DIMENSION REDUCTION VIA ADAPTIVE SLICING

Tao Wang

Shanghai Jiao Tong University

Supplementary Material

THE ONLINE SUPPLEMENTARY MATERIAL CONTAINS ADDITIONAL SIMULATIONS AND ALL PROOFS.

S1 Additional simulations

Example A1 for SIR. We first generated $\boldsymbol{X}$ from a multivariate Gaussian distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{ij})$ with $\Sigma_{ij} = 0.5^{|i-j|}$. We then generated Y according to the following model:

$$Y = \sin(\boldsymbol{\eta}^\top \boldsymbol{X} + \epsilon), \tag{S1-1}$$

where $\boldsymbol{\eta} = (1, 0, \ldots, 0)^\top \in \mathbb{R}^{p\times 1}$, and ϵ is standard normal and is independent of $\boldsymbol{X}$. In this example $\mathcal{S}_{Y|\boldsymbol{X}} = \text{span}(\boldsymbol{\eta})$, and the optimal slicing scheme does not exist.

Example A2 for SAVE. We first generated $\boldsymbol{X}$ from a multivariate Gaussian distribution with mean vector zero and covariance matrix $\boldsymbol{\Sigma} =$

(Σ_{ij}) with $\Sigma_{ij} = 0.5^{|i-j|}$. We then generated Y according to the following model:

$$Y = (\boldsymbol{\eta}^\top \boldsymbol{X})^2 + \epsilon, \tag{S1-2}$$

where $\boldsymbol{\eta} = (1, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^{p\times 1}$, and ϵ is standard normal and is independent of $\boldsymbol{X}$. In this example $\mathcal{S}_{Y|\boldsymbol{X}} = \text{span}(\boldsymbol{\eta})$, and the optimal slicing scheme does not exist.

Example A3 for SAVE. We first simulated Y uniformly on the interval $[0, 5]$. Given $Y = y$, we then generated $\boldsymbol{X}$ from the model

$$\boldsymbol{X} = \boldsymbol{\eta}_1 \mathbf{C}\boldsymbol{h}(y) + 0.5\boldsymbol{\varepsilon} + 0.3s(y)\boldsymbol{\eta}_2\epsilon, \tag{S1-3}$$

where $\boldsymbol{\eta}_1 = (1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^{p\times 1}, \boldsymbol{\eta}_2 = (0, 0, 1, 1, 0, \ldots, 0)^\top \in \mathbb{R}^{p\times 1}, \mathbf{C} = (2, -2, \ldots, 2, -2) \in \mathbb{R}^{1\times G_0}$, $\boldsymbol{h}(y) \in \mathbb{R}^{G_0\times 1}$ is a vector of slice indicator functions, and $(\boldsymbol{\varepsilon}^\top, \epsilon)^\top \in \mathbb{R}^{p+1}$ is multivariate Gaussian with zero mean and identity covariance matrix and is independent of Y. We set $G_0 = 10$ and constructed $\boldsymbol{h}$ via quantile slicing of observed responses with G_0 slices. Let $\mathcal{S}_g$ denote the gth slice. To specify a heteroscedastic error structure, we define $s(y) = g$ if $y \in \mathcal{S}_{2g-1} \cup \mathcal{S}_{2g}$, for $g = 1, \ldots, 5$. By Proposition 3.2, $\mathcal{S}_{Y|\boldsymbol{X}} = \text{span}(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$. In this example, there is an optimal slicing scheme in location and scale: G_0 slices with equal number of observations in each slice.

Table S1-1: Means and standard deviations (in parentheses) of the vector correlation coefficient for SIR-AS and its various competitors, based on 200 data applications, are reported for Example A1.

SIR				FSIR			
$G = 5$	$G = 10$	$G = 20$	CUME	$H = 10$	$H = 20$	$H = 30$	SIR-AS
0.949	0.948	0.944	0.953	0.951	0.950	0.949	0.945
(0.027)	(0.030)	(0.033)	(0.024)	(0.025)	(0.026)	(0.028)	(0.029)

Table S1-2: Means and standard deviations (in parentheses) of the vector correlation coefficient for SAVE-AS and its various competitors, based on 200 data applications, are reported for Examples A2 and A3.

	SAVE				FSAVE			
Model	$G = 5$	$G = 10$	$G = 20$	CUVE	$H = 10$	$H = 20$	$H = 30$	SAVE-AS
(S1-2)	0.969	0.960	0.947	0.983	0.970	0.960	0.954	0.954
	(0.015)	(0.021)	(0.028)	(0.008)	(0.016)	(0.023)	(0.025)	(0.024)
(S1-3)	0.030	0.786	0.698	0.036	0.214	0.533	0.507	0.772
	(0.024)	(0.081)	(0.115)	(0.026)	(0.145)	(0.134)	(0.151)	(0.086)

S2 Appendix

PROOF OF LEMMA 3.1. This is a corollary of the Courant–Fischer theorem.

