DIMENSION FOLDING PCA AND PFC FOR MATRIX-VALUED PREDICTORS

Shanshan Ding and R. Dennis Cook

University of Minnesota and University of Minnesota

Supplementary Material

This note contains background and proofs of all propositions and corollaries proposed in the paper.

S1 Matrix normal distribution

A matrix-valued distribution (De Waal, 1985) is a probability distribution of a random matrix. The matrix normal distribution is a generalization of the multivariate normal distribution to matrix-valued random variables. Let $X = (X_{ij}), i = 1, ..., p_L, j = 1, ..., p_R$, be a matrix-valued variable. Its expected value and covariance matrix are defined as $\mathrm{E}[X] = (\mathrm{E}[X_{ij}]) = \mu$ and $\mathrm{var}(X) = \mathrm{E}[\mathrm{vec}(X - \mathrm{E}[X])\mathrm{vec}^T(X - \mathrm{E}[X])] = \Sigma$. Then X has a matrix normal distribution if its covariance can be decomposed as the Kronecker product of two positive definite matrices Ω and M, and $\mathrm{vec}(X)$ follows a multivariate normal distribution with mean $\mathrm{vec}(\mu)$ and covariance matrix $\Sigma = \Omega \otimes M$. The matrix normal distribution is denoted as $\mathbf{N}_{p_L \times p_R}(\mu, \Omega, M)$. Its density function is defined through the distribution of $\mathrm{vec}(X)$ and is given by

$$\begin{aligned} f_X(x) &= f_{\mathrm{vec}(X)}(\mathrm{vec}(x)) \\ &= (2\pi)^{-\frac{p_L p_R}{2}} |\Omega|^{-\frac{p_L}{2}} |M|^{-\frac{p_R}{2}} \exp\{-\frac{1}{2}\mathrm{tr}(\Omega^{-1}(x-\mu)^T M^{-1}(x-\mu))\}. \end{aligned} \tag{S1.1}$$

The second moments of X are $\mathrm{E}[(X-\mu)(X-\mu)^T] = M\mathrm{tr}(\Omega)$ and $\mathrm{E}[(X-\mu)^T(X-\mu)] = \Omega\mathrm{tr}(M)$. Thus, $\Omega = \mathrm{E}[(X-\mu)^T(X-\mu)]/\mathrm{tr}(M)$ is called the row covariance matrix and $M = \mathrm{E}[(X-\mu)(X-\mu)^T]/\mathrm{tr}(\Omega)$ is called the column covariance matrix. The rows or columns of X are independent if and only if Ω or M is diagonal. In addition, if both Ω and M are scalar matrices, X is called isotropic, which means that X has an isotropic variance.

The MLE algorithm for the matrix normal distribution was proposed by Dutilleul (1999). The MLE of μ is $\bar{x}$. For fixed M, the MLE $\hat{\Omega}$ is given by

$$\hat{\Omega} = \frac{1}{np_L}\sum_{i=1}^{n}(X_i - \bar{X})^T M^{-1}(X_i - \bar{X}); \tag{S1.2}$$

and for fixed Ω, the MLE $\hat{M}$ is

$$\hat{M} = \frac{1}{np_R}\sum_{i=1}^{n}(X_i - \bar{X})\Omega^{-1}(X_i - \bar{X})^T. \tag{S1.3}$$

Dutilleul (1999) showed that the MLEs of Ω and M estimated from (S1.2) and (S1.3) are positive definite if and only if $n \geq \max(p_L/p_R, p_R/p_L) + 1$, so a large sample size is not required in order to invert the estimated covariance matrices, as long as the relative ratios of the two dimensions are not too large.

Based on this result, for the general dimension folding PFC model with a log likelihood function (3.6), the MLE $\hat{\Omega}$ is given by

$$\hat{\Omega} = \frac{1}{np_L}\sum_{i=1}^{n}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T)^T M^{-1}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T), \tag{S1.4}$$

for fixed Γ_1, Γ_2, β_1, β_2 and M; and the MLE $\hat{M}$ is

$$\hat{M} = \frac{1}{np_R}\sum_{i=1}^{n}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T)\Omega^{-1}(X_i - \bar{X} - \Gamma_2\beta_2 f_i\beta_1^T\Gamma_1^T)^T, \tag{S1.5}$$

for fixed Γ_1, Γ_2, β_1, β_2 and Ω.

