Sufficient Dimension Reduction for Classification

Xin Chen, Jingjing Wu, Zhigang Yao and Jia Zhang

Southern University of Science and Technology, University of Calgary,

National University of Singapore and Southwestern University of Finance and Economics

Supplementary Material

The supplementary material includes all the theoretical proof of the main paper.

S1 Proof of Theorem 1

Proof. The following lemma is required.

Lemma 1. *Under the linearity and coverage condition, it holds that*

$$\textit{if } \; Y \perp\!\!\!\perp \boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X}, \quad \textit{then } \; \boldsymbol{\beta}_{0i} \perp S_{\mathbf{X}|Y}. \tag{S1.1}$$

Proof. We use the method of contradiction to prove the argument in (S1.1). Without loss of generality, we assume $\mathrm{E}(\mathbf{X}) = \mathbf{0}$ and $\mathrm{Cov}(\mathbf{X}) = \mathbf{I}_p$ in the following proof. If $\boldsymbol{\beta}_{0i} \not\perp S_{\mathbf{X}|Y}$, then $\boldsymbol{\beta}_{0i} = \boldsymbol{\gamma}_{0i} + \boldsymbol{\lambda}_{0i}$, where $\boldsymbol{\gamma}_{0i} \in S_{\mathbf{X}|Y}$, $\boldsymbol{\gamma}_{0i} \neq \mathbf{0}$, and $\boldsymbol{\lambda}_{0i} \in S_{\mathbf{X}|Y}^{\perp}$. We consider

$$\mathrm{E}(\boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y) + \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) =: T_1 + T_2\,.$$

Let $\boldsymbol{\mu}_Y = \mathrm{E}(\mathbf{X}|Y)$ and $\boldsymbol{\Sigma}_{\mathrm{E}(\mathbf{X}|Y)} = \mathrm{Cov}\{\mathrm{E}(\mathbf{X}|Y)\}$. For T_1, noting that $\mathrm{E}(\boldsymbol{\mu}_Y) = \mathrm{E}(\mathbf{X}) = \mathbf{0}$ and $\boldsymbol{\gamma}_{0i} \in S_{\mathbf{X}|Y}$, we have

$$\mathrm{Var}\{\mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y)\} = \mathrm{Var}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\boldsymbol{\mu}_Y) = \boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathrm{E}(\boldsymbol{\mu}_Y\boldsymbol{\mu}_Y^{\mathrm{T}})\boldsymbol{\gamma}_{0i} = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{E}(\mathbf{X}|Y)}\boldsymbol{\Gamma}\boldsymbol{\alpha}_1\,, \tag{S1.2}$$

where $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times d_*}$ is an orthonormal basis of the central subspace $S_{\mathbf{X}|Y}$ and $\boldsymbol{\alpha}_1 \in \mathbb{R}^{d_*}$ does not equal $\mathbf{0}$. If the linearity condition and the coverage condition hold, we know that (Li, 1991; Tan et al., 2020)

$$\boldsymbol{\Sigma}_{\mathrm{E}(\mathbf{X}|Y)} = \sum_{i=1}^{d_*}\lambda_i\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^{\mathrm{T}} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}\,,$$

where $\boldsymbol{\Gamma} = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_{d_*})$ and $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{d_*})$ with $\lambda_1 > \ldots > \lambda_{d_*} > 0$. Therefore,

$$\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{E}(\mathbf{X}|Y)}\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma} = \boldsymbol{\Lambda}\,.$$

Combining (S1.2), we obtain $\mathrm{Var}\{\mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y)\} = \boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Lambda}\boldsymbol{\alpha}_1 \geq \lambda_{d_*}\|\alpha_1\|^2 > 0$, which implies that

$$P\{\mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X})\} < 1\,. \tag{S1.3}$$

For T_2, we have

$$\mathrm{Var}\{\mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y)\} = \mathrm{Var}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\boldsymbol{\mu}_Y) = \boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathrm{E}(\boldsymbol{\mu}_Y\boldsymbol{\mu}_Y^{\mathrm{T}})\boldsymbol{\lambda}_{0i} = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Gamma}^{\perp,\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{E}(\mathbf{X}|Y)}\boldsymbol{\Gamma}^{\perp}\boldsymbol{\alpha}_2\,,$$

where $\boldsymbol{\Gamma}^{\perp} \in \mathbb{R}^{p\times(p-d_*)}$ is an orthonormal basis of $S_{\mathbf{X}|Y}^{\perp}$ such that $\boldsymbol{\Gamma}^{\perp,\mathrm{T}}\boldsymbol{\Gamma} = \mathbf{0}$. Recall the linearity condition and coverage condition imply that $\boldsymbol{\Sigma}_{\mathrm{E}(\mathbf{X}|Y)} =$

$\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}$, and then it holds that

$$\mathrm{Var}\{\mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y)\} = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Gamma}^{\perp,\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{E}(\mathbf{X}|Y)}\boldsymbol{\Gamma}^{\perp}\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_2^{\mathrm{T}}\boldsymbol{\Gamma}^{\perp,\mathrm{T}}\boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma}^{\perp}\boldsymbol{\alpha}_2 = 0\,,$$

which leads to

$$P\{\mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X})\} = 1\,. \tag{S1.4}$$

Combining (S1.3) and (S1.4), it holds that

$$\begin{aligned}
&P\{\mathrm{E}(\boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X})\} \\
&= P\{\mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y) + \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}) + \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X})\} \\
&= P\{\mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y) + \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}) + \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}), \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X})\} \\
&\quad + P\{\mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y) + \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}) + \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}), \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) \neq \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X})\} \\
&\leq P\{\mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\gamma}_{0i}^{\mathrm{T}}\mathbf{X})\} + P\{\mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X}|Y) \neq \mathrm{E}(\boldsymbol{\lambda}_{0i}^{\mathrm{T}}\mathbf{X})\} \\
&< 1\,.
\end{aligned}$$

From the relationship between mean independence and independence, we obtain

$$P\{\mathrm{E}(\boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{E}(\boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X})\} < 1 \;\Rightarrow\; Y \not\perp\!\!\!\perp \boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X}\,.$$