PROOF OF PROPOSITION 3.1. Consider first the least squares loss function $L_{SIR}(\mathbf{B}, \mathbf{C})$. For fixed $\mathbf{B}$, the minimizer is $\hat{\mathbf{C}}_g = \mathbf{B}^\top \hat{\boldsymbol{\mu}}_g$, and the minimum is

$$\begin{aligned} L_{SIR}(\mathbf{B}, \hat{\mathbf{C}}) &= \sum_{g=1}^{G} \frac{n_g}{n} \|\hat{\boldsymbol{\mu}}_g - \mathbf{B}\mathbf{B}^\top \hat{\boldsymbol{\mu}}_g\|_2^2 \\ &= \sum_{g=1}^{G} \frac{n_g}{n} \mathrm{trace}\{(\mathbf{I}_p - \mathbf{B}\mathbf{B}^\top)\hat{\boldsymbol{\mu}}_g \hat{\boldsymbol{\mu}}_g^\top\} \\ &= \mathrm{trace}(\hat{\mathbf{M}}_{SIR}) - \mathrm{trace}(\mathbf{B}^\top \hat{\mathbf{M}}_{SIR} \mathbf{B}). \end{aligned}$$

Thus, minimizing $L_{SIR}(\mathbf{B}, \hat{\mathbf{C}})$ over $\mathbf{B} \in \mathcal{G}_{p,d}$ is equivalent to maximizing $\mathrm{trace}(\boldsymbol{\alpha}^\top \hat{\mathbf{M}}_{SIR} \boldsymbol{\alpha})$ over $\boldsymbol{\alpha} \in \mathcal{G}_{p,d}$.

Consider now $L_{SAVE}(\mathbf{B}, \mathbf{F})$. For fixed $\mathbf{B}$, the minimizer is $\hat{\mathbf{F}}_g = \mathbf{B}^\top(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_g)$, and the minimum is

$$\begin{aligned} L_{SAVE}(\mathbf{B}, \hat{\mathbf{F}}) &= \sum_{g=1}^{G} \frac{n_g}{n} \|\mathrm{vec}(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_g) - \mathrm{vec}\{\mathbf{B}\mathbf{B}^\top(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_g)\}\|_2^2 \\ &= \sum_{g=1}^{G} \frac{n_g}{n} \mathrm{trace}\{(\mathbf{I}_p - \mathbf{B}\mathbf{B}^\top)(\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_g)^2\} \\ &= \mathrm{trace}(\hat{\mathbf{M}}_{SAVE}) - \mathrm{trace}(\mathbf{B}^\top \hat{\mathbf{M}}_{SAVE} \mathbf{B}). \end{aligned}$$

Thus, minimizing $L_{SAVE}(\mathbf{B}, \hat{\mathbf{F}})$ over $\mathbf{B} \in \mathcal{G}_{p,d}$ is equivalent to maximizing $\mathrm{trace}(\boldsymbol{\alpha}^\top \hat{\mathbf{M}}_{SAVE} \boldsymbol{\alpha})$ over $\boldsymbol{\alpha} \in \mathcal{G}_{p,d}$. The proof is complete.

LEMMA S2.1. *Let $\mathbf{A}$ be a $p \times d$ semi-orthogonal matrix, and let $\mathbf{A}_0$ be an orthogonal complement of $\mathbf{A}$ such that $(\mathbf{A}, \mathbf{A}_0)$ is $p \times p$ orthogonal. Then, for any $p \times p$ positive definite matrix $\mathbf{B}$,* $\det(\mathbf{A}^\top \mathbf{B} \mathbf{A}) = \det(\mathbf{B}) \det(\mathbf{A}_0^\top \mathbf{B}^{-1} \mathbf{A}_0)$.

Proof of Lemma S2.1. Note that

$$
\begin{aligned}
\det(\mathbf{B}) &= \det\{(\mathbf{A},\mathbf{A}_0)^\top \mathbf{B}(\mathbf{A},\mathbf{A}_0)\} \\
&= \det(\mathbf{A}^\top\mathbf{B}\mathbf{A})\det\{\mathbf{A}_0^\top\mathbf{B}\mathbf{A}_0 - \mathbf{A}_0^\top\mathbf{B}\mathbf{A}(\mathbf{A}^\top\mathbf{B}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{B}\mathbf{A}_0\}.
\end{aligned}
$$

It is easy to show that

$$\mathbf{B} = \mathbf{A}_0(\mathbf{A}_0^\top\mathbf{B}^{-1}\mathbf{A}_0)^{-1}\mathbf{A}_0^\top + \mathbf{B}\mathbf{A}(\mathbf{A}^\top\mathbf{B}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{B}.$$

Hence,

$$(\mathbf{A}_0^\top\mathbf{B}^{-1}\mathbf{A}_0)^{-1} = \mathbf{A}_0^\top\mathbf{B}\mathbf{A}_0 - \mathbf{A}_0^\top\mathbf{B}\mathbf{A}(\mathbf{A}^\top\mathbf{B}\mathbf{A})^{-1}\mathbf{A}^\top\mathbf{B}\mathbf{A}_0.$$

Consequently, $\det(\mathbf{B}) = \det(\mathbf{A}^\top\mathbf{B}\mathbf{A})\det\{(\mathbf{A}_0^\top\mathbf{B}^{-1}\mathbf{A}_0)^{-1}\}$. The proof is complete.

