu}$  are identical, with corresponding eigenvalues  $(1 + \hat{\lambda}_i)$  and  $\hat{\lambda}_i$ , i = 1, ..., p.

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

The conclusion follows because  $\hat{\Sigma}_{res} = \hat{\Sigma} - \hat{\Sigma}_{cu} > 0$  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  $S_d(\hat{\Sigma}, \hat{\Sigma}_{cu})$  under the inverse model (1) no matter what the real form of  $\mathbf{f}_y$  or the nature of  $\boldsymbol{\varepsilon}$  is. Since  $S_d(\hat{\Sigma}, \hat{\Sigma}_{cu})$  is the span of  $\hat{\Sigma}^{-1}$  times the first d eigenvectors of  $\hat{\Sigma}_{cu}$ , which equals the span of the first d eigenvectors of  $\hat{\Sigma}_{cu}$ , which equals the span of the first d eigenvectors of  $\hat{\Sigma}^{-1}\hat{\Sigma}_{cu}$ , it is sufficient to consider the property of  $\hat{\Sigma}^{-1}\hat{\Sigma}_{cu}$ .

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

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

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

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

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

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

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

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

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

we can conclude that

$$\begin{split} &P\left(\frac{1}{m} \left| \sum_{i=1}^{m} \hat{\boldsymbol{\Sigma}}_{\tilde{y}_{i}} - \sum_{i=1}^{m} \boldsymbol{\Sigma}^{1/2} \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}}) \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}})^{T} \boldsymbol{\Sigma}^{1/2} \right| < \epsilon \right) \\ &\geq P\left(\frac{1}{m} \sum_{i=1}^{m} \left| \hat{\boldsymbol{\Sigma}}_{\tilde{y}_{i}} - \boldsymbol{\Sigma}^{1/2} \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}}) \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}})^{T} \boldsymbol{\Sigma}^{1/2} \right| < \epsilon \right) \\ &\geq P\left(\forall i \in \{1, ..., m\}, \left| \hat{\boldsymbol{\Sigma}}_{\tilde{y}_{i}} - \boldsymbol{\Sigma}^{1/2} \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}}) \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}})^{T} \boldsymbol{\Sigma}^{1/2} \right| < \epsilon \right) \\ &= P\left( \max_{i} \left| \hat{\boldsymbol{\Sigma}}_{\tilde{y}_{i}} - \boldsymbol{\Sigma}^{1/2} \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}}) \operatorname{Corr}(\mathbf{X}, f_{Y;\tilde{y}_{i}})^{T} \boldsymbol{\Sigma}^{1/2} \right| < \epsilon \right) \rightarrow 1, \end{split}$$

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

Combined with model (2) in the article we have

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

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

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

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

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

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

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 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

- 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.