We complete the proof for the argument in (S1.1). □

Now we begin the proof. By the property of the MV index, we have $\boldsymbol{\beta}_{0i}^{\mathrm{T}}\mathbf{X} \perp\!\!\!\perp Y$ for $i = d+1, \ldots, p$. From Lemma 1, it holds that $\boldsymbol{\beta}_{0i} \in S_{Y|\mathbf{X}}^{\perp}$ for

$i = d+1, \ldots, p$, where $S_{Y|\mathbf{X}}^{\perp}$ denotes the orthogonal complement space of $S_{Y|\mathbf{X}}$. Noting that $\boldsymbol{\beta}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_{0i} = 1$, $\boldsymbol{\beta}_{0i}^{\mathrm{T}}\boldsymbol{\beta}_{0j} = 0$ for $i, j = 1, \ldots, p$ and $i \neq j$ and $\mathrm{span}(\boldsymbol{\beta}_{0(d+1)}, \ldots, \boldsymbol{\beta}_{0p}) \subseteq S_{Y|\mathbf{X}}^{\perp}$, we easily obtain $S_{Y|\mathbf{X}} \subseteq \mathrm{span}(\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0d})$.

For any $k < d$, if $\mathrm{span}(\boldsymbol{\beta}_{01}, \ldots, \boldsymbol{\beta}_{0k}) \supseteq S_{Y|\mathbf{X}}$, then $Y \perp\!\!\!\perp \mathbf{X} | \boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}, \ldots, \boldsymbol{\beta}_{0k}^{\mathrm{T}}\mathbf{X}$. Since $\mathbf{X} \sim N(\mathbf{0}, \mathbf{I})$, it holds that $\boldsymbol{\beta}_{0d}^{\mathrm{T}}\mathbf{X} \perp\!\!\!\perp (\boldsymbol{\beta}_{01}^{\mathrm{T}}\mathbf{X}, \ldots, \boldsymbol{\beta}_{0k}^{\mathrm{T}}\mathbf{X})$. The above two arguments imply that $Y \perp\!\!\!\perp \boldsymbol{\beta}_{0d}^{\mathrm{T}}\mathbf{X}$ (Cook (1994), Corollary 5.1), which contradicts the assumption that $\mathrm{MV}(\boldsymbol{\beta}_{0d}^{\mathrm{T}}\mathbf{X}|Y) \neq 0$. We complete the proof.

□

S2 Proof of Corollary 2

Proof. Let Z denote the variable $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ and Z_i denote the variable $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y = i$ $(i = 1, -1)$ for $\boldsymbol{\beta} \in \mathbb{R}^p$ satisfying $\|\boldsymbol{\beta}\| = 1$. Then, $Z_1 \sim N(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})$, $Z_{-1} \sim N(-\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}, \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})$ and

$$
\begin{aligned}
F(Z) &= p_1 \cdot F_1(Z) + p_{-1} \cdot F_{-1}(Z) \\
&= p_1 \cdot \Phi\left((Z - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu})/\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}\right) + p_{-1} \cdot \Phi\left((Z + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu})/\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}\right),
\end{aligned}
$$

where $F_i(Z)$ denotes the conditional distribution function of $Z|Y = i$ for $i = 1, -1$. Then,

$$\begin{aligned}
&\mathrm{MV}(\boldsymbol{\beta}) \\
&= p_1 \int \{F_1(z) - F(z)\}^2 \mathrm{d}F(z) + p_{-1} \int \{F_{-1}(z) - F(z)\}^2 \mathrm{d}F(z) \\
&= \left(p_1 p_{-1}^2 + p_{-1} p_1^2\right) \left[p_1 \int \{F_1(z) - F_{-1}(z)\}^2 \mathrm{d}F_1(z) + p_{-1} \int \{F_1(z) - F_{-1}(z)\}^2 \mathrm{d}F_{-1}(z) \right] \\
&= p_1 p_{-1} \left[p_1 \int \left\{ \Phi(t) - \Phi(t + 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}/\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}) \right\}^2 \mathrm{d}\Phi(t) \right. \\
&\quad \left. + p_{-1} \int \left\{ \Phi(t) - \Phi\left(t - 2\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}/\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}\right) \right\}^2 \mathrm{d}\Phi(t) \right].
\end{aligned}$$

Hence, we only need to maximize $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}/\sqrt{\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}}$. The solution is exactly the optimal weight of LDA, i.e. $\boldsymbol{\beta}_{01} \propto \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$, where $\boldsymbol{\beta}_{01}$ is defined in Theorem 1. Thus, if we choose $d = 1$ in the MMV procedure, then MMV+LDA equals LDA at the population level.

Then we try to find $\boldsymbol{\beta}_{02}$. We always assume $\boldsymbol{\Sigma}$ is positive definite, and thus for any non-zero $\boldsymbol{\beta}$, $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta} > 0$. Since now we maximize $(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}$ subject to $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = 1$ and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 0$, let

$$L = (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu} + \lambda(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} - 1) + \pi\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\,.$$

Take first partial derivative w.r.t. $\boldsymbol{\beta}$ and set it to zero. We then have

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-3/2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma} + 2\lambda\boldsymbol{\beta}^{\mathrm{T}} + \pi\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} = 0\,.$$

Then

$$\begin{aligned} 0 &= \frac{\partial L}{\partial \boldsymbol{\beta}}\boldsymbol{\beta} \\ &= (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta} - (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-1/2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu} + 2\lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} + \pi\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\beta \\ &= 2\lambda\,, \end{aligned}$$

i.e. $\lambda = 0$. Now $(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{3/2}\frac{\partial L}{\partial \boldsymbol{\beta}} = 0$ is reduced to

$$0 = (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma} + (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{3/2}\pi\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}. \tag{S2.1}$$