Proof of Proposition 3.2. By Proposition 2 of Cook and Forzani (2009), $\text{span}(\boldsymbol{\eta}) = \mathcal{S}_{Y|\boldsymbol{X}}$. If we estimate the unknown parameters by maximum likelihood, then Theorem 2 of Cook and Forzani (2009) shows that the profile log-likelihood function takes the form

$$l(\boldsymbol{\eta}) = c - \frac{1}{2}\sum_{g=1}^{G} n_g \log\det(\boldsymbol{\eta}^\top\mathbf{S}_g\boldsymbol{\eta}) + \frac{n}{2}\log\det(\boldsymbol{\eta}^\top\mathbf{S}\boldsymbol{\eta}),$$

where c is an irrelevant constant. Hence, at the population level, $\boldsymbol{\eta}$ minimizes

$$\sum_{g=1}^{G}\pi_g \log\det\{\boldsymbol{\eta}^\top\text{Cov}(\boldsymbol{X} \mid Y = g)\boldsymbol{\eta}\} - \log\det\{\boldsymbol{\eta}^\top\text{Cov}(\boldsymbol{X})\boldsymbol{\eta}\}.$$

From Proposition 3 of Cook and Forzani (2009), we know that $\mathrm{span}(\boldsymbol{\eta}) = \mathcal{S}_{SAVE}$. This completes the first part of the proof.

Assume for now that $\boldsymbol{X} \mid (Y = g) \sim N(\boldsymbol{\mu}_g, \boldsymbol{\Sigma})$. One can show that the corresponding profile log-likelihood function

$$l(\boldsymbol{\eta}) = c - \frac{n}{2} \log\det(\boldsymbol{\eta}^\top \mathbf{S}_W \boldsymbol{\eta}) + \frac{n}{2} \log\det(\boldsymbol{\eta}^\top \mathbf{S} \boldsymbol{\eta}),$$

where $\mathbf{S}_W = \sum_{g=1}^{G} \sum_{i:y_i=g} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_g)^\top / n$, and c is an unimportant constant. Consequently, at the population level, $\boldsymbol{\eta}$ minimizes

$$\log\det[\boldsymbol{\eta}^\top \mathrm{E}\{\mathrm{Cov}(X \mid Y)\}\boldsymbol{\eta}] - \log\det\{\boldsymbol{\eta}^\top \mathrm{Cov}(\boldsymbol{X})\boldsymbol{\eta}\}.$$

By Lemma S2.1,

$$\begin{aligned}
& \log\det[\boldsymbol{\eta}^\top \mathrm{E}\{\mathrm{Cov}(X \mid Y)\}\boldsymbol{\eta}] - \log\det\{\boldsymbol{\eta}^\top \mathrm{Cov}(\boldsymbol{X})\boldsymbol{\eta}\} \\
= \ & \log\det[\mathrm{E}\{\mathrm{Cov}(X \mid Y)\}] + \log\det(\boldsymbol{\eta}_0^\top [\mathrm{E}\{\mathrm{Cov}(X \mid Y)\}]^{-1} \boldsymbol{\eta}_0) \\
& - \log\det\{\mathrm{Cov}(\boldsymbol{X})\} - \log\det[\boldsymbol{\eta}_0^\top \{\mathrm{Cov}(\boldsymbol{X})\}^{-1} \boldsymbol{\eta}_0] \\
\geq \ & \log\det[\mathrm{E}\{\mathrm{Cov}(X \mid Y)\}] - \log\det\{\mathrm{Cov}(\boldsymbol{X})\},
\end{aligned}$$

where the inequality follows from the fact that $\mathrm{E}\{\mathrm{Cov}(X \mid Y)\} \leq \mathrm{Cov}(\boldsymbol{X})$. Let $\boldsymbol{\alpha}$ be a basis matrix for $\mathcal{S}_{SIR}$. It suffices to show that $\boldsymbol{\alpha}_0^\top [\mathrm{E}\{\mathrm{Cov}(X \mid Y)\}]^{-1} \boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_0^\top \{\mathrm{Cov}(\boldsymbol{X})\}^{-1} \boldsymbol{\alpha}_0$, where $\boldsymbol{\alpha}_0$ is an orthogonal complement of $\boldsymbol{\alpha}$ such that $(\boldsymbol{\alpha}, \boldsymbol{\alpha}_0)$ is $p \times p$ orthogonal.

