Supplementary Material: Model Averaging Estimation for Partially Linear Functional Score Models

Shishi Liu[1], Chunming Zhang[2], Hao Zhang[4], Rou Zhong[3,4] and Jingxiao Zhang[3,4,*]

[1]*School of Economics, Hangzhou Dianzi University*

[2]*Department of Statistics, University of Wisconsin*

[3]*Center for Applied Statistics, Renmin University of China*

[4]*School of Statistics, Renmin University of China*

Supplementary Material

This supplementary file contains additional simulations, additional details in real application, and all technical proofs. Here we omit the notation $a.s.$ (almost surely) for convenience.

S1 Additional Simulations

Example 3. $M_0 = 4$ and $\boldsymbol{\theta} = (1.5, 0.7, 0.2, -0.4)^T$. $\mathbf{Z}_i$ is a 4×1 vector that follows a multivariate normal distribution with zero means and a variance-covariance matrix $\Sigma = (0.5^{|a-b|})_{4\times 4}$. The functional predictor $X_i(t)$ is obtained by

$$X_i(t) = \sum_{l=1}^{4} \zeta_{il}\psi_l(t), \quad t \in [0, 1],$$

where $\psi_l(t) = \sqrt{2}\sin(\pi l t)$, $l = 1, \ldots, 4$, and ζ_{il} is i.i.d and simulated from $N(0, l^{-3/2})$, $i = 1, \ldots, n$. The random error term ε_i is i.i.d. and follows $N(0, \eta^2)$.

η controls the signal-to-noise ratio and we vary it such that $R^2 = var(\mu_i)/var(Y_i)$ ranges from 0.1 to 0.9, where $var(\mu_i)$ and $var(Y_i)$ denote the variances of μ_i and Y_i, respectively. And the non-linear effect of $X_i(t)$ is introduced by

$$\mathbf{f}(\boldsymbol{\xi}_i) = \exp\Big\{\sum_{l=1}^{4} \xi_{il}/l\Big\},$$

where $\xi_{il} = \Phi(\lambda_l^{-1/2}\zeta_{il})$. All settings are identical to that of Example 1 except for the non-linear effect of $X(t)$.

Example 4. $M_0 = 50$ and $\theta_j = j^{-1/2}$. The case where $\mathbf{z}$ and $X(t)$ being correlated is considered. Simulate $(\mathbf{Z}_i, \zeta_{i1}) \sim MN(0, \Sigma)$, where $\Sigma = (0.5^{|a-b|})_{51\times 51}$. The functional predictor $X_i(t)$ is simulated by

$$X_i(t) = \sum_{l=1}^{5} \zeta_{il}\psi_l(t), \quad t \in [0, 10],$$

where $\psi_l(t) = \cos(\pi lt/5)/\sqrt{5}$, $l = 1, \dots, 5$, and ζ_{il} is i.i.d. and follows $N(0, l^{-2})$, $i = 1, \dots, n$, $l = 2, \dots, 5$. The independent ε_i's are heteroscedastic as $\varepsilon_i \sim N\big(0, \eta^2(u_i^2 + 0.01)\big)$, where u_i follows $U[-1, 1]$. Still varying η such that R^2 varies between 0.1 and 0.9. And the non-linear effect of $X_i(t)$ is generated from

$$\mathbf{f}(\boldsymbol{\xi}_i) = 2\big(\xi_{i1}\xi_{i2} - \frac{1}{2}\big) + \big(\xi_{i3} - \frac{1}{2}\big)^2 - \frac{1}{12} + \frac{1}{4}\big(\xi_{i4} - \frac{1}{2}\big) + \frac{1}{5}\big(\xi_{i5} - \frac{1}{2}\big),$$

where $\xi_{il} = \Phi(\lambda_l^{-1/2}\zeta_{il})$.

Identical to Example 1, we still omit z_4 and ξ_4 in preparing candidate models in Example 3, so all candidate models are misspecified. With different specifications of which elements in $\{z_1, z_2, z_3\}$ and $\{\xi_1, \xi_2, \xi_3\}$ are included in the model, we have a total number of $M = (2^3 - 1)(2^3 - 1) = 49$ candidate models for Ex-