Multiply both sides of above equation by $\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}$ from right, together with $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} = 1$ and $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 0$, we have

$$\begin{aligned} 0 &= (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} + (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{3/2}\pi\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} \\ &= 0 - \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\mu} + \pi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{3/2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta} \\ &= \pi(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{3/2}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta} - \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta}\,, \end{aligned} \tag{S2.2}$$

and thus

$$\pi = (\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta})^{-3/2}\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta}\right)^{-1}(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta})\,. \tag{S2.3}$$

Note that none of π, $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta}$ and $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta}$ can be 0. If $\pi = 0$, then (S2.2) gives $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta} = 0$. Plugging this and $\pi = 0$ back into (S2.1) gives $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\beta}\boldsymbol{\mu}^{\mathrm{T}} = 0$, i.e. $\boldsymbol{\mu} = 0$, which contradicts the assumption $\boldsymbol{\mu} \neq 0$ and thus $\pi \neq 0$. If $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta} = 0$, then (S2.2) gives $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta} = 0$. Multiplying $\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta}$ from the right to (S2.1) and plugging in $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta} = 0$ and $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta} = 0$, we have $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-3}\boldsymbol{\beta} = 0$

as $\pi \neq 0$. If we repeat this deduction then we have $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-i}\boldsymbol{\beta} = 0$, $i = 0, 1, \cdots$. This can not hold for general $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ unless $\boldsymbol{\beta} = 0$, which contradicts $\boldsymbol{\beta} = 1$, and thus $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta} \neq 0$. If $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta} = 0$, then $\pi = 0$ which contradicts $\pi \neq 0$. Therefore (S2.3) holds, is well-defined and $\pi \neq 0$. Now plug the π back into (S2.1) and we have

$$\begin{aligned} 0 &= \boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{\Sigma}\boldsymbol{\beta}\boldsymbol{\mu}^{\mathrm{T}}\right) - \boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{\mu}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}\right) + \left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta}\right)^{-1}\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta}\right)\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} \\ &= \boldsymbol{\beta}^{\mathrm{T}}U - \boldsymbol{\beta}^{\mathrm{T}}U^{\mathrm{T}} + W, \text{ say,} \end{aligned} \tag{S2.4}$$

and thus $\boldsymbol{\beta}^{\mathrm{T}}U = \boldsymbol{\beta}^{\mathrm{T}}U^{\mathrm{T}} - W$. By multiplying this equation from the right by $\boldsymbol{\beta}$ and applying it again, we have

$$\begin{aligned} \boldsymbol{\beta}^{\mathrm{T}}U\boldsymbol{\beta} &= \boldsymbol{\beta}^{\mathrm{T}}U^{\mathrm{T}}\boldsymbol{\beta} - W\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{\beta}^{\mathrm{T}}U\right)^{\mathrm{T}} - W\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}}\left(\boldsymbol{\beta}^{\mathrm{T}}U^{\mathrm{T}} - W\right)^{\mathrm{T}} - W\boldsymbol{\beta} \\ &= \boldsymbol{\beta}^{\mathrm{T}}U\boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}}W^{\mathrm{T}} - W\boldsymbol{\beta}\,, \end{aligned}$$

i.e. $0 = W\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}W^{\mathrm{T}} = \left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}\boldsymbol{\beta}\right)^{-1}\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta}\right)\left(\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right)$. Since $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\beta} \neq 0$, we have $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = 0$. Its solution is of the form $\boldsymbol{\beta} = \boldsymbol{\Sigma}(V - V^{\mathrm{T}})\boldsymbol{\mu}$ with V a $p \times p$ matrix. Plugging this $\boldsymbol{\beta}$ back into (S2.4), we have

$$\begin{aligned} 0 &= \boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^{3}(V - V^{\mathrm{T}})\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^{2} \\ &\quad + \left\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(V - V^{\mathrm{T}})\boldsymbol{\mu}\right\}^{-1}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}(V - V^{\mathrm{T}})\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

By the arbitrariness of $\boldsymbol{\mu}$, the above equation gives

$$\begin{aligned} 0 \ &= (V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^3(V - V^{\mathrm{T}})\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - (V^{\mathrm{T}} - V)\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^2 \\ &\quad + \left\{\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(V - V^{\mathrm{T}})\boldsymbol{\mu}\right\}^{-1}\boldsymbol{\Sigma}(V - V^{\mathrm{T}})\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

Multiplying it by $\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}$ from the left, we have

$$0 \ = \ \boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}(V^{\mathrm{T}}-V)\boldsymbol{\Sigma}^3(V-V^{\mathrm{T}})\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}-\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-2}(V^{\mathrm{T}}-V)\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}}-V)\boldsymbol{\Sigma}^2+\boldsymbol{\mu}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1},$$

from which, by the arbitrariness of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ again, we have

$$0 \ = \ \boldsymbol{\Sigma}^{-1}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^3(V - V^{\mathrm{T}})\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} - \boldsymbol{\Sigma}^{-1}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^2 + I.$$

Thus

$$I \ = \ \boldsymbol{\Sigma}^{-1}(V^{\mathrm{T}}-V)\boldsymbol{\Sigma}\left[\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^2 - \left\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^2\right\}^{\mathrm{T}}\right] \ = \ \boldsymbol{\Sigma}^{-1}(V^{\mathrm{T}}-V)\boldsymbol{\Sigma}S, \text{ say.}$$

If we take transpose it follows that

$$\begin{aligned} I \ &= \ S^{\mathrm{T}}\boldsymbol{\Sigma}(V - V^{\mathrm{T}})\boldsymbol{\Sigma}^{-1} \\ &= \ -S\boldsymbol{\Sigma}\left\{-(V^{\mathrm{T}} - V)\right\}\boldsymbol{\Sigma}^{-1} \\ &= \ S\boldsymbol{\Sigma}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^{-1}. \end{aligned}$$