Write $\mathbf{A} = \mathrm{E}\{\mathrm{Cov}(X \mid Y)\}$ and $\mathbf{B} = \mathrm{Cov}(\boldsymbol{X})$. We have $\mathbf{A} \leq \mathbf{B}$.

Furthermore,

$$\mathrm{span}(\mathbf{B}^{-1/2}\mathbf{A}^{1/2}) = \mathrm{span}(\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2}) = \mathrm{span}(\mathbf{M}_{SIR}) = \mathbf{B}^{1/2}\mathcal{S}_{SIR}.$$

It follows that $\mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2} = \mathbf{B}^{1/2}\boldsymbol{\alpha}\mathbf{C}\boldsymbol{\alpha}^\top\mathbf{B}^{1/2}$, where $\mathbf{C}$ is a $d \times d$ positive definite matrix. Hence,

$$\begin{aligned}
\mathbf{A} &= \mathbf{B} - (\mathbf{B} - \mathbf{A}) \\
&= \mathbf{B} - \mathbf{B}^{1/2}(\mathbf{I}_p - \mathbf{B}^{-1/2}\mathbf{A}\mathbf{B}^{-1/2})\mathbf{B}^{1/2} \\
&= \mathbf{B} - \mathbf{B}^{1/2}(\mathbf{I}_p - \mathbf{B}^{1/2}\boldsymbol{\alpha}\mathbf{C}^{1/2}\mathbf{C}^{1/2}\boldsymbol{\alpha}^\top\mathbf{B}^{1/2})\mathbf{B}^{1/2}.
\end{aligned}$$

By the matrix inversion lemma,

$$(\mathbf{I}_p - \mathbf{B}^{1/2}\boldsymbol{\alpha}\mathbf{C}^{1/2}\mathbf{C}^{1/2}\boldsymbol{\alpha}^\top\mathbf{B}^{1/2})^{-1} = \mathbf{I}_p + \mathbf{B}^{1/2}\boldsymbol{\alpha}\mathbf{C}^{1/2}(\mathbf{I}_d - \mathbf{C}^{1/2}\boldsymbol{\alpha}^\top\mathbf{B}\boldsymbol{\alpha}\mathbf{C}^{1/2})^{-1}\mathbf{C}^{1/2}\boldsymbol{\alpha}^\top\mathbf{B}^{1/2}$$

and

$$\begin{aligned}
\mathbf{A}^{-1} &= \mathbf{B}^{-1} + \mathbf{B}^{-1/2}[\{(\mathbf{I}_p - \mathbf{B}^{1/2}\boldsymbol{\alpha}\mathbf{C}^{1/2}\mathbf{C}^{1/2}\boldsymbol{\alpha}^\top\mathbf{B}^{1/2})\}^{-1} - \mathbf{I}_p]\mathbf{B}^{-1/2} \\
&= \mathbf{B}^{-1} + \boldsymbol{\alpha}\mathbf{C}^{1/2}(\mathbf{I}_d - \mathbf{C}^{1/2}\boldsymbol{\alpha}^\top\mathbf{B}\boldsymbol{\alpha}\mathbf{C}^{1/2})^{-1}\mathbf{C}^{1/2}\boldsymbol{\alpha}^\top.
\end{aligned}$$

Consequently, $\boldsymbol{\alpha}_0^\top\mathbf{A}^{-1}\boldsymbol{\alpha}_0 = \boldsymbol{\alpha}_0^\top\mathbf{B}^{-1}\boldsymbol{\alpha}_0$. The proof is complete.

Proof of Theorem 4.1. It suffices to show that, as $n \to \infty$,

$$P\left\{\max_{\mathcal{S}\in\mathfrak{S}_+\cup\mathfrak{S}_-} \mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}}) < \mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}})\right\} \to 1.$$

We first consider the case of an over-slicing scheme $\mathcal{S} \in \mathfrak{S}_+$. For simplicity, assume that $\mathcal{S} = \{\mathcal{B}_{11}, \mathcal{B}_{12}, \mathcal{B}_{02}, \ldots, \mathcal{B}_{0G_0}\}$, where $\mathcal{B}_{11}$ and $\mathcal{B}_{12}$ are two sub-slices formed from $\mathcal{B}_{01}$. We have

$$\begin{aligned}\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}})-\mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}}) &= f_{\mathcal{B}_{11}}\mathrm{trace}(\tilde{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}^\top\tilde{\boldsymbol{\alpha}})+f_{\mathcal{B}_{12}}\mathrm{trace}(\tilde{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}}^\top\tilde{\boldsymbol{\alpha}})\\ &\quad -f_{\mathcal{B}_{01}}\mathrm{trace}(\tilde{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}^\top\tilde{\boldsymbol{\alpha}})-\frac{\log(n)}{n}d.\end{aligned}$$

It is easy to show that

$$f_{\mathcal{B}_{01}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}^\top=f_{\mathcal{B}_{11}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}^\top+f_{\mathcal{B}_{12}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}}^\top-\frac{f_{\mathcal{B}_{11}}f_{\mathcal{B}_{12}}}{f_{\mathcal{B}_{01}}}(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}-\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}-\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})^\top.$$