S2 Proofs

Proof of Propositions 1 and 3. We demonstrate the proof of Proposition 1 first. The condition $\nu|X \sim \nu \mid \Gamma_2^T X\Gamma_1$ is equivalent to $(X|\Gamma_2^T X\Gamma_1, \nu) \sim X|\Gamma_2^T X\Gamma_1$, where '$\sim$' stands for equivalence in distribution. Treating ν as a parameter matrix and X as data, we can show that $\Gamma_2^T X\Gamma_1$ is a sufficient statistic for $X|\nu$. Since Γ_1 and Γ_2 have the smallest column dimensions, it is equivalent to prove that $\Gamma_2^T X\Gamma_1$ is a minimum sufficient statistic for $X|\nu$. To show this, let $f(X|\nu)$ be the conditional density function of $X|\nu$, we consider the the log likelihood ratio based on model (2.2):

$$\log\frac{f(X|\nu)}{f(Z|\nu)} = -\frac{1}{2}\mathrm{tr}[(X-\mu)^T(X-\mu) - (Z-\mu)^T(Z-\mu)] + \mathrm{tr}[(\nu(\Gamma_1^T(X-Z)^T\Gamma_2)].$$

It can be seen that $\log f(X|\nu)/f(Z|\nu)$ is a constant in ν if and only if $(\Gamma_1^T(X-Z)^T\Gamma_2) = 0$. Thus, $\Gamma_2^T X\Gamma_1$ is a minimum sufficient statistic and the condition $\nu|X \sim \nu \mid \Gamma_2^T X\Gamma_1$ holds. Similarly, it can be shown that $(\Gamma_1 \otimes \Gamma_2)^T\mathrm{vec}(X)$ is a minimum sufficient statistic for $\mathrm{vec}(X)|\nu$ based the log likelihood ratio in (2.3).

For proposition 3, when the random error is isotropic, the result can be directly obtained from proposition 1. When the random error has a general matrix normal distribution, let $Z = M^{-\frac{1}{2}}X\Omega^{-\frac{1}{2}}$, then $\mathrm{vec}(Z) = (\Omega \otimes M)^{-\frac{1}{2}}\mathrm{vec}(X)$ has covariance $I_{p_Lp_R}$. Transforming model (3.2) into Z scale, we have $\mathcal{S}_{Y|\circ Z\circ} = (\Omega \otimes M)^{-\frac{1}{2}}\mathrm{Span}(\Gamma_1 \otimes$

Γ_2). Based on Proposition 1 in Li, et al. (2010), $\mathcal{S}_{Y|\circ X^\circ} = (\Omega^{-\frac{1}{2}} \otimes M^{-\frac{1}{2}})\mathcal{S}_{Y|\circ Z^\circ} = \mathrm{Span}(\Omega^{-1}\Gamma_1) \otimes \mathrm{Span}(M^{-1}\Gamma_2)$.

Proof of Propositions 2 and 4. We first prove Proposition 2. It is easy to see that

$$\begin{aligned}\sum_{i=1}^{n} \mathrm{tr}[(X_i - G_2\omega_i G_1^T)^T(X_i - G_2\omega_i G_1^T)] = {} & \mathrm{tr}(\sum_{i=1}^{n} X_i^T X_i) - 2\mathrm{tr}(\sum_{i=1}^{n} X_i^T G_2\omega_i G_1^T) \\ & + \mathrm{tr}(\sum_{i=1}^{n} \omega_i^T \omega_i)\end{aligned} \tag{S2.1}$$