Combining the above two equations gives $\boldsymbol{\Sigma}^{-1}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}S = S\boldsymbol{\Sigma}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^{-1}$ which has the solutions $S = \boldsymbol{\Sigma}^{-2}$ and $S = 0$. If $S = \boldsymbol{\Sigma}^{-2}$, then by the fact $S^{\mathrm{T}} = -S$ we have $\boldsymbol{\Sigma}^{-2} = -\boldsymbol{\Sigma}^{-2}$ which is not valid. Thus we have $S = 0$. From this we have $\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^2 = \left\{\boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}}(V^{\mathrm{T}} - V)\boldsymbol{\Sigma}^2\right\}^{\mathrm{T}}$ which has solutions $V^{\mathrm{T}} - V = \boldsymbol{\Sigma}^{-2}$ and $V^{\mathrm{T}} - V = 0$. If $V^{\mathrm{T}} - V = \boldsymbol{\Sigma}^{-2}$,

then $(V^{\mathrm{T}} - V)^{\mathrm{T}} = (\boldsymbol{\Sigma}^{-2})^{\mathrm{T}}$, i.e. $V - V^{\mathrm{T}} = \boldsymbol{\Sigma}^{-2}$, i.e. $-\boldsymbol{\Sigma}^{-2} = \boldsymbol{\Sigma}^{-2}$ is a contradiction. Thus $V^{\mathrm{T}} - V = 0$ which gives $\boldsymbol{\beta} = \boldsymbol{\Sigma}(V - V^{\mathrm{T}})\boldsymbol{\mu} = 0$.

Therefore $\boldsymbol{\beta}_{02} = 0$ for general $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. For some specific $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, there might exist nonzero $\boldsymbol{\beta}_{02}$, but always $\boldsymbol{\beta}_{02}^{\mathrm{T}}\boldsymbol{\mu} = 0$. This gives the conclusion that for multivariate normal, the true $d = 1$ and only the first $\boldsymbol{\beta}_{01}$ contributes to the variation among different classes.

□

S3 Proof of Corollary 3

Proof. From $Y \perp\!\!\!\perp \mathbf{X}|\mathbf{BX}$ and $\mathbf{B}\boldsymbol{\gamma} = \mathbf{0}$, we get $Y \perp\!\!\!\perp \boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}|\mathbf{BX}$ (Proposition 4.3 in Cook (1998)). If we also have $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X} \perp\!\!\!\perp \mathbf{BX}$, then by Lemma 4.3 of Dawid (1979) and Proposition 4.6 of Cook (1998), we obtain $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X} \perp\!\!\!\perp Y$ which implies $\mathrm{MV}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}|Y) = 0$ according to the property of the MV index. When $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X} \perp\!\!\!\perp \mathbf{BX}$ if and only if

$$\mathrm{Cov}(\boldsymbol{\gamma}^{\mathrm{T}}\mathbf{X}, \mathbf{BX}) = \boldsymbol{\gamma}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B}^{\mathrm{T}} = \mathbf{0}\,. \tag{S3.1}$$

If $\boldsymbol{\Sigma} = \mathbf{I}$, then (S3.1) holds by $\mathbf{B}\boldsymbol{\gamma} = \mathbf{0}$. Let $\boldsymbol{\Gamma}$ denote the vector space spanned by all $\boldsymbol{\gamma}$ satisfying (S3.1) and $\mathbf{B}\boldsymbol{\gamma} = \mathbf{0}$, then we can easily get $\dim(\boldsymbol{\Gamma}) = p - k$ which implies $d = k$ for d defined in Theorem 1. If $\boldsymbol{\Sigma} \neq \mathbf{I}$ and $2k \leq p$, then $\dim(\boldsymbol{\Gamma}) \geq p - 2k$ by noting that $\boldsymbol{\Sigma}$ is positive definite.

Hence $d \leq 2k$. We complete the proof. □

S4 Proof of Proposition 4

The following lemma is needed.

Lemma 2. *Under Conditions 1, 3 and 4, for any $\boldsymbol{\beta} \in B(\kappa_{01})$, it holds that*

(1) *if R is fixed, $\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y) \to \mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ in probability as $n \to \infty$.*

(2) *if R is diverging with n and satisfies $R = O(n^{\delta})$ with $0 < \delta \leq 1/2$, $\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y) \to \mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ in probability as $n \to \infty$.*

Proof. Denote $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ by X with support $\mathbb{R}_X$ and the transformed samples $\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j\}_{j=1}^n$ by $\{X_j\}_{j=1}^n$. By the definitions of $\mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ and $\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$,

we have

$$
\begin{aligned}
&\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y) - \mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y) = \mathrm{MV}_n(X|Y) - \mathrm{MV}(X|Y) \\
&= \frac{1}{n}\sum_{j=1}^{n}\sum_{r=1}^{R}\hat{p}_r\{\hat{F}_{hr}(X_j) - \hat{F}_h(X_j)\}^2 - \sum_{r=1}^{R}p_r\int\{F_r(x) - F(x)\}^2\mathrm{d}F(x) \\
&= \sum_{r=1}^{R}\hat{p}_r\int\{\hat{F}_{hr}(x) - \hat{F}_h(x)\}^2\mathrm{d}\hat{F}(x) - \sum_{r=1}^{R}p_r\int\{F_r(x) - F(x)\}^2\mathrm{d}F(x) \\
&= \sum_{r=1}^{R}\hat{p}_r\left[\int\{\hat{F}_{hr}(x) - \hat{F}_h(x)\}^2\mathrm{d}\hat{F}(x) - \int\{F_r(x) - F(x)\}^2\mathrm{d}F(x)\right] \\
&\quad + \sum_{r=1}^{R}(\hat{p}_r - p_r)\int\{F_r(x) - F(x)\}^2\mathrm{d}F(x) \\
&= \sum_{r=1}^{R}\hat{p}_r\int[\{\hat{F}_{hr}(x) - \hat{F}_h(x)\}^2 - \{F_r(x) - F(x)\}^2]\mathrm{d}\hat{F}(x) \\
&\quad + \sum_{r=1}^{R}\hat{p}_r\int\{F_r(x) - F(x)\}^2\mathrm{d}\{\hat{F}(x) - F(x)\} \\
&\quad + \sum_{r=1}^{R}(\hat{p}_r - p_r)\int\{F_r(x) - F(x)\}^2\mathrm{d}F(x) \\
&=: A_1 + A_2 + A_3\,.
\end{aligned}
$$