Consequently,

$$\begin{aligned}&\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}})-\mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}})\\ =\;&\frac{f_{\mathcal{B}_{11}}f_{\mathcal{B}_{12}}}{f_{\mathcal{B}_{01}}}\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}-\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}}-\hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})^\top\tilde{\boldsymbol{\alpha}}\}-\frac{\log(n)}{n}d.\end{aligned}$$

Since $\hat{\boldsymbol{\mu}}_{\mathcal{B}_{1s}}=\boldsymbol{\mu}_{\mathcal{B}_{01}}+O_P(n^{-1/2})$, $s=1,2$, we obtain

$$\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}})-\mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}})=O_P\left(\frac{1}{n}\right)-\frac{\log(n)}{n}d.$$

Similarly, we can show that this result holds for any $\mathcal{S}\in\mathfrak{S}_+$. Therefore, as $n\to\infty$,

$$P\left\{\max_{\mathcal{S}\in\mathfrak{S}_+}\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}})<\mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}})\right\}\to 1. \tag{S2-4}$$

Now consider the case where $\mathcal{S}$ is under-slicing, that is, $\mathcal{S}\in\mathfrak{S}_-$. For simplicity, assume that $\mathcal{S}=\{\mathcal{B}_{0*},\mathcal{B}_{03},\ldots,\mathcal{B}_{0G_0}\}$. Here $\mathcal{B}_{0*}$ is a new slice constructed by merging $\mathcal{B}_{01}$ and $\mathcal{B}_{02}$. We have

$$\begin{aligned}\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}})-\mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}}) &= f_{\mathcal{B}_{0*}}\mathrm{trace}(\tilde{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\mu}}_{\mathcal{B}_{0*}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{0*}}^\top\tilde{\boldsymbol{\alpha}})-f_{\mathcal{B}_{01}}\mathrm{trace}(\tilde{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}^\top\tilde{\boldsymbol{\alpha}})\\ &\quad -f_{\mathcal{B}_{02}}\mathrm{trace}(\tilde{\boldsymbol{\alpha}}^\top\hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}}^\top\tilde{\boldsymbol{\alpha}})+\frac{\log(n)}{n}d.\end{aligned}$$

Again, it is easy to see that

$$f_{\mathcal{B}_{0*}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{0*}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{0*}}^\top = f_{\mathcal{B}_{01}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}}^\top + f_{\mathcal{B}_{02}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}}\hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}}^\top - \frac{f_{\mathcal{B}_{01}}f_{\mathcal{B}_{02}}}{f_{\mathcal{B}_{0*}}}(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})^\top.$$

It follows that

$$\begin{aligned}
&\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}}) \\
= \;& -\frac{f_{\mathcal{B}_{01}}f_{\mathcal{B}_{02}}}{f_{\mathcal{B}_{0*}}}\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})^\top\tilde{\boldsymbol{\alpha}}\} + \frac{\log(n)}{n}d \\
= \;& -\frac{\pi_{\mathcal{B}_{01}}\pi_{\mathcal{B}_{02}}}{\pi_{\mathcal{B}_{0*}}}\mathrm{trace}\{\boldsymbol{\alpha}_0{}^\top(\boldsymbol{\mu}_{\mathcal{B}_{01}} - \boldsymbol{\mu}_{\mathcal{B}_{02}})(\boldsymbol{\mu}_{\mathcal{B}_{01}} - \boldsymbol{\mu}_{\mathcal{B}_{02}})^\top\boldsymbol{\alpha}_0\} + O_P\left(\frac{1}{\sqrt{n}}\right),
\end{aligned}$$

where for a slice $\mathcal{B}$, $\pi_{\mathcal{B}} = \sum_{k\in\mathcal{B}}\pi_k$. By the definition of $\mathcal{S}_0$, $\boldsymbol{\mu}_{\mathcal{B}_{01}} \neq \boldsymbol{\mu}_{\mathcal{B}_{02}}$. Hence, there exists a constant $c < 0$ such that $\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}}) < c$, with probability tending to 1 as $n \to \infty$. Together with the strategy from the first part, we can show that this result holds for any $\mathcal{S} \in \mathfrak{S}_-$. Thus, as $n \to \infty$,

$$P\left\{\max_{\mathcal{S}\in\mathfrak{S}_-}\mathrm{BIC}_1(\mathcal{S};\tilde{\boldsymbol{\alpha}}) < \mathrm{BIC}_1(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}})\right\} \to 1. \tag{S2-5}$$

Combining (S2-4) and (S2-5), the proof is complete.