Minimizing (S2.1) over G_1, G_2 and ω_i is the same as minimizing $L = \mathrm{tr}(\sum\limits_{i=1}^{n} \omega_i^T\omega_i) - 2\mathrm{tr}(\sum\limits_{i=1}^{n} X_i^T G_2\omega_i G_1^T)$. For fixed G_1 and G_2, to obtain the minimizer ν_i over ω_i, we take the first derivative of L corresponding to ω_i and have $\partial L/\partial\omega_i = 2\omega_i - 2(G_2^T X_i G_1)$. Since the second derivate of L on ω_i is positive, the minimum L is obtained when $\hat{\nu}_i = G_2^T X_i G_1$, $i = 1, ..., n$. Thus, the objective function L becomes

$$L = -\mathrm{tr}[G_1^T(\sum_{i=1}^{n} X_i^T \mathrm{P}_2 X_i)G_1], \tag{S2.2}$$

where $\mathrm{P}_2 = G_2G_2^T$. For fixed G_2, L is minimized by choosing the columns of the minimizer $\hat{\Gamma}_1$ over G_1 to be the d_R eigenvectors of $\sum\limits_{i=1}^{n} X_i^T\mathrm{P}_2X_i$ (or $\sum\limits_{i=1}^{n} X_i^T\mathrm{P}_2X_i/n$) corresponding to its d_R largest nonzero eigenvalues. Similarly, (S2.2) can be written as $L = -\mathrm{tr}[G_2^T(\sum\limits_{i=1}^{n} X_i\mathrm{P}_1X_i^T)G_2]$, where $\mathrm{P}_1 = G_1G_1^T$. Then for fixed G_1, the minimizer $\hat{\Gamma}_2$ over G_2 is obtained when its columns are composed by the d_L eigenvectors of $\sum\limits_{i=1}^{n} X_i\mathrm{P}_1X_i^T$ (or $\sum\limits_{i=1}^{n} X_i\mathrm{P}_1X_i^T/n$) corresponding to its d_L largest nonzero eigenvalues.

To prove Proposition 4, for fixed G_1 and b_1, let $f^* = f(Y)b_1^T$ and $G_{20} \in \mathbb{R}^{p_L\times(p_L-d_L)}$ be the orthogonal compliment of G_2, then we have

$$\begin{aligned}&\mathrm{E}_n\{\mathrm{tr}[(X - G_2b_2f(Y)b_1^TG_1^T)^T(X - G_2b_2f(Y)b_1^TG_1^T)]\} \\ =&\mathrm{E}_n\{\mathrm{tr}[(X - G_2b_2f^*G_1^T)^T(G_2G_2^T + G_{20}G_{20}^T)(X - G_2b_2f^*G_1^T)]\} \\ =&\mathrm{E}_n\{\mathrm{tr}[(G_2^TX - b_2f^*G_1^T)(G_2^TX - b_2f^*G_1^T)^T]\} + \mathrm{E}_n\{\mathrm{tr}[(G_{20}^TX)^T(G_{20}^TX)]\}\end{aligned} \tag{S2.3}$$

We first find the minimizer $\hat{\beta}_2$ over b_2 assuming other terms are fixed. By taking the first derivative of the last equation in (S2.3) corresponding to b_2, we have $\partial L_1/\partial b_2 = -2G_2^T\mathrm{E}_n(XG_1f^{*^T}) + 2b_2\mathrm{E}_n(f^*f^{*^T})$, then $\hat{\beta}_2 = G_2^T\mathrm{E}_n(XG_1f^{*^T})[\mathrm{E}_n(f^*f^{*^T})]^{-1}$. Replacing b_2 with $\hat{\beta}_2$, the objective function (3.5) becomes

$$\begin{aligned}&\mathrm{E}_n\{\mathrm{tr}[(X - G_2\hat{\beta}_2f^*G_1^T)^T(X - G_2\hat{\beta}_2f^*G_1^T)]\} \\ =&\mathrm{E}_n[\mathrm{tr}(XX^T)] - \mathrm{tr}\{P_{G_2}\mathrm{E}_n(XG_1f^{*^T})[\mathrm{E}_n(f^*f^{*^T})]^{-1}\mathrm{E}_n(XG_1f^{*^T})^T\}.\end{aligned}$$

Therefore, the minimizer $\hat{\Gamma}_2$ over G_2 has its columns formed by the first d_L eigenvectors of

$$\mathrm{E}_n(XG_1{f^*}^T)[\mathrm{E}_n(f^*{f^*}^T)]^{-1}\mathrm{E}_n(f^*G_1^TX^T),$$

and correspondingly $\hat{\beta}_2 = \hat{\Gamma}_2^T\mathrm{E}_n(XG_1{f^*}^T)[\mathrm{E}_n(f^*{f^*}^T)]^{-1}$.