For the first term A_1,

$$
\begin{aligned}
|A_1| \;&\le 2\max_{1\le r\le R}\int|\{\hat{F}_{hr}(x) - F_r(x)\} - \{\hat{F}_h(x) - F(x)\}|\mathrm{d}\hat{F}(x) \\
&\le 2\max_{1\le r\le R}\sup_{x\in\mathbb{R}_X}\{|\hat{F}_{hr}(x) - F_r(x)| + |\hat{F}_h(x) - F(x)|\} \\
&=: 2(B_1 + B_2)\,,
\end{aligned}
$$

where the second inequality is obtained by $\int\mathrm{d}\hat{F}(x) = 1$. We then consider

the term B_1,

$$B_1 = \max_{1\le r\le R} \sup_{x\in\mathbb{R}_X} |\hat{F}_{hr}(x) - F_r(x)| = \max_{1\le r\le R} O_{\mathrm{p}}(n_r^{-\alpha}) = O_{\mathrm{p}}(n^{-\alpha(1-\delta)}),$$

where the second equality is implied by Theorem 2.2 of Cheng (2017) with any $0 < \alpha < 1/2$ under Conditions 3 and 4, and the last equality is given by Condition 1 and the fact that $|\hat{p}_r - p_r| = O_{\mathrm{p}}(n^{-1/2})$ holds uniformly for $r = 1, \ldots, R$. For the second term B_2, also by Theorem 2.2 of Cheng (2017), we obtain $B_2 = \sup_{x\in\mathbb{R}_X} |\hat{F}_h(x) - F(x)| = O_{\mathrm{p}}(n^{-\alpha})$.

We turn to the term A_2,

$$\begin{aligned}
|A_2| &= \sum_{r=1}^{R} \hat{p}_r \int \{F_r(x) - F(x)\}^2 \mathrm{d}|\hat{F}(x) - F(x)| \\
&\le \max_r \int \{F_r(x) - F(x)\}^2 \mathrm{d}|\hat{F}(x) - F(x)| \\
&\le \int \mathrm{d}|\hat{F}(x) - F(x)| \\
&\le 2 \sup_{x\in\mathbb{R}_X} |\hat{F}(x) - F(x)| \\
&= O_{\mathrm{p}}(n^{-\alpha}),
\end{aligned}$$

where the last equality is based on the extended Glivenko-Cantelli lemma (Fabian, 1985) with α defined above.

For the last term A_3, we have $|A_3| = O_{\mathrm{p}}(n^{-\alpha})$ with any $0 < \alpha < 1/2$ by Lemma A.4 of Cui, Li and Zhong (2015). To sum up, $|\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y) - \mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)| = |A_1 + A_2 + A_3| \le O_{\mathrm{p}}(n^{-\alpha(1-\delta)}) + O_{\mathrm{p}}(n^{-\alpha}) = O_{\mathrm{p}}(n^{-\alpha(1-\delta)})$.

Thus, we complete the proof. □

We first prove $\widehat{\boldsymbol{\beta}}_1 \to_p \boldsymbol{\beta}_{01}$. From Lemma 2, we obtain that for any $\boldsymbol{\beta}_1 \in B(\kappa_1)$ with $\kappa_1 \in (0, \kappa_{01}]$, $\mathrm{MV}_n(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}|Y) \to_p \mathrm{MV}(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{X}|Y)$. For any $\epsilon > 0$, let $\{\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^M\}$ be an $\epsilon/\sqrt{p}$-net of $B(\kappa_{01})$ with $M = (2\kappa_{01}\sqrt{p}/\epsilon+1)^p$. Since M is fixed, by Lemma 2 we obtain

$$\max_{1\le j\le M} |\mathrm{MV}_n(\boldsymbol{\beta}^j) - \mathrm{MV}(\boldsymbol{\beta}^j)| \to_p 0\,,$$

as $n \to \infty$. For any $\boldsymbol{\beta} \in B(\kappa_{01})$, there exists a $m \in \{1, 2, \ldots, M\}$ such that $\|\boldsymbol{\beta}-\boldsymbol{\beta}^m\| \le \epsilon/\sqrt{p}$. Then by Condition 5, $|\mathrm{MV}_n(\boldsymbol{\beta})-\mathrm{MV}_n(\boldsymbol{\beta}^m)| = o_{\mathrm{p}}(1)$ and $|\mathrm{MV}(\boldsymbol{\beta}) - \mathrm{MV}(\boldsymbol{\beta}^m)| \to 0$. Therefore, combining the above three equations, we obtain

$$\sup_{\boldsymbol{\beta}\in B(\kappa_{01})} |\mathrm{MV}_n(\boldsymbol{\beta}) - \mathrm{MV}(\boldsymbol{\beta})| \to_p 0\,.$$

Then by Condition 2, it is easy to see that for any sufficiently small $\kappa_1 > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\beta}_1\in\partial B(\kappa_1)\cap\Gamma_1} \mathrm{MV}_n(\boldsymbol{\beta}_1) \le \mathrm{MV}_n(\boldsymbol{\beta}_{01})\right) \to 1$$

as $n \to \infty$. Thus, there exists a local maximum local point $\widehat{\boldsymbol{\beta}}_1 \in \partial B(\kappa_1)\cap\Gamma_1$ with probability approaching to 1, which means that $\mathbb{P}(\|\widehat{\boldsymbol{\beta}}_1-\boldsymbol{\beta}_{01}\| < \kappa_1) \to 1$.