Proof of Theorem 4.2. It suffices to show that, as $n \to \infty$,

$$P\left\{\max_{\mathcal{S}\in\mathfrak{S}_+\cup\mathfrak{S}_-}\mathrm{BIC}_2(\mathcal{S};\tilde{\boldsymbol{\alpha}}) < \mathrm{BIC}_2(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}})\right\} \to 1.$$

We first consider the case of an over-slicing scheme $\mathcal{S} \in \mathfrak{S}_+$. For simplicity, assume that $\mathcal{S} = \{\mathcal{B}_{11}, \mathcal{B}_{12}, \mathcal{B}_{02}, \ldots, \mathcal{B}_{0G_0}\}$, where $\mathcal{B}_{11}$ and $\mathcal{B}_{12}$ are two sub-slices formed from $\mathcal{B}_{01}$. Let $\mathrm{df}_0 = d + d(d+1)/2$. We have

$$
\begin{aligned}
& \mathrm{BIC}_2(\mathcal{S}; \tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0; \tilde{\boldsymbol{\alpha}}) \\
= \; & f_{\mathcal{B}_{11}} \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{11}})^2 \tilde{\boldsymbol{\alpha}}\} + f_{\mathcal{B}_{12}} \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{12}})^2 \tilde{\boldsymbol{\alpha}}\} \\
& - f_{\mathcal{B}_{01}} \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (\mathbf{I}_p - \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{01}})^2 \tilde{\boldsymbol{\alpha}}\} - \frac{\log(n)}{n} \mathrm{df}_0 \\
= \; & -2\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (f_{\mathcal{B}_{11}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{11}} + f_{\mathcal{B}_{12}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{12}} - f_{\mathcal{B}_{01}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{01}}) \tilde{\boldsymbol{\alpha}}\} \\
& + \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (f_{\mathcal{B}_{11}} \hat{\boldsymbol{\Sigma}}^2_{\mathcal{B}_{11}} + f_{\mathcal{B}_{12}} \hat{\boldsymbol{\Sigma}}^2_{\mathcal{B}_{12}} - f_{\mathcal{B}_{01}} \hat{\boldsymbol{\Sigma}}^2_{\mathcal{B}_{01}}) \tilde{\boldsymbol{\alpha}}\} - \frac{\log(n)}{n} \mathrm{df}_0 \\
= \; & T_1 + T_2 - \frac{\log(n)}{n} \mathrm{df}_0.
\end{aligned}
$$

It is easy to show that

$$
f_{\mathcal{B}_{01}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{01}} = f_{\mathcal{B}_{11}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{11}} + f_{\mathcal{B}_{12}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{12}} + \frac{f_{\mathcal{B}_{11}} f_{\mathcal{B}_{12}}}{f_{\mathcal{B}_{01}}} (\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})^\top \tag{S2-6}
$$

Since $\hat{\boldsymbol{\mu}}_{\mathcal{B}_{1s}} = \boldsymbol{\mu}_{\mathcal{B}_{01}} + O_P(n^{-1/2}), s = 1, 2$, we obtain

$$
T_1 = 2 \frac{f_{\mathcal{B}_{11}} f_{\mathcal{B}_{12}}}{f_{\mathcal{B}_{01}}} \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})^\top \tilde{\boldsymbol{\alpha}}\} = O_P\left(\frac{1}{n}\right).
$$

A simple calculation shows that

$$
\begin{aligned}
T_2 \; = \; & 2\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (f_{\mathcal{B}_{11}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{11}} + f_{\mathcal{B}_{12}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{12}} - f_{\mathcal{B}_{01}} \hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{01}}) \boldsymbol{\Sigma}_{\mathcal{B}_{01}} \tilde{\boldsymbol{\alpha}}\} \\
& + \mathrm{trace}[\tilde{\boldsymbol{\alpha}}^\top \{f_{\mathcal{B}_{11}} (\hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{11}} - \boldsymbol{\Sigma}_{\mathcal{B}_{01}})^2 + f_{\mathcal{B}_{12}} (\hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{12}} - \boldsymbol{\Sigma}_{\mathcal{B}_{01}})^2 - f_{\mathcal{B}_{01}} (\hat{\boldsymbol{\Sigma}}_{\mathcal{B}_{01}} - \boldsymbol{\Sigma}_{\mathcal{B}_{01}})^2\} \tilde{\boldsymbol{\alpha}}] \\
= \; & T_{21} + T_{22}.
\end{aligned}
$$

By (S2-6),

$$
T_{21} = -2 \frac{f_{\mathcal{B}_{11}} f_{\mathcal{B}_{12}}}{f_{\mathcal{B}_{01}}} \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{11}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{12}})^\top \boldsymbol{\Sigma}_{\mathcal{B}_{01}} \tilde{\boldsymbol{\alpha}}\} = O_P\left(\frac{1}{n}\right).
$$