Similarly, given G_2 and b_2, let $f^* = b_2f(Y)$ and we have

$$\begin{aligned}&\mathrm{E}_n\{\mathrm{tr}[(X - G_2b_2f(Y)b_1^TG_1^T)(X - G_2b_2f(Y)b_1^TG_1^T)^T]\}\\ =&\mathrm{E}_n\{\mathrm{tr}[(X^T - G_1b_1{f^*}^TG_2^T)^T(X^T - G_1b_1{f^*}^TG_2^T)]\}.\end{aligned}$$

The same procedure for estimating $\hat{\Gamma}_2$ and $\hat{\beta}_2$ can be applied to obtain $\hat{\Gamma}_1$ and $\hat{\beta}_1$. Hence the columns of $\hat{\Gamma}_1$ consist of the first d_R eigenvectors of the matrix

$$\mathrm{E}_n(X^TG_2f^*)[\mathrm{E}_n({f^*}^Tf^*)]^{-1}\mathrm{E}_n({f^*}^TG_2^TX),$$

and $\hat{\beta}_1 = \hat{\Gamma}_1^T\mathrm{E}_n(X^TG_2f^*)[\mathrm{E}_n({f^*}^Tf^*)]^{-1}$.

Proof of Proposition 5 and Corollary 1. To prove Proposition 5 (i), for fixed Ω, Γ_1 and β_1, let $X^* = X\Omega^{-\frac{1}{2}}$, and $f^* = f(Y)\beta_1^T\Gamma_1^T\Omega^{-\frac{1}{2}}$. The log likelihood function (3.6) under centered predictors becomes

$$l(\mathcal{S}_{\Gamma_2}, \beta_2, M) = C - \frac{np_R}{2}\mathrm{log}|M| - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\{(X_i^* - \Gamma_2\beta_2f_i^*)^TM^{-1}(X_i^* - \Gamma_2\beta_2f_i^*)\},$$

where $C = -\frac{np_Lp_R}{2}\mathrm{log}(2\pi) - \frac{np_L}{2}\mathrm{log}|\Omega|$. Treating Γ_2 and M fixed, by taking derivatives of the log likelihood corresponding to β_2, it is easy to obtain that

$$\hat{\beta}_2 = (\Gamma_2^TM^{-1}\Gamma_2)^{-1}\Gamma_2^TM^{-1}\mathbb{X}_L^T\mathbb{F}_L(\mathbb{F}_L^T\mathbb{F}_L)^{-1}.$$

Substituting $\hat{\beta}_2$ back, after some algebra we have

$$l(\mathcal{S}_{\Gamma_2}, M) = C - \frac{np_R}{2}\mathrm{log}|M| - \frac{np_R}{2}\{\mathrm{tr}(M^{-\frac{1}{2}}\tilde{M}M^{-\frac{1}{2}}) - \mathrm{tr}(P_{M^{-\frac{1}{2}}\Gamma_2}M^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}M^{-\frac{1}{2}}\}.$$

Now treating M fixed, the log likelihood is maximized when the columns of $M^{-\frac{1}{2}}\Gamma_2$ contain the first d_L eigenvectors of $M^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}M^{-\frac{1}{2}}$. Since $\hat{M}_{\mathrm{res}} = \tilde{M} - \hat{\Sigma}_{\mathrm{fit}_L}$, then the log likelihood reduces to

$$\begin{aligned}&l(M)\\ &= C - \frac{np_R}{2}\mathrm{log}|M| - \frac{np_R}{2}\{\mathrm{tr}(M^{-\frac{1}{2}}\hat{M}_{\mathrm{res}}M^{-\frac{1}{2}}) - \mathrm{tr}((I - P_{M^{-\frac{1}{2}}\Gamma_2})M^{-\frac{1}{2}}\hat{\Sigma}_{\mathrm{fit}_L}M^{-\frac{1}{2}}\}\\ &= C - \frac{np_R}{2}\mathrm{log}|M| - \frac{np_R}{2}\mathrm{tr}(M^{-1}\hat{M}_{\mathrm{res}}) - \frac{np_R}{2}\sum_{i=d_L+1}^{p_L}\lambda_i(M^{-1}\hat{\Sigma}_{\mathrm{fit}_L}).\end{aligned}$$