Next, Let p be diverging with n and $p^{p/2}n^{-\alpha(1-\delta)} = o(1)$. Notice that Lemma 2 still holds when p is diverging, because the dimension of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ stays to be 1 whether the dimension of $\mathbf{X}$ diverges or not. Recall that

$|\mathrm{MV}_n(\boldsymbol{\beta}) - \mathrm{MV}(\boldsymbol{\beta})| = O_{\mathrm{p}}(n^{-\alpha(1-\delta)})$ (See Lemma 2). Then

$$\begin{aligned}
&\max_{1\le i\le M} |\mathrm{MV}_n(\boldsymbol{\beta}_j) - \mathrm{MV}(\boldsymbol{\beta}_j)| \\
&\le \sum_{i=1}^{M} |\mathrm{MV}_n(\boldsymbol{\beta}_j) - \mathrm{MV}(\boldsymbol{\beta}_j)| \\
&= O_{\mathrm{p}}(Mn^{-\alpha(1-\delta)}) \\
&= o_{\mathrm{p}}(1)\,.
\end{aligned}$$

By the same arguments for the fixed p case, we complete the proof.

□

S5 Proof of Theorem 5

The following lemma is needed.

Lemma 3. *Under Conditions 1, 3 and 4, for any $\boldsymbol{\beta} \in \mathbb{C}^p$, it holds that*

(1) *if R is fixed, $\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y) \to \mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ in probability as $n \to \infty$.*

(2) *if R is diverging with n and satisfies $R = O(n^{\delta})$ with $0 < \delta \le 1/2$, $\mathrm{MV}_n(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y) \to \mathrm{MV}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}|Y)$ in probability as $n \to \infty$.*

Proof. The proof is similar to that of Lemma 2. The only difference comes from the complex $\boldsymbol{\beta}$ and the corresponding complex random variable $X =: \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$ with support $\mathbb{C}_X$. Denote $X =: a + ib$ where $\mathbf{Z} := (a, b)^{\mathrm{T}}$ is a real random vector. By the definition of the cumulative distribution function of a complex variable, that is, the joint distribution function of the real part

and the imaginary part of the variable, we obtain

$$\sup_{\mathbf{z}\in\mathbb{R}^2} |\hat{F}_h(\mathbf{z}) - F(\mathbf{z})| = O_{a.s.}(n^{-1/2}(\log n)^{1/2}) .$$

This convergence rate comes from Theorem 3 of Liu and Yang (2008). Then, by the same arguments of Lemma 2, we complete the proof. □

The following gives the proof of Theorem 5.

For simplicity, we omit $i = 1$ in the subscripts of $\boldsymbol{\beta}$ and θ. Then $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \lambda)^{\mathrm{T}} \in \mathbb{R}^{p+1}$, and $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\lambda})^{\mathrm{T}}$ is the maximizer of $L_{nh}(\boldsymbol{\theta})$. Hence, $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\beta}}^{\mathrm{T}}, \hat{\lambda})^{\mathrm{T}}$ is a stationary point of $L_{nh}(\boldsymbol{\theta})$, that is, $L_{nh}^{'}(\widehat{\boldsymbol{\theta}}) = 0$. Similarly, since $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \lambda_0)^{\mathrm{T}}$ is the maximizer of $L(\boldsymbol{\theta})$, then $\boldsymbol{\theta}_0 = (\boldsymbol{\beta}_0^{\mathrm{T}}, \lambda_0)^{\mathrm{T}}$ is a stationary point of $L(\boldsymbol{\theta})$.

We then prove $n^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to_d N(0, \mathbf{V})$ where the covariance matrix $\mathbf{V}$ will be given in the proof below. By the Taylor expansion, we have

$$0 = L_{nh}^{'}(\widehat{\boldsymbol{\theta}}) = L_{nh}^{'}(\boldsymbol{\theta}_0) + L_{nh}^{''}(\boldsymbol{\theta}_0)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + R(\boldsymbol{\theta}^*) , \qquad \text{(S5.1)}$$

where $\boldsymbol{\theta}^*$ satisfies $\|\boldsymbol{\theta}^* - \boldsymbol{\theta}_0\| \le \|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$ and $\boldsymbol{\theta}^* = (\boldsymbol{\beta}^{*\mathrm{T}}, \lambda^*)^{\mathrm{T}}$. With regular calculation, we obtain

$$L_{nh}^{'}(\boldsymbol{\theta}_0) = \begin{pmatrix} \mathrm{MV}_n^{'}(\boldsymbol{\beta}_0) + 2\lambda_0\boldsymbol{\beta}_0 \\ \\ \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\beta}_0 - 1 \end{pmatrix} ,$$

and

$$L_{nh}^{''}(\boldsymbol{\theta}_0) = \begin{pmatrix} \mathrm{MV}_n^{''}(\boldsymbol{\beta}_0) + 2\lambda_0 \mathbf{I}_p & 2\boldsymbol{\beta}_0 \\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0 \end{pmatrix},$$

where $\mathbf{I}_p$ denotes the identity matrix of dimension $p \times p$. The remainder term $R(\boldsymbol{\theta}^*)$ contains the third derivative of $L_{nh}(\boldsymbol{\theta})$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^*$. Let $\mathbf{T}_n = L_{nh}^{'''}(\boldsymbol{\theta}^*)$, an array of dimension $(p+1)\times(p+1)\times(p+1)$, and for each $j = 1, \ldots, (p+1)$, $T_n(j,;,;)$ is a matrix of dimension $(p+1) \times (p+1)$. Hence, we can write

$$R(\boldsymbol{\theta}^*) = \frac{1}{2} \begin{pmatrix} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{T}_n(1,;,;)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{T}_n(2,;,;)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \\ \cdot \\ \cdot \\ \cdot \\ (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{T}_n(p+1,;,;)(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \end{pmatrix}.$$

Then, based on the explicit expressions of the derivatives given above and

(S5.1), we obtain

$$
-\begin{pmatrix} \mathrm{MV}_n^{''}(\boldsymbol{\beta}_0)+2\lambda_0\mathbf{I}_p & 2\boldsymbol{\beta}_0 \\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \times \sqrt{n}\begin{pmatrix} \mathrm{MV}_n^{'}(\boldsymbol{\beta}_0)+2\lambda_0\boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\beta}_0-1 \end{pmatrix} =
$$

$$
\left[\mathbf{I}_{p+1}+\frac{1}{2}\begin{pmatrix} \mathrm{MV}_n^{''}(\boldsymbol{\beta}_0)+2\lambda_0\mathbf{I}_p & 2\boldsymbol{\beta}_0 \\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \times \begin{pmatrix} (\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(1,;,;) \\ (\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(2,;,;) \\ \cdot \\ \cdot \\ \cdot \\ (\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(p+1,;,;) \end{pmatrix}\right]
$$

$$
\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\,. \tag{S5.2}
$$

Next, our proof is divided into two parts:

Part 1:

$$
\begin{pmatrix} \mathrm{MV}_n^{''}(\boldsymbol{\beta}_0)+2\lambda_0\mathbf{I}_p & 2\boldsymbol{\beta}_0 \\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \times \sqrt{n}\begin{pmatrix} \mathrm{MV}_n^{'}(\boldsymbol{\beta}_0)+2\lambda_0\boldsymbol{\beta}_0 \\ \boldsymbol{\beta}_0^{\mathrm{T}}\boldsymbol{\beta}_0-1 \end{pmatrix} \to N(\mathbf{0},\mathbf{V})\,,
$$

where $\mathbf{V}$ will be defined later.

Part 2:

$$\left(\mathbf{I}_{p+1} + \frac{1}{2} \begin{pmatrix} \mathrm{MV}_n^{''}(\boldsymbol{\beta}_0) + 2\lambda_0 \mathbf{I}_p & 2\boldsymbol{\beta}_0 \\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \times \begin{pmatrix} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{T}_n(1,;,;) \\ (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{T}_n(2,;,;) \\ \cdot \\ \cdot \\ \cdot \\ (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathbf{T}_n(p+1,;,;) \end{pmatrix} \right)$$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) =_d \sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\,,$$

where $=_d$ indicates convergence in distribution.

Proof of Part 1. If $\boldsymbol{\theta} = (\boldsymbol{\beta}^{\mathrm{T}}, \lambda)^{\mathrm{T}}$ is a complex vector, the above arguments still hold. By Condition 6 and Lemma 3, it follows from the convergence of analytic functions that

$$L_{nh}^{''}(\boldsymbol{\theta}_0)^{-1} = \begin{pmatrix} \mathrm{MV}_n^{''}(\boldsymbol{\beta}_0) + 2\lambda_0 \mathbf{I}_p & 2\boldsymbol{\beta}_0 \\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \to_p L^{''}(\boldsymbol{\theta}_0)^{-1} =: \mathbf{A}\,. \quad \text{(S5.3)}$$

Denote $\boldsymbol{\beta} = (b_1, \ldots, b_p)^{\mathrm{T}}$ and $T_{jn} = \sqrt{n}\partial \mathrm{MV}_n(\boldsymbol{\beta})/\partial b_j|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}$. Notice that $\boldsymbol{\beta}_0^T \boldsymbol{\beta}_0 - 1 = 0$. Then it suffices to prove the asymptotic normality of T_{jn} for each $j = 1, \ldots, p$. It is worth noting that the elements of $-\mathbf{A}\sqrt{n}L_{nh}^{'}(\boldsymbol{\theta}_0)$ are linear combinations of T_{jn}s. We consider the $j = 1$ case.

By the definition of $\mathrm{MV}_n(\boldsymbol{\beta})$, we have

$$T_{1n} = -\sqrt{n}\bigg(\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial b_1}\bigg[\sum_{r=1}^{R}\hat{p}_r\{\hat{F}_{hr}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) - \hat{F}_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}^2\bigg]\bigg)\bigg|_{\boldsymbol{\beta}=\boldsymbol{\beta}_0}.$$

By Lemma 3,

$$\begin{aligned}&\sum_{i=1}^{n}\bigg[\sum_{r=1}^{R}\hat{p}_r\{\hat{F}_{hr}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) - \hat{F}_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}^2\bigg]\\ =&\sum_{i=1}^{n}\bigg[\sum_{r=1}^{R}p_r\{F_r(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) - F(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\}^2\bigg]\{1 + u_n(\boldsymbol{\beta}) + iv_n(\boldsymbol{\beta})\},\end{aligned}$$

where $\boldsymbol{\beta} \in C(\kappa_0)$, $i^2 = -1$, and $u_n(\boldsymbol{\beta}) + iv_n(\boldsymbol{\beta}) = o_p(1)$ is uniform in $\boldsymbol{\beta} \in C(\kappa_0)$ when $n \to \infty$ with $u_n(\boldsymbol{\beta})$ and $v_n(\boldsymbol{\beta})$ being real functions of $\boldsymbol{\beta}$.

By Cauchy's residue theorem, we have

$$T_{1n} = \frac{1}{\sqrt{n}}\frac{1}{2\pi i}\oint_{C_1}\frac{\sum_{i=1}^{n}\sum_{r=1}^{R}\hat{p}_r\{\hat{F}_{hr}(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i) - \hat{F}_h(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}^2}{(b_1 - b_{01})^2}\mathrm{d}b_1\,,$$

where $\tilde{\boldsymbol{\beta}} = (b_1, b_{02}, \ldots, b_{0p})^{\mathrm{T}}$ with $\boldsymbol{\beta}_0 = (b_{01}, \ldots, b_{0p})^{\mathrm{T}}$, and C_1 satisfies $\{b_1 \in \mathbb{C} : \|b_1 - b_{01}\| = r\}$ with $r < \kappa_0$. Define

$$S_{1n} = \frac{1}{\sqrt{n}}\frac{1}{2\pi i}\oint_{C_1}\frac{\sum_{i=1}^{n}\sum_{r=1}^{R}p_r\{F_r(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i) - F(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}^2}{(b_1 - b_{01})^2}\mathrm{d}b_1\,.$$

Then,

$$T_{1n} - S_{1n} = \frac{1}{\sqrt{n}}\frac{1}{2\pi i}\oint_{C_1}\frac{\sum_{i=1}^{n}\sum_{r=1}^{R}p_r\{F_r(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i) - F(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}^2\{u_n(\tilde{\boldsymbol{\beta}}) + iv_n(\tilde{\boldsymbol{\beta}})\}}{(b_1 - b_{01})^2}\mathrm{d}b_1. \tag{S5.4}$$