Note that $\hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}} = \mathbf{\Sigma}_{\mathcal{B}_{01}} + O_P(n^{-1/2})$ and $\hat{\mathbf{\Sigma}}_{\mathcal{B}_{1s}} = \mathbf{\Sigma}_{\mathcal{B}_{01}} + O_P(n^{-1/2})$, $s = 1, 2$. It follows that

$$T_{22} = O_P\left(\frac{1}{n}\right).$$

Consequently,

$$\mathrm{BIC}_2(\mathcal{S}; \tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0; \tilde{\boldsymbol{\alpha}}) = O_P\left(\frac{1}{n}\right) - \frac{\log(n)}{n}\mathrm{df}_0.$$

Similarly, we can show that this result holds for any $\mathcal{S} \in \mathfrak{S}_+$. Therefore, as $n \to \infty$,

$$P\left\{\max_{\mathcal{S}\in\mathfrak{S}_+} \mathrm{BIC}_2(\mathcal{S}; \tilde{\boldsymbol{\alpha}}) < \mathrm{BIC}_2(\mathcal{S}_0; \tilde{\boldsymbol{\alpha}})\right\} \to 1. \tag{S2-7}$$

Consider now the case where $\mathcal{S}$ is under-slicing, that is, $\mathcal{S} \in \mathfrak{S}_-$. For simplicity, assume that $\mathcal{S} = \{\mathcal{B}_{0*}, \mathcal{B}_{03}, \ldots, \mathcal{B}_{0G_0}\}$. Here $\mathcal{B}_{0*}$ is a new slice constructed by merging $\mathcal{B}_{01}$ and $\mathcal{B}_{02}$. We have

$$\begin{aligned}
& \mathrm{BIC}_2(\mathcal{S}; \tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0; \tilde{\boldsymbol{\alpha}}) \\
= \; & f_{\mathcal{B}_{0*}}\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top(\mathbf{I}_p - \hat{\mathbf{\Sigma}}_{\mathcal{B}_{0*}})^2\tilde{\boldsymbol{\alpha}}\} - f_{\mathcal{B}_{01}}\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top(\mathbf{I}_p - \hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}})^2\tilde{\boldsymbol{\alpha}}\} \\
& - f_{\mathcal{B}_{02}}\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top(\mathbf{I}_p - \hat{\mathbf{\Sigma}}_{\mathcal{B}_{02}})^2\tilde{\boldsymbol{\alpha}}\} + \frac{\log(n)}{n}\mathrm{df}_0 \\
= \; & 2\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top(f_{\mathcal{B}_{01}}\hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}} + f_{\mathcal{B}_{02}}\hat{\mathbf{\Sigma}}_{\mathcal{B}_{02}} - f_{\mathcal{B}_{0*}}\hat{\mathbf{\Sigma}}_{\mathcal{B}_{0*}})\tilde{\boldsymbol{\alpha}}\} \\
& - \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top(f_{\mathcal{B}_{01}}\hat{\mathbf{\Sigma}}^2_{\mathcal{B}_{01}} + f_{\mathcal{B}_{02}}\hat{\mathbf{\Sigma}}^2_{\mathcal{B}_{02}} - f_{\mathcal{B}_{0*}}\hat{\mathbf{\Sigma}}^2_{\mathcal{B}_{0*}})\tilde{\boldsymbol{\alpha}}\} + \frac{\log(n)}{n}\mathrm{df}_0.
\end{aligned}$$

If the optimal slicing scheme $\mathcal{S}_0$ is in location, then $\hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}} = \mathbf{\Sigma}_{\mathcal{B}_{01}} + O_P(n^{-1/2})$

and $\hat{\mathbf{\Sigma}}_{\mathcal{B}_{1s}} = \mathbf{\Sigma}_{\mathcal{B}_{01}} + O_P(n^{-1/2})$, $s = 1, 2$. Hence

$$
\begin{aligned}
& \mathrm{BIC}_2(\mathcal{S}; \tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0; \tilde{\boldsymbol{\alpha}}) \\
= \; & 2\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (f_{\mathcal{B}_{01}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}} + f_{\mathcal{B}_{02}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{02}} - f_{\mathcal{B}_{0*}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{0*}}) \tilde{\boldsymbol{\alpha}}\} + O_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

Again, it is easy to see that

$$
f_{\mathcal{B}_{0*}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{0*}} = f_{\mathcal{B}_{01}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}} + f_{\mathcal{B}_{02}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{02}} + \frac{f_{\mathcal{B}_{01}} f_{\mathcal{B}_{02}}}{f_{\mathcal{B}_{0*}}} (\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})^\top \tag{S2-8}
$$

It follows that

$$
\begin{aligned}
& \mathrm{BIC}_2(\mathcal{S}; \tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0; \tilde{\boldsymbol{\alpha}}) \\
= \; & -2\frac{f_{\mathcal{B}_{01}} f_{\mathcal{B}_{02}}}{f_{\mathcal{B}_{0*}}} \mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})(\hat{\boldsymbol{\mu}}_{\mathcal{B}_{01}} - \hat{\boldsymbol{\mu}}_{\mathcal{B}_{02}})^\top \tilde{\boldsymbol{\alpha}}\} + O_P\left(\frac{1}{\sqrt{n}}\right) \\
= \; & -2\frac{\pi_{\mathcal{B}_{01}} \pi_{\mathcal{B}_{02}}}{\pi_{\mathcal{B}_{0*}}} \mathrm{trace}\{\boldsymbol{\alpha}^\top (\boldsymbol{\mu}_{\mathcal{B}_{01}} - \boldsymbol{\mu}_{\mathcal{B}_{02}})(\boldsymbol{\mu}_{\mathcal{B}_{01}} - \boldsymbol{\mu}_{\mathcal{B}_{02}})^\top \boldsymbol{\alpha}\} + O_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}
$$