The MLE of M is $\hat{M} = \hat{M}_{\mathrm{res}} + \hat{M}_{\mathrm{res}}^{\frac{1}{2}}\hat{U}_L\hat{D}_L\hat{U}_L^T\hat{M}_{\mathrm{res}}^{\frac{1}{2}}$. This proof can be done in the same way as for Theorem 3.1 in Cook and Forzani (2008). Thus it is omitted. Substitute $\hat{M}$

back to the estimate of $M^{-\frac{1}{2}}\Gamma_2$, we have $\hat{\Gamma}_2 = \hat{M}^{\frac{1}{2}}$ times the first d_L eigenvectors of $\hat{M}^{-\frac{1}{2}}\hat{\Sigma}_{\text{fit}_L}\hat{M}^{-\frac{1}{2}}$ and further $\hat{\beta}_2 = \hat{\Gamma}_2^T\hat{M}^{-1}\mathbb{X}_L^T\mathbb{F}_L(\mathbb{F}_L^T\mathbb{F}_L)^{-1}$.

The results in Proposition 5 (ii) can be simply obtained by taking transpose of (3.1) and then following the above procedure.

To prove Corollary 1, let $A = (I_{p_L} + \hat{D}_L)^{-1}$ and $\tilde{U}_L = \hat{M}_{\text{res}}^{-\frac{1}{2}}\hat{U}_L A^{\frac{1}{2}}$. Applying Lemma A.1 in Cook and Forzani (2008), we have $\text{Span}(\tilde{U}_L) = \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$. Since A is a full rank diagonal matrix, $\text{Span}(\tilde{U}_L)$ is equal to $\text{Span}(\hat{M}_{\text{res}}^{-\frac{1}{2}}\hat{U}_L)$, where $\hat{U}_L$ are the first d_L eigenvectors of $\hat{M}_{\text{res}}^{-\frac{1}{2}}\hat{\Sigma}_{\text{fit}_L}\hat{M}_{\text{res}}^{-\frac{1}{2}}$. This implies that $\mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_L}(\hat{M}_{\text{res}}, \hat{\Sigma}_{\text{fit}_L})$. Similarly, one can show that $\mathcal{S}_{d_R}(\hat{\Omega}, \hat{\Sigma}_{\text{fit}_R}) = \mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}}, \hat{\Sigma}_{\text{fit}_R})$. Thus the second form holds. Since $\hat{\Sigma}_{\text{fit}_L} = \tilde{M} - \hat{M}_{\text{res}}$, it is easy to see that $\tilde{M}^{-1}\hat{\Sigma}_{\text{fit}_L}$ and $\hat{M}_{\text{res}}^{-1}\hat{\Sigma}_{\text{fit}_L}$ have the same eigenvectors. This provides the result: $\mathcal{S}_{d_L}(\tilde{M}, \hat{\Sigma}_{\text{fit}_L}) = \tilde{M}^{-\frac{1}{2}}\mathcal{S}_{d_L}(\tilde{M}^{-\frac{1}{2}}\hat{\Sigma}_{\text{fit}_L}\tilde{M}^{-\frac{1}{2}}) = \mathcal{S}_{d_L}(\tilde{M}^{-1}\hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_L}(\hat{M}_{\text{res}}, \hat{\Sigma}_{\text{fit}_L})$. Similarly, $\mathcal{S}_{d_R}(\hat{\Omega}_{\text{res}}, \hat{\Sigma}_{\text{fit}_R}) = \mathcal{S}_{d_R}(\tilde{\Omega}, \hat{\Sigma}_{\text{fit}_R})$. The third form is proved. The last two forms hold since $\tilde{\Omega} = \hat{\Omega}_{\text{res}} + \hat{\Sigma}_{\text{fit}_R}$ and $\tilde{M} = \hat{M}_{\text{res}} + \hat{\Sigma}_{\text{fit}_L}$.