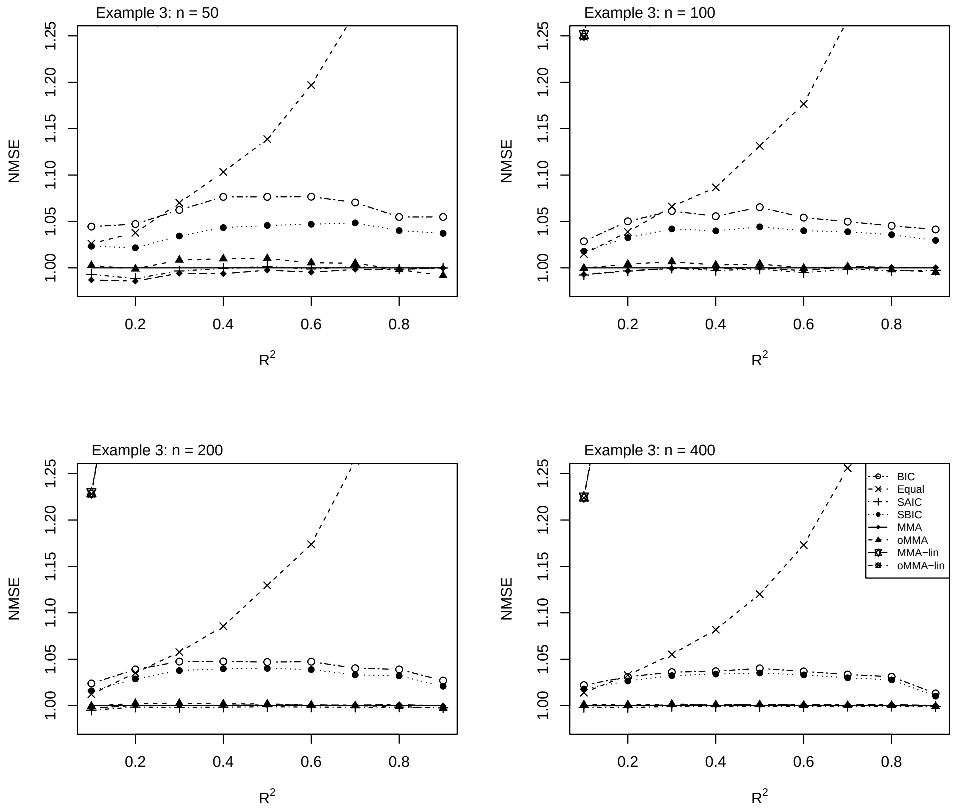

Figure S1: Normalized mean squared error (NMSE) comparisons for Example 3.

amples 1 and 2. As for Example 4, similar to that of Example 2, a pre-screening is conducted first, and candidate models are then constructed in a nested way.

In Example 3, both MMA-type and AIC-type estimators in Figure S1 demonstrate superiority over other competing estimators. Particularly, when sample size n is limited, MMA exhibits an advantage over AIC and SAIC methods. With increasing n, these estimators tend to behave similarly in most cases for R^2. This underscores the effectiveness of MMA approach. Additionally, it is observed that the oracle MMA (oMMA) is less effective for small values of R^2 and n, which

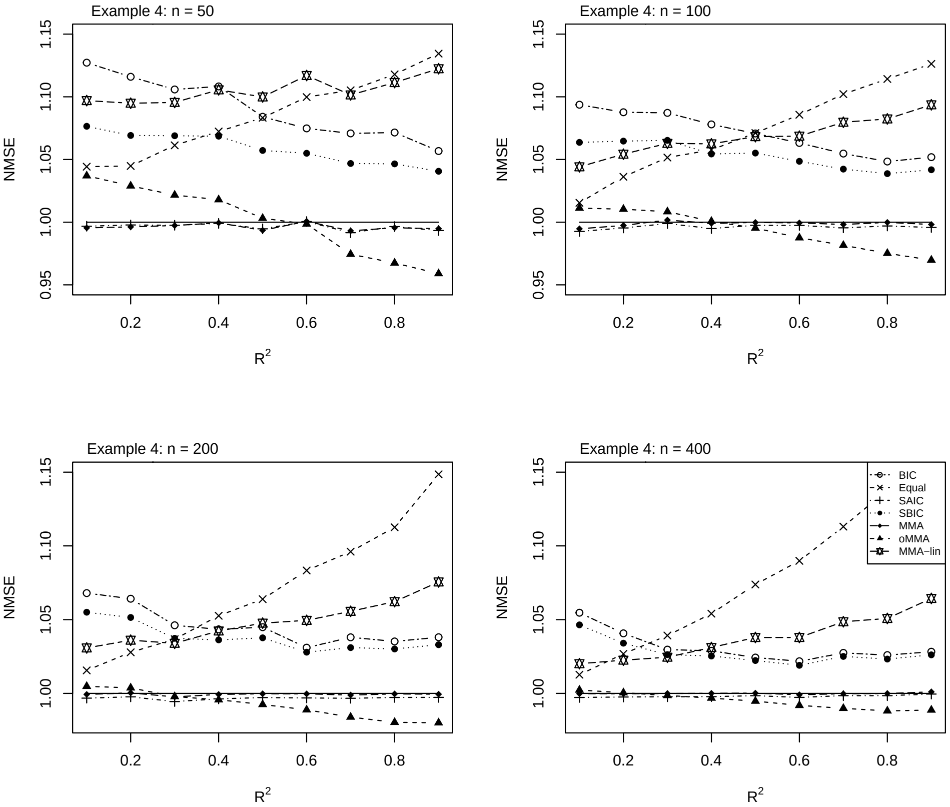

Figure S2: Normalized mean squared error (NMSE) comparisons for Example 4.

can be attributed to the signal-to-noise ratio in data , as discussed earlier in the manuscript. Also, Figure S1 illustrates that SAIC and SBIC outperform their model selection counterparts, AIC and BIC. The differences tend to decrease as R^2 or n grows. Finally, it is worth noting that the lines for MMA-lin and oracle MMA-lin (oMMA-lin) mostly fall outside the subfigures; only two points of each can be seen at the top left of the subfigures. This implies that partially linear functional linear models are inefficient in detecting non-linear effects, highlighting the necessity of exploring nonparametric modeling. Similar trends to those in Figure 1 can also be observed in Figure S1.

In Example 4, as depicted in Figure S2, the oracle MMA (o-MMA) demonstrates a clear edge over other competing methods for medium and large R^2 values. Additionally, MMA outperforms the others for small R^2 and n values. As n increases, the differences between MMA and AIC-type methods diminish. On the other hand, MMA-lin deteriorates as R^2 increases, coinciding with the strengthening signal of the non-linear component. Notably, equally weighting still performs the poorest for most large R^2 values. Similar trends to those in Figure 2 can also be observed in Figure S2.

S2 Details in Real Application

We excluded one covariate, phosphorous ($mg{\cdot}kg^{-1}$), from the original set of 19 covariates due to its incompleteness. The sample marginal correlations of remaining covariates with our response variable, the total carbon percentage, are presented in Table 1, ranging from -0.3508 to 0.6523. To prepare candidate model, we firstly screened out scalar variables whose absolute marginal correlations exceed 0.1, and then adopted the nested fashion described in the manuscript.

Through $D = 1000$ repetitions, we finally obtained D mean squared prediction error (MSPE) values. Figures S3 and S4 present boxplots and empirical cumulative distribution functions of MSPEs for each approach. It is evident from Figure S4 that a visual stochastic dominance relationship exists between

Table 1: Sample marginal Correlations to the total carbon percentage (Y).

	marginal correlation
Exchangeable Magnesium ($cmol_c \cdot kg^{-1}$)	0.6523
Exchangeable Bases ($cmol_c \cdot kg^{-1}$)	0.6124
Exchangeable Potassium ($cmol_c \cdot kg^{-1}$)	0.6093
Exchangeable Calcium ($cmol_c \cdot kg^{-1}$)	0.5698
Boron Concentration ($mg \cdot kg^{-1}$)	0.5022
Soil Electrical Conductivity ($dS \cdot m^{-1}$)	0.4172
Sulphur ($mg \cdot kg^{-1}$)	0.3890
Soil pH in Water	0.2202
Iron Concentration ($mg \cdot kg^{-1}$)	0.2143
Copper Concentration ($mg \cdot kg^{-1}$)	0.1963
Zinc ($mg \cdot kg^{-1}$)	0.1733
Exchangeable Calcium-to-Magnesium Ratio	0.0826
Exchangeable Manganese ($mg \cdot kg^{-1}$)	0.0749
Exchangeable Acidity ($cmol_c \cdot kg^{-1}$)	-0.0147
Exchangeable Aluminium ($mg \cdot kg^{-1}$)	-0.0533
Exchangeable Sodium ($cmol_c \cdot kg^{-1}$)	-0.0991
Exchangeable Sodium Ratio (%)	-0.3088
Exchangeable Sodium Percentage (%)	-0.3508

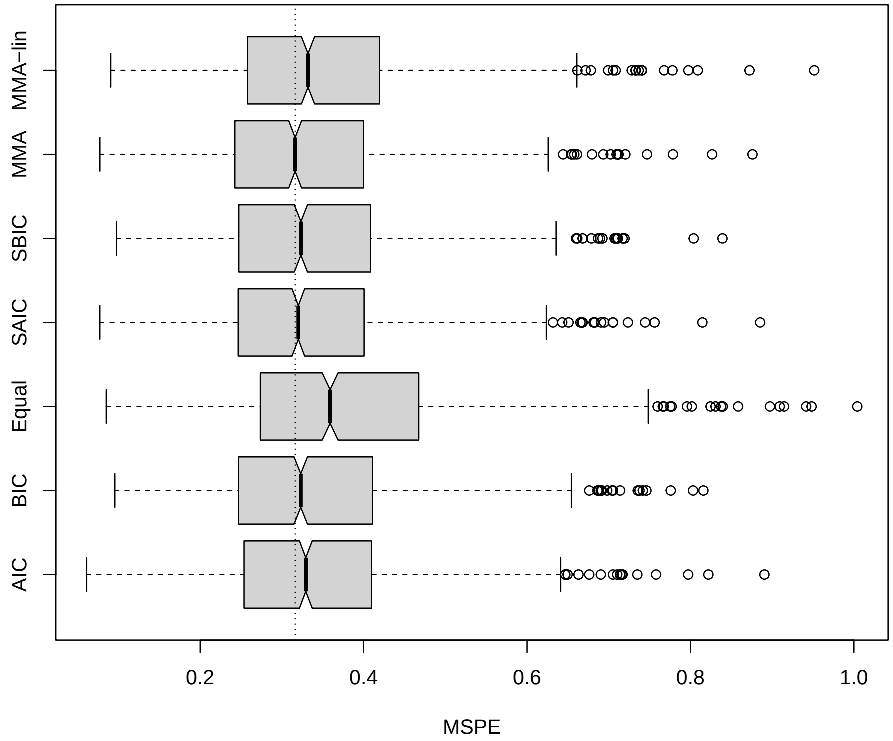

Figure S3: Boxplots of MSPE across $D = 1000$ repetitions for the soil dataset. The vertical dashed line represents the median MSPE value of MMA.

MMA and other competitors, indicating the preference for MMA on this dataset. Moreover, Figure S4 reveals that linear modeling may be underspecified, and the equally weighting scheme may be inefficient in practice.

Additionally, the remaining results for the data-driven method using paired Mann-Whitney-Wilcoxon (MWW) test are listed in Table 2. The alternative hypothesis indicates that the other method is less accurate than MMA. A small statistic value (relative to $D(D + 1)/2 \approx 250, 250$) of the MWW test suggests

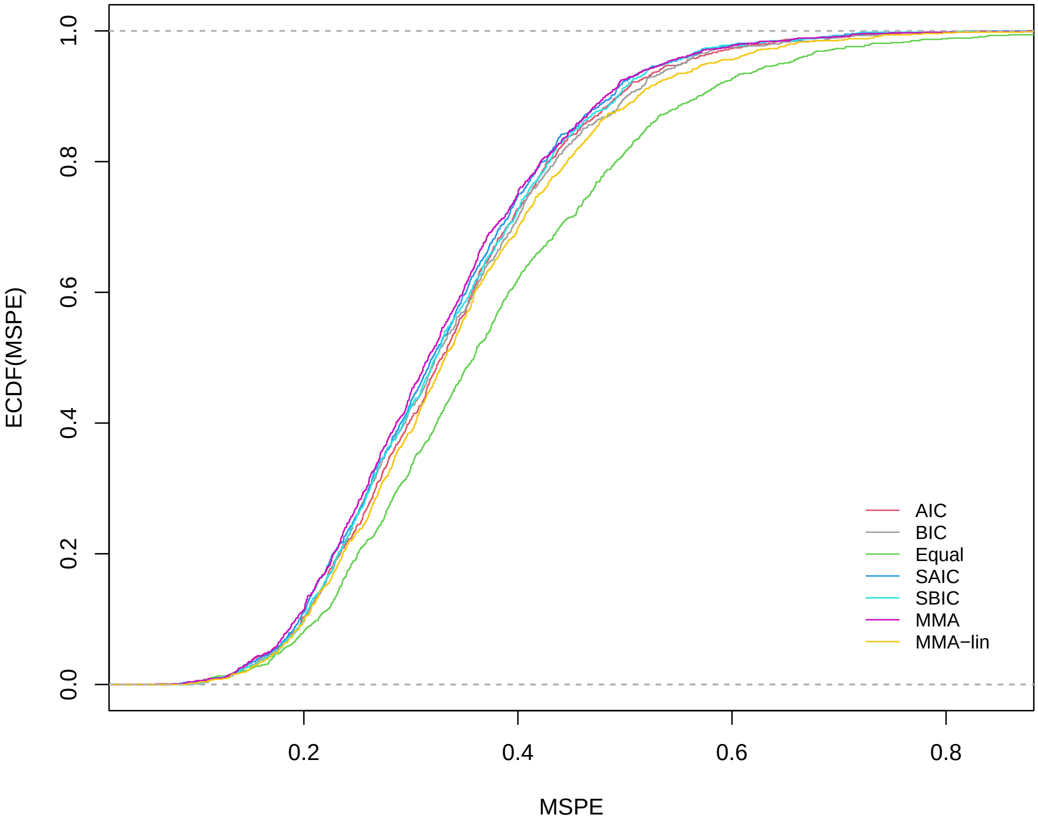

Figure S4: ECDFs of MSPE based on $D = 1000$ values for the soil dataset.

that the respective method is less accurate than MMA. The p-values have also been adjusted using the BH method, as employed in the manuscript. Again, both the MWW statistics and corresponding adjusted p-values in Table 2 demonstrate that MMA outperforms other competitors in terms of prediction accuracy for this dataset.

Table 2: Results of test statistics and adjusted p-values for the data-driven approach using paired MWW test.

	MMA/AIC	MMA/BIC	MMA/Equal	MMA/SAIC	MMA/SBIC	MMA/MMA-lin
mww stat.	1.31E+05	1.53E+05	1.16E+05	2.03E+05	1.79E+05	1.89E+05
mww p.val	4.63E-39	1.09E-26	1.12E-48	9.17E-08	4.54E-15	9.61E-12

S3 Some Lemmas

The estimation error of the transformed FPC score is of order $O_p(n^{-1/2}l)$ as shown in [4], and we list the result here while omitting the detailed proof.

Lemma S3.1. *Suppose the transformation function $\Phi(\cdot)$ has bounded derivative. Under Assumption 1, there exists a constant $C > 0$ such that $\mathbb{E}(\widehat{\xi}_{il} - \xi_{il})^2 \leq Cl^2/n$ uniformly for $l \leq J_n$, where $J_n = \lfloor (2C_\lambda O_p(1))^{-1/(1+\alpha)} n^{1/(2+2\alpha)} \rfloor$.*

Lemma S3.2. *Under Assumptions 1, 2 and 3(c), we have $\lambda_{\max}(\mathbf{K}_{(m)}) = O(1)$, $\lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) = O_p(1)$, and $\lambda_{\max}(\mathbf{P}_{(m)}) = O(1)$, $\lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) = O_p(1)$, for all $m = 1, \ldots, M$.*

Proof. For any square matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ (see [1]), we have

$$\begin{aligned} &\lambda_{\max}(\mathbf{M}_1\mathbf{M}_2) \leq \lambda_{\max}(\mathbf{M}_1)\lambda_{\max}(\mathbf{M}_2), \\ \text{and} \quad &\lambda_{\max}(\mathbf{M}_1 + \mathbf{M}_2) \leq \lambda_{\max}(\mathbf{M}_1) + \lambda_{\max}(\mathbf{M}_2). \end{aligned} \tag{S3.1}$$

These two inequalities will be frequently used in the following proofs.

By an inequality of Reisz (see [2]), we obtain that

$$\lambda^2_{\max}(\mathbf{K}_{(m)}) \leq \max_i \sum_{j=1}^n |\mathbf{K}_{(m),ij}| \cdot \max_j \sum_{i=1}^n |\mathbf{K}_{(m),ij}|,$$

which implies that $\lambda_{\max}(\mathbf{K}_{(m)}) = O(1)$. Hence,

$$\lambda_{\max}(\mathbf{P}_{(m)}) = \lambda_{\max}(\widetilde{\mathbf{P}}_{(m)})\big(1 + \lambda_{\max}(\mathbf{K}_{(m)})\big) + \lambda_{\max}(\mathbf{K}_{(m)})$$

$$= \big(1 + \lambda_{\max}(\mathbf{K}_{(m)})\big) + \lambda_{\max}(\mathbf{K}_{(m)}) = O(1).$$

From Lemma S3.1, we obtain that

$$\widehat{\xi}_{il} - \xi_{il} = O_p(n^{-1/2}l),$$

$$\widehat{\xi}_{il} - \widehat{\xi}_{jl} = \xi_{il} - \xi_{jl} + O_p(n^{-1/2}l), \quad \text{for} k \leq J_n.$$

Applying Taylor series expansion and Assumption 2,

$$\widehat{\mathbf{K}}_{(m),ij} = \frac{\mathcal{K}(\widehat{\boldsymbol{\xi}}_{(m),i} - \widehat{\boldsymbol{\xi}}_{(m),j})}{\sum_{j'=1}^{n} \mathcal{K}(\widehat{\boldsymbol{\xi}}_{(m),i} - \widehat{\boldsymbol{\xi}}_{(m),j'})}$$

$$= \Big\{\mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j}) + \sum_{l=1}^{q_m} k'(\xi_{il} - \xi_{jl}) \prod_{u \neq l} k(\xi_{iu} - \xi_{ju})(\widehat{\xi}_{il} - \xi_{il} + \xi_{jl} - \widehat{\xi}_{jl})$$

$$+ o_p(n^{-\frac{1}{2}}q_m)\Big\} / \Big\{ \sum_{j'=1}^{n} [\mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j'}) + O_p(n^{-\frac{3}{2}}q_m)] \Big\}$$

$$= \frac{\mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j})}{\sum_{j'=1}^{n} \mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j'})} + \Big\{ \sum_{l=1}^{q_m} k'(\xi_{il} - \xi_{jl}) \prod_{u \neq l} k(\xi_{iu} - \xi_{ju})(\widehat{\xi}_{il} - \xi_{il} + \xi_{jl} - \widehat{\xi}_{jl}) \Big\}$$

$$/ \Big\{ \sum_{j'=1}^{n} \mathcal{K}(\boldsymbol{\xi}_{(m),i} - \boldsymbol{\xi}_{(m),j'}) \Big\} + o_p(n^{-\frac{1}{2}}q_m)$$

$$= \mathbf{K}_{(m),ij} + O_p(n^{-\frac{3}{2}}q_m), \qquad q_m \leq J_n,$$

i.e., $\widehat{\mathbf{K}}_{(m),ij} = \mathbf{K}_{(m),ij} + O_p(n^{-\frac{3}{2}}q_m)$. Note that q_m is no larger than J_n and it is common for kernel smoothing to restrict the dimension of $\boldsymbol{\xi}$ to handle the curse of dimensionality. By Assumptions 2 and 3, we can show that

$$\max_i \sum_{j=1}^{n} |\widehat{\mathbf{K}}_{(m),ij}| = \max_i \sum_{j=1}^{n} |\mathbf{K}_{(m),ij}| + O_p(n^{-\frac{1}{2}}q_m) = O_p(1),$$

$$\max_j \sum_{i=1}^{n} |\widehat{\mathbf{K}}_{(m),ij}| = \max_j \sum_{i=1}^{n} |\mathbf{K}_{(m),ij}| + O_p(n^{-\frac{1}{2}}q_m) = O_p(1),$$

uniformly for $m = 1, \ldots, M$. Similarly,

$$\lambda^2_{\max}(\widehat{\mathbf{K}}_{(m)}) \le \max_i \sum_{j=1}^n |\widehat{\mathbf{K}}_{(m),ij}| \max_j \sum_{i=1}^n |\widehat{\mathbf{K}}_{(m),ij}| = O_p(1),$$

$$\begin{aligned}\lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) &= \lambda_{\max}(\overline{\mathbf{P}}_{(m)})\big(1 + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)})\big) + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) \\ &= \big(1 + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)})\big) + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) = O_p(1).\end{aligned}$$

In addition,

$$\max_i \sum_{j=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}| = O_p(n^{-\frac{1}{2}} q_m),$$

$$\max_j \sum_{i=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}| = O_p(n^{-\frac{1}{2}} q_m),$$

$$\lambda^2_{\max}(\widehat{\mathbf{K}}_{(m)} - \mathbf{K}_{(m)}) \le \max_i \sum_{j=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}| \max_j \sum_{i=1}^n |\widehat{\mathbf{K}}_{(m),ij} - \mathbf{K}_{(m),ij}|,$$

which leads to

$$\lambda_{\max}(\widehat{\mathbf{K}}_{(m)} - \mathbf{K}_{(m)}) = O_p(n^{-\frac{1}{2}} q_m). \tag{S3.2}$$

□

Lemma S3.3 corresponds to Lemma 1 in the paper.

Lemma S3.3. *Under Assumptions 1, 2 and 3(c)(e), we have*

$$\lambda_{\max}\big(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}\big) = O_p(n^{-\frac{1}{2}} q_m),$$

for all $m = 1, \ldots, M$.

Proof. Take a decomposition as

$$\begin{aligned}\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)} = (\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)}) + (\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)}) + \widetilde{\mathbf{P}}_{(m)}(\widehat{\mathbf{K}}_{(m)} - \mathbf{K}_{(m)}) \\ + (\overline{\mathbf{P}}_{(m)} - \widetilde{\mathbf{P}}_{(m)})\mathbf{K}_{(m)} + (\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)})(\mathbf{K}_{m} - \widehat{\mathbf{K}}_{(m)}).\end{aligned} \tag{S3.3}$$

Recalling Eq.(S3.1), it suffices to determine the order of $\lambda_{\max}(\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)})$ and $\lambda_{\max}(\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)})$. We have already quantified $\lambda_{\max}(\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)})$ in Eq.(S3.2). Remind that $\widetilde{\mathbf{P}}_{(m)}$ and $\overline{\mathbf{P}}_{(m)}$ are projection matrices related to $\widetilde{\mathbf{Z}}_{(m)} = (\mathbf{I} - \mathbf{K}_{(m)})\mathbf{Z}_{(m)}$ and $\widehat{\mathbf{Z}}_{(m)} = (\mathbf{I} - \widehat{\mathbf{K}}_{(m)})\mathbf{Z}_{(m)}$, i.e. $\widetilde{\mathbf{P}}_{(m)} = \widetilde{\mathbf{Z}}_{(m)}(\widetilde{\mathbf{Z}}_{(m)}^T\widetilde{\mathbf{Z}}_{(m)})^{-1}\widetilde{\mathbf{Z}}_{(m)}^T$ and $\overline{\mathbf{P}}_{(m)} = \widehat{\mathbf{Z}}_{(m)}(\widehat{\mathbf{Z}}_{(m)}^T\widehat{\mathbf{Z}}_{(m)})^{-1}\widehat{\mathbf{Z}}_{(m)}^T$. We simplify the notations as $\widetilde{\mathbf{P}}_{(m)} = \mathbf{H}_{(\mathbf{I}-\mathbf{K}_{(m)})\mathbf{Z}_{(m)}}$ and $\overline{\mathbf{P}}_{(m)} = \mathbf{H}_{(\mathbf{I}-\widehat{\mathbf{K}}_{(m)})\mathbf{Z}_{(m)}}$, where $\mathbf{H}_A$ represents the (normalised) projection operator generated from A. It is obviously observed that the deviation between two projection operators $\widetilde{\mathbf{P}}_{(m)}$ and $\overline{\mathbf{P}}_{(m)}$ derives from the difference between $\mathbf{K}_{(m)}$ and $\widehat{\mathbf{K}}_{(m)}$, so using Assumption 3(e) it holds that

$$\begin{aligned}&\|\widetilde{\mathbf{P}}_{(m)} - \overline{\mathbf{P}}_{(m)}\| \le \|\mathbf{H}_{\mathbf{I}-\mathbf{K}_{(m)}} - \mathbf{H}_{\mathbf{I}-\widehat{\mathbf{K}}_{(m)}}\| = \|\mathbf{H}_{\mathbf{K}_{(m)}} - \mathbf{H}_{\widehat{\mathbf{K}}_{(m)}}\| \\ &\le C\big(\|\mathbf{K}_{(m)} - \widehat{\mathbf{K}}_{(m)}\|\big)\big(\lambda_{\max}(\mathbf{K}_{(m)}) + \lambda_{\max}(\widehat{\mathbf{K}}_{(m)})\big) + 2\lambda_{\max}(\mathbf{K}_{(m)})\lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) = O_p(n^{-\frac{1}{2}}q_m),\end{aligned} \tag{S3.4}$$

where C is related to the smallest nonzero singular value of $\mathbf{K}_{(m)}$.

Finally according to Eq.(S3.3), combining Eqs.(S3.1), (S3.2) and (S3.4), we have

$$\lambda_{\max}(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}) = O_p(n^{-\frac{1}{2}}q_m),$$

which completes the proof. □

S4 Proof of Theorem 1

Proof. Firstly, it follows from Lemma S3.2 that

$$\begin{aligned}\sup_{\boldsymbol{\omega}} \lambda_{\max}\big(\widehat{\mathbf{P}}(\boldsymbol{\omega})\big) &= \sup_{\boldsymbol{\omega}} \lambda_{\max}\Big(\sum_{m=1}^{M} \omega_m \widehat{\mathbf{P}}_{(m)}\Big) \leq \sup_{\boldsymbol{\omega}} \sum_{m=1}^{M} \omega_m \lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) \\ &\leq \max_{1\leq m\leq M} \lambda_{\max}(\widehat{\mathbf{P}}_{(m)}) = O_p(1),\end{aligned} \tag{S4.5}$$

and similarly,

$$\sup_{\boldsymbol{\omega}} \lambda_{\max}\big(\mathbf{P}(\boldsymbol{\omega})\big) = O_p(1). \tag{S4.6}$$

Let $\widehat{\mathbf{A}}(\boldsymbol{\omega}) = \mathbf{I} - \widehat{\mathbf{P}}(\boldsymbol{\omega})$ and $\mathbf{A}(\boldsymbol{\omega}) = \mathbf{I} - \mathbf{P}(\boldsymbol{\omega})$. From the definition of $L_n(\boldsymbol{\omega})$, $\widehat{C}_n(\boldsymbol{\omega})$ and $R_n(\boldsymbol{\omega})$, we have

$$\begin{aligned}\widehat{C}_n(\boldsymbol{\omega}) = {}& L_n(\boldsymbol{\omega}) + \|\varepsilon\|^2 - 2\boldsymbol{\mu}^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} - 2\varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\varepsilon - 2\varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} \\ & - 2\varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T\boldsymbol{\mu} - 2\varepsilon^T(\widehat{\mathbf{P}}(\boldsymbol{\omega}) - \mathbf{P}(\boldsymbol{\omega}))^T\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\mu} + 2\varepsilon^T\mathbf{P}^T(\boldsymbol{\omega})(\widehat{\mathbf{P}}(\boldsymbol{\omega}) - \mathbf{P}(\boldsymbol{\omega}))\boldsymbol{\mu} \\ & + \boldsymbol{\mu}^T(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} + \varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\varepsilon \\ & + \varepsilon^T(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega}))^T(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\mu} + 2\varepsilon^T\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\mu} \\ & - 2[\varepsilon^T\mathbf{P}(\boldsymbol{\omega})\varepsilon - tr(\mathbf{P}(\boldsymbol{\omega})\mathbf{\Omega})] - 2[tr(\mathbf{P}(\boldsymbol{\omega})\mathbf{\Omega}) - tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\mathbf{\Omega})],\end{aligned}$$

and

$$L_n(\boldsymbol{\omega}) - R_n(\boldsymbol{\omega}) = \varepsilon^T\mathbf{P}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\varepsilon - tr(\mathbf{P}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\mathbf{\Omega}) - 2\varepsilon^T\mathbf{P}^T(\boldsymbol{\omega})\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\mu}.$$

So similar to the proof of Theorem 2.1 of [1], in order to prove Eq.(4) in Theorem 1, we need only to verify that

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T\mathbf{A}(\boldsymbol{\omega})\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.7}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T\mathbf{P}(\boldsymbol{\omega})\varepsilon - tr(\mathbf{P}(\boldsymbol{\omega})\mathbf{\Omega})|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.8}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \mathbf{P}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\varepsilon - tr(\mathbf{P}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.9}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\boldsymbol{\mu}^T \mathbf{A}^T(\boldsymbol{\omega})\mathbf{P}(\boldsymbol{\omega})\varepsilon|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.10}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.11}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) \boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.12}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \big(\widehat{\mathbf{P}}(\boldsymbol{\omega}) - \mathbf{P}(\boldsymbol{\omega})\big)^T \mathbf{P}(\boldsymbol{\omega})\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.13}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \mathbf{P}^T(\boldsymbol{\omega}) \big(\widehat{\mathbf{P}}(\boldsymbol{\omega}) - \mathbf{P}(\boldsymbol{\omega})\big)\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.14}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.15}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\boldsymbol{\mu}^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.16}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\boldsymbol{\mu}^T \big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.17}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)\varepsilon|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.18}$$

$$\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)\varepsilon|}{R_n(\boldsymbol{\omega})} = o_p(1), \tag{S4.19}$$

and

$$\sup_{\boldsymbol{\omega}} \frac{|tr(\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega}) - tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} = o_p(1). \tag{S4.20}$$

Note that Eqs.(S4.7)–(S4.10) do not include any $\widehat{\cdot}$ terms. From Eq.(S4.6) and Assumption 3, Eqs.(S4.7)–(S4.10) can be shown by using the same steps as in the proof of Theorem 1 of [3].

For proving Eq.(S4.19), by (S3.1), it is seen that

$$\begin{aligned}
&\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T \big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)\varepsilon|}{R_n(\boldsymbol{\omega})} \\
&\leq \eta_n^{-1} \frac{1}{2} \sup_{\boldsymbol{\omega}} \left| \varepsilon^T \Big[\big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) + \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) \Big] \varepsilon \right| \\
&\leq \eta_n^{-1} \frac{1}{2} \|\varepsilon\|^2 \cdot \sup_{\boldsymbol{\omega}} \lambda_{\max} \Big[\big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) + \big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T \big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) \Big] \\
&\leq \eta_n^{-1} \|\varepsilon\|^2 \cdot \sup_{\boldsymbol{\omega}} \lambda_{\max}\big(\mathbf{P}(\boldsymbol{\omega}) + \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) \lambda_{\max}\big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) \\
&\leq \eta_n^{-1} \|\varepsilon\|^2 \cdot \sup_{\boldsymbol{\omega}} \big[\lambda_{\max}(\mathbf{P}(\boldsymbol{\omega})) + \lambda_{\max}(\widehat{\mathbf{P}}(\boldsymbol{\omega}))\big] \cdot \sum_{m=1}^{M} \omega_m \lambda_{\max}(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}) \\
&\leq n\eta_n^{-1} \cdot \frac{\|\varepsilon\|^2}{n} \cdot \sup_{\boldsymbol{\omega}} \big[\lambda_{\max}(\mathbf{P}(\boldsymbol{\omega})) + \lambda_{\max}(\widehat{\mathbf{P}}(\boldsymbol{\omega}))\big] \cdot \max_{1 \leq m \leq M} \lambda_{\max}(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}) \\
&= o_p(1),
\end{aligned}$$

where the last step is from Eqs.(S4.5)–(S4.6), Assumption 3 and Lemma S3.3. By Lemma S3.3 and Assumption 3, we can prove Eqs.(S4.15)–(S4.18) in a similar way.

For Eq.(S4.12),

$$
\begin{aligned}
&\sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T\big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} \\
&\leq \eta_n^{-1}\|\boldsymbol{\mu}\| \cdot \sup_{\boldsymbol{\omega}} \|\big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big)^T\varepsilon\| \\
&\leq \eta_n^{-1} \cdot \|\boldsymbol{\mu}\| \cdot \sup_{\boldsymbol{\omega}} \lambda_{\max}\big(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega})\big) \cdot \|\varepsilon\| \\
&\leq n\eta_n^{-1} \cdot \frac{\|\boldsymbol{\mu}\|}{\sqrt{n}} \frac{\|\varepsilon\|}{\sqrt{n}} \cdot \max_{1 \leq m \leq M} \lambda_{\max}\big(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}\big) \\
&= o_p(1),
\end{aligned}
$$

where the last step is from Lemma S3.3 and Assumption 3. Similarly, we can verify Eqs.(S4.11), (S4.13)–(S4.14) by Lemma S3.3, Assumption 3 and Eqs.(S4.5)–(S4.6).

Now we consider the last Eq.(S4.20). Note that $\boldsymbol{\Omega}$ is a diagonal matrix,

$$
\begin{aligned}
&\sup_{\boldsymbol{\omega}} \frac{|tr(\mathbf{P}(\boldsymbol{\omega})\boldsymbol{\Omega}) - tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} \\
&= \sup_{\boldsymbol{\omega}} \frac{|tr[(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))\boldsymbol{\Omega}]|}{R_n(\boldsymbol{\omega})} \\
&\leq \eta_n^{-1} \sup_{\boldsymbol{\omega}} |tr(\mathbf{P}(\boldsymbol{\omega}) - \widehat{\mathbf{P}}(\boldsymbol{\omega}))|\lambda_{\max}(\boldsymbol{\Omega}) \\
&\leq n\eta_n^{-1} \max_{1 \leq m \leq M} \lambda_{\max}\big(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}\big)\lambda_{\max}(\boldsymbol{\Omega}) \\
&= o_p(1),
\end{aligned}
$$

where the last step is from Lemma S3.3 and Assumption 3. This completes the proof of Theorem 1. □

S5 Proof of Theorem 2

Proof. Note that

$$\widehat{C}_n(\boldsymbol{\omega})|_{\boldsymbol{\Omega}=\widehat{\boldsymbol{\Omega}}} = \widehat{C}_n(\boldsymbol{\omega}) + 2tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\widehat{\boldsymbol{\Omega}}) - 2tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega}).$$

From the result of Theorem 1, to prove Eq.(5) in Theorem 2, it suffices to prove that

$$\sup_{\boldsymbol{\omega}} \frac{|tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\widehat{\boldsymbol{\Omega}}) - tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})} = o_p(1). \tag{S5.21}$$

Let $\mathbf{Q}_{(m)} = diag(\rho_{11}^{(m)}, \ldots, \rho_{nn}^{(m)})$ and $\mathbf{Q}(\boldsymbol{\omega}) = \sum_{m=1}^{M} \omega_m \mathbf{Q}_{(m)}$. To prove Eq.(S5.21), we decompose the left-hand side of Eq.(S5.21) into four parts as follows.

$$\begin{aligned}
&\sup_{\boldsymbol{\omega}} \frac{|tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\widehat{\boldsymbol{\Omega}}) - tr(\widehat{\mathbf{P}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})}\\
&= \sup_{\boldsymbol{\omega}} \frac{|(\mathbf{Y} - \widehat{\mathbf{P}}_{(M^*)}\mathbf{Y})^T\widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{Y} - \widehat{\mathbf{P}}_{(M^*)}\mathbf{Y}) - tr(\widehat{\mathbf{Q}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})}\\
&= \sup_{\boldsymbol{\omega}} \frac{|(\boldsymbol{\mu} + \varepsilon)^T(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T\widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})(\boldsymbol{\mu} + \varepsilon) - tr(\widehat{\mathbf{Q}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})}\\
&\leq \sup_{\boldsymbol{\omega}} \frac{|\boldsymbol{\mu}^T(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T\widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})} + \sup_{\boldsymbol{\omega}} \frac{2|\varepsilon^T(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T\widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}|}{R_n(\boldsymbol{\omega})}\\
&\quad + \sup_{\boldsymbol{\omega}} \frac{|\varepsilon^T(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})^T\widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I} - \widehat{\mathbf{