Since $\boldsymbol{\mu}_{\mathcal{B}_{01}} \neq \boldsymbol{\mu}_{\mathcal{B}_{02}}$, there exists a constant $c_1 < 0$ such that $\mathrm{BIC}_2(\mathcal{S}; \tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0; \tilde{\boldsymbol{\alpha}}) < c_1$, with probability tending to 1 as $n \to \infty$.

If the optimal slicing scheme is in scale, then $\hat{\boldsymbol{\mu}}_{\mathcal{B}_{0s}} = \boldsymbol{\mu}_{\mathcal{B}_{0*}} + O_P(n^{-1/2})$, $s = 1, 2$. By (S2-8),

$$
f_{\mathcal{B}_{0*}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{0*}} = f_{\mathcal{B}_{01}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}} + f_{\mathcal{B}_{02}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{02}} + O_P\left(\frac{1}{n}\right)
$$

and

$$
\hat{\mathbf{\Sigma}}_{\mathcal{B}_{0*}} = \frac{f_{\mathcal{B}_{01}}}{f_{\mathcal{B}_{0*}}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{01}} + \frac{f_{\mathcal{B}_{02}}}{f_{\mathcal{B}_{0*}}} \hat{\mathbf{\Sigma}}_{\mathcal{B}_{02}} + O_P\left(\frac{1}{n}\right).
$$

Consequently,

$$\begin{aligned}
&\mathrm{BIC}_2(\mathcal{S};\tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}}) \\
= \;& -\mathrm{trace}\{\tilde{\boldsymbol{\alpha}}^\top (f_{\mathcal{B}_{01}}\hat{\boldsymbol{\Sigma}}^2_{\mathcal{B}_{01}} + f_{\mathcal{B}_{02}}\hat{\boldsymbol{\Sigma}}^2_{\mathcal{B}_{02}} - f_{\mathcal{B}_{0*}}\hat{\boldsymbol{\Sigma}}^2_{\mathcal{B}_{0*}})\tilde{\boldsymbol{\alpha}}\} + O_P\left(\frac{1}{\sqrt{n}}\right) \\
= \;& -\pi_{\mathcal{B}_{0*}}\mathrm{trace}\left[\boldsymbol{\alpha}^\top \left\{\frac{\pi_{\mathcal{B}_{01}}}{\pi_{\mathcal{B}_{0*}}}\boldsymbol{\Sigma}^2_{\mathcal{B}_{01}} + \frac{\pi_{\mathcal{B}_{02}}}{\pi_{\mathcal{B}_{0*}}}\boldsymbol{\Sigma}^2_{\mathcal{B}_{02}} - \left(\frac{\pi_{\mathcal{B}_{01}}}{\pi_{\mathcal{B}_{0*}}}\boldsymbol{\Sigma}_{\mathcal{B}_{01}} + \frac{\pi_{\mathcal{B}_{02}}}{\pi_{\mathcal{B}_{0*}}}\boldsymbol{\Sigma}_{\mathcal{B}_{02}}\right)^2\right\}\boldsymbol{\alpha}\right] + O_P\left(\frac{1}{\sqrt{n}}\right).
\end{aligned}$$

Since $\boldsymbol{\Sigma}_{\mathcal{B}_{01}} \neq \boldsymbol{\Sigma}_{\mathcal{B}_{02}}$, by Jensen's inequality, there exists a constant $c_2 < 0$ such that $\mathrm{BIC}_2(\mathcal{S};\tilde{\boldsymbol{\alpha}}) - \mathrm{BIC}_2(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}}) < c_2$, with probability tending to 1 as $n \to \infty$.

Together with the strategy from the first part, we can show that the above results holds for any $\mathcal{S} \in \mathfrak{S}_-$. Thus, as $n \to \infty$,

$$P\left\{\max_{\mathcal{S}\in\mathfrak{S}_-} \mathrm{BIC}_2(\mathcal{S};\tilde{\boldsymbol{\alpha}}) < \mathrm{BIC}_2(\mathcal{S}_0;\tilde{\boldsymbol{\alpha}})\right\} \to 1. \tag{S2-9}$$

Combining (S2-7) and (S2-9), the proof is complete.

Bibliography

Cook, R. D. and L. Forzani (2009). Likelihood-based sufficient dimension reduction. *Journal of the American Statistical Association 104*(485), 197–208.