Proof of Proposition 6. Recall that $g = \beta_2 f(Y)\beta_1^T$ is the true fitting function and $l = \kappa_2 h(Y)\kappa_1^T$ is the selected fitting function. Let h_i denote $h(Y_i)$ and h_Y denote $h(Y)$. By applying conventional PFC model, we can obtain proper initial values for our algorithm to prove the consistency of our estimates. To do so, we choose a nonzero vector $v \in \mathbb{R}^{p_L}$. Recall that $f(Y)$ is a diagonal fitted matrix with dimensions $r \times r$. Based on model (3.1), we have $X^Tv = \Gamma_1\beta_1 f(Y)\beta_2^T\Gamma_2^Tv + \varepsilon^Tv = \Gamma_1\beta_1 f(Y)\omega + \varepsilon^Tv$, where $\omega = \beta_2^T\Gamma_2^Tv$ is a r dimensional vector and $\text{var}(\varepsilon^Tv) = a\Omega$ with a constant $a = v^TMv$. Let $\tilde{f} \in \mathbb{R}^r$ denote a vector containing the diagonal elements in $f(Y)$, then $f(Y)$ can be written as $\text{diag}(\tilde{f})$. Since $\text{diag}(\tilde{f})\omega = \text{diag}(\omega)\tilde{f}$, it follows that $X^Tv = \Gamma_1\beta_1\text{diag}(\omega)\tilde{f} + \varepsilon^Tv = \Gamma\beta\tilde{f} + \varepsilon^Tv$, where $\Gamma = \Gamma_1$ and $\beta = \beta_1\text{diag}(\omega)$. This forms a conventional PFC model and the unknown parameters Γ, β and Ω can be estimated based on it. Conventional PFC provides $\sqrt{n}$ consistent estimator for the true subspace $\text{Span}(\Omega^{-1}\Gamma)$, even when the function $\tilde{f}$ is misspecified by $\tilde{h}$ but they are sufficiently correlated (Cook and Forzani (2008)), where $\text{diag}(\tilde{h}) = h(Y)$. Thus, we can apply conventional PFC to get proper initial values of Γ_1, κ_1 and Ω as $\hat{\Gamma}$, $\hat{\kappa}$ and $\hat{\Omega}$. Let $X^* = X_i\hat{\Omega}^{-\frac{1}{2}}$, $h^* = h_Y\hat{\kappa}^T\hat{\Gamma}^T\hat{\Omega}^{-\frac{1}{2}}$. Then $\hat{\Sigma}_{\text{fit}_L} = (\sum_{i=1}^n X_i^*{h_i^*}^T/n)(\sum_{i=1}^n h_i^*{h_i^*}^T/n)^{-1}(\sum_{i=1}^n h_i^*{X_i^*}^T/n)/p_R$ converges to $\Sigma_{\text{fit}_L} = \text{E}(X\Omega^{-1}\Gamma_1\kappa h_Y^T)Q^{-1}\text{E}(X\Omega^{-1}\Gamma_1\kappa h_Y^T)^T/p_R$, where $Q = \text{var}_c(h_Y\kappa^T) = \text{E}(h_Y\kappa^T\kappa h_Y)$, $\kappa = \kappa_1\text{diag}(\omega)$ and $\omega = \kappa_2^T\Gamma_2^Tv$. Using (3.1), we have $\text{E}(X\Omega^{-1}\Gamma_1\kappa h_Y^T) = \text{E}(\Gamma_2 g\kappa h_Y^T) = \Gamma_2\text{cov}_c(g, h_Y\kappa^T) = \Gamma_2\text{cov}_c(g, h_Y\text{diag}(\omega)\kappa_1^T) = \Gamma_2 V\text{diag}(\omega)$, where $V = \text{cov}_c(g, h_Y\kappa_1^T)$. Thus, $\Sigma_{\text{fit}_L} = \Gamma_2 V\text{diag}(\omega)Q^{-1}\text{diag}(\omega)V^T\Gamma_2^T/p_R$. As early defined, $\tilde{M} = \sum_{i=1}^n X_i^*{X_i^*}^T/np_R = \sum_{i=1}^n X_i\Omega^{-1}X_i^T/np_R$. It follows that $\tilde{M}$ converges at $\sqrt{n}$ rate to $M^* = \text{E}[X\Omega_1^{-1}X^T]/p_R = (\Gamma_2\text{var}_c(g)\Gamma_2^T + M)/p_R$. The last equation is obtained based on (3.1).

From Corollary 1, we know $\mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L}) = \mathcal{S}_{d_L}(\tilde{M}, \hat{\Sigma}_{\text{fit}_L})$, that is equivalent to $\mathcal{S}_{d_L}(\tilde{M}^{-1}\hat{\Sigma}_{\text{fit}_L})$. Hence $\mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\text{fit}_L})$ converges to $\mathcal{S}_{d_L}({M^*}^{-1}\Sigma_{\text{fit}_L})$ at $\sqrt{n}$ rate. Using the fact that $(\Gamma_2 C\Gamma_2^T + M)^{-1} = M^{-1} - M^{-1}\Gamma_2(C^{-1} + \Gamma_2^TM^{-1}\Gamma_2)^{-1}\Gamma_2^TM^{-1}$, we have $\text{Span}({M^*}^{-1}\Sigma_{\text{fit}_L}) = \text{Span}\{(\Gamma_2\text{var}_c(g)\Gamma_2^T + M)^{-1}\Gamma_2 V\text{diag}(\omega)Q^{-1}\text{diag}(\omega)V^T\Gamma_2^T\} \subseteq$

$\mathrm{Span}\{(\Gamma_2 \mathrm{var}_c(g)\Gamma_2^T + M)^{-1}\Gamma_2\} = \mathrm{Span}(M^{-1}\Gamma_2)$. Since Γ_2 has full rank d_L and $\mathrm{diag}(\omega)$ has full rank r (Its diagonal elements are all nonzeros with probability one.), we have $\mathrm{Span}({M^*}^{-1}\Sigma_{\mathrm{fit}_L}) = \mathrm{Span}(M^{-1}\Gamma_2)$ if and only if the rank of $V = \mathrm{cov}_c(g, h_Y\kappa_1^T)$ is equal to d_L. Since $\rho_L = \mathrm{var}_c^{-\frac{1}{2}}(g)\mathrm{cov}_c(g, l)\mathrm{var}_c^{-\frac{1}{2}}(l) = \mathrm{var}_c^{-\frac{1}{2}}(g)\mathrm{cov}_c(g, h_Y\kappa_1^T)\kappa_2^T\mathrm{var}_c^{-\frac{1}{2}}(l)$ and κ_2 has rank d_L, the rank of ρ_L is equal to the rank of $\mathrm{cov}_c(g, h_Y\kappa_1^T)$.

Similarly, by following the above steps one can show that $\mathcal{S}_{d_R}(\hat{\Omega}^{-1/2}, \hat{\Sigma}_{\mathrm{fit}_R})$ converges to $\mathrm{Span}(\Omega^{-1}\Gamma_1)$ at $\sqrt{n}$ rate if and only if $\mathrm{cov}_r(g, h_Y^T\kappa_2^T)$ or, equivalently, ρ_R has rank d_R, based on the fact that $\mathrm{Span}(\hat{M}^{-1}\hat{\Gamma}_2) = \mathcal{S}_{d_L}(\hat{M}, \hat{\Sigma}_{\mathrm{fit}_L})$ is $\sqrt{n}$ consistent to $\mathrm{Span}(M^{-1}\Gamma_2)$.

Proof of Proposition 7. Assume that $E(X) = 0$. Let $Z = M^{-\frac{1}{2}}X\Omega^{-\frac{1}{2}}$ and let $\mathcal{S}_{Y|\circ Z\circ} = \mathrm{Span}(\alpha_1 \otimes \alpha_2)$. Under the elliptically symmetric condition, we have

$$\begin{aligned}(\Omega \otimes M)^{-\frac{1}{2}}\mathrm{E}[\mathrm{vec}(X)|Y] = \mathrm{E}[\mathrm{vec}(Z)|Y] &= \mathrm{E}\{\mathrm{E}[\mathrm{vec}(Z)|(\alpha_1 \otimes \alpha_2)^T\mathrm{vec}(Z), Y]|Y\} \\ &= \mathrm{E}\{\mathrm{E}[\mathrm{vec}(Z)|(\alpha_1 \otimes \alpha_2)^T\mathrm{vec}(Z)]|Y\} \qquad \text{(S2.4)} \\ &= P_{\alpha_1\otimes\alpha_2}\mathrm{E}[\mathrm{vec}(Z)|Y].\end{aligned}$$

Thus, $(\Omega \otimes M)^{-\frac{1}{2}}\mathrm{E}[\mathrm{vec}(X)|Y] \in \mathcal{S}_{Y|\circ Z\circ}$. From model (3.1), we can observe that $\mathrm{E}[\mathrm{vec}(Z)|Y] = (\Omega \otimes M)^{-\frac{1}{2}}(\Gamma_1 \otimes \Gamma_2)(\beta_1 \otimes \beta_2)\mathrm{vec}(f(Y))$. Hence $(\Omega \otimes M)^{-\frac{1}{2}}\mathrm{Span}(\Gamma_1 \otimes \Gamma_2) = \mathrm{Span}\{\mathrm{E}[\mathrm{vec}(Z)|Y] : \text{over all } Y\} \subseteq \mathcal{S}_{Y|\circ Z\circ}$. By the invariance property $\mathcal{S}_{Y|\circ Z\circ} = (\Omega \otimes M)^{\frac{1}{2}}\mathcal{S}_{Y|\circ X\circ}$, we have $\mathcal{S}_{fPFC} = (\Omega \otimes M)^{-1}\mathrm{Span}(\Gamma_1 \otimes \Gamma_2) = \mathrm{Span}\{\zeta = (\Omega \otimes M)^{-1}\mathrm{E}[\mathrm{vec}(X)|Y] : \text{over all } Y\} \subseteq \mathcal{S}_{Y|\circ X\circ}$.

Dimension folding SIR can be formulated with $f(Y)$ specified by $h(Y) = \mathrm{diag}\{I(Y \in J_1) - \frac{n_1}{n}, ..., I(Y \in J_{h-1}) - \frac{n_{h-1}}{n}\}^T = \mathrm{diag}(I(\tilde{Y} = 1) - \frac{n_1}{n}, ..., I(\tilde{Y} = h-1)/(h-1) - \frac{n_{h-1}}{n})^T$. Then $\tilde{\zeta} = (\Omega \otimes M)^{-1}\mathrm{E}[\mathrm{vec}(X)|\tilde{Y}] = (\Omega \otimes M)^{-1}(\Gamma_1 \otimes \Gamma_2)(\beta_1 \otimes \beta_2)\mathrm{vec}(h(Y))$. It follows that $\mathrm{Span}(\tilde{\zeta}) = \mathrm{Span}\{(\Omega \otimes M)^{-1}\mathrm{E}[\mathrm{vec}(X)|\tilde{Y}] : \text{over all } \tilde{Y}\} \subseteq \mathrm{Span}\{(\Omega \otimes M)^{-1}(\Gamma_1 \otimes \Gamma_2)\} = \mathcal{S}_{fPFC}$. According to Theorem 1 in Li et al. (2010), the dimension folding SIR subspace $\mathcal{S}_{fSIR}$ is equal to the Kronecker envelope $\mathcal{E}^{\otimes}(\zeta)$, which is the Kronecker product of the two smallest subspaces $\mathcal{S}_{\circ\zeta} \otimes \mathcal{S}_{\zeta\circ}$ such that $\mathrm{Span}(\zeta) \subseteq \mathcal{S}_{\circ\zeta} \otimes \mathcal{S}_{\zeta\circ}$. Therefore, $\mathcal{S}_{fSIR} \subseteq \mathcal{S}_{fPFC}$.