Let

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\sum_{r=1}^{R}p_r\{F_r(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i) - F(\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i)\}^2 =: R_n(b_1) + iI_n(b_1)\,.$$

Noting that the left-hand side of (S5.4) is real, we consider the real part of the other hand, that is

$$\frac{1}{2\pi}\int_0^{2\pi}\frac{(R_n\cos\mu+I_n\sin\mu)u_n+(R_n\sin\mu-I_n\cos\mu)v_n}{r}\mathrm{d}\mu\,, \tag{S5.5}$$

where the arguments of R_n, I_n, u_n, v_n are $b_{01}+re^{i\mu}$. By the mean value theorem, we obtain

$$(S5.5)=\frac{(R_{0n}\cos\mu_0+I_{0n}\sin\mu_0)u_{0n}+(R_{0n}\sin\mu_0-I_{0n}\cos\mu_0)v_{0n}}{r}\,, \tag{S5.6}$$

where the arguments of $R_{0n}, I_{0n}, u_{0n}, v_{0n}$ are all $b_{01}+re^{i\mu_0}$ with $\mu_0\in[0,2\pi]$. By the definition of S_{1n}, we have

$$S_{1n}=-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{\partial}{\partial b_1}\sum_{r=1}^{R}p_r\{F_r(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)-F(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\}^2=-\frac{\partial}{\partial b_1}\{R_n(b_{01})+iI_n(b_{01})\}\,. \tag{S5.7}$$

By central limit theorem, S_{1n} is asymptotically normally distributed. Then, noticing $u_n(\boldsymbol{\beta})+iv_n(\boldsymbol{\beta})=o_{\mathrm{p}}(1)$ for $\boldsymbol{\beta}\in C(\kappa_0)$ and letting $r\to 0$, $|T_{1n}-S_{1n}|\le 2/r(|R_{0n}|+|I_{0n}|)(|u_{0n}|+|v_{0n}|)=o_{\mathrm{p}}(1)$ as $n\to\infty$. Using Slutsky's theorem, we complete the proof for the asymptotic normality of T_{1n}. From (S5.7), we can further derive the form of $\mathbf{V}$:

$$\mathbf{V}=\begin{pmatrix}\mathbf{V}_1 & * \\ * & *\end{pmatrix}\,,$$

where $\mathbf{V}_1 = \mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_1 \in \mathbb{R}^{p\times p}$,

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_1 & * \\ * & * \end{pmatrix} \tag{S5.8}$$

for $\mathbf{A}$ defined in (S5.3), $\boldsymbol{\Sigma} = \mathbb{E}\{\alpha(\mathbf{X}_i)\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}}\}+4\lambda_0\boldsymbol{\beta}_0\mathbb{E}\{\alpha(\mathbf{X}_i)\mathbf{X}_i^{\mathrm{T}}\}+4\lambda_0^2\boldsymbol{\beta}_0\boldsymbol{\beta}_0^{\mathrm{T}}$ with $\alpha(\mathbf{X}_i) = 2\sum_{r=1}^{R} p_r\{F(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i) - F_r(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\}\{f(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i) - f_r(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i)\}$. We complete the proof of Part 1.

Proof of Part 2. By Conditions 6-7, we can easily obtain

$$\mathbf{I}_{p+1}+\frac{1}{2}\begin{pmatrix} \mathrm{MV}_n^{''}(\boldsymbol{\beta}_0) + 2\lambda_0\mathbf{I}_p & 2\boldsymbol{\beta}_0 \\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0 \end{pmatrix}^{-1} \times \begin{pmatrix} (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(1,;,;) \\ (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(2,;,;) \\ \cdot \\ \cdot \\ \cdot \\ (\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(p+1,;,;) \end{pmatrix} \to_p \mathbf{I}_{p+1}\,.$$

Hence, by Slutsky's theorem,

$$\left[\mathbf{I}_{p+1}+\frac{1}{2}\begin{pmatrix}\mathrm{MV}_n^{''}(\boldsymbol{\beta}_0)+2\lambda_0\mathbf{I}_p & 2\boldsymbol{\beta}_0\\ 2\boldsymbol{\beta}_0^{\mathrm{T}} & 0\end{pmatrix}^{-1}\times\begin{pmatrix}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(1,;,;)\\ (\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(2,;,;)\\ \cdot\\ \cdot\\ \cdot\\ (\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)^{\mathrm{T}}\mathbf{T}_n(p+1,;,;)\end{pmatrix}\right]$$

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)=_d\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\,.$$

Therefore, combining Part 1 and Part 2, we have $\sqrt{n}(\widehat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0)\to_d N(\mathbf{0},\mathbf{V})$. Because $\widehat{\boldsymbol{\theta}}=(\widehat{\boldsymbol{\beta}}^{\mathrm{T}},\hat{\lambda})^{\mathrm{T}}$, by the property of multivariate normal distribution, we complete the proof for the case $i=1$ with the covariance matrix $\mathbf{V}_1$ being the $p\times p$ sub-matrix at the top right-hand corner of $\mathbf{V}$. □

References

Cheng, F. (2017). Strong uniform consistency rates of kernel estimators of cumulative distribution functions. *Communications in Statistics-Theory and Methods*, **46(14)**, 6803-6807.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177–189.

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

REFERENCES

Cui, H., Li, R. and Zhong, W. (2015). Model-Free Feature Screening for Ultrahigh Dimensional Discriminant Analysis. *Journal of the American Statistical Association*, **110**, 630–641.

Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society*, **B**, 1-31.

Fabian, V. (1985). *Introduction to probability and mathematical statistics*. John Wiley and Sons Incorporated.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316–342.

Liu, R. and Yang, L. (2008). Kernel estimation of multivariate cumulative distribution function. *Journal of Nonparametric Statistics*, **20(8)**, 661-677.

Tan, K., Shi, L., & Yu, Z. (2020). Sparse SIR: Optimal rates and adaptive estimation. *The Annals of Statistics*, **48**, 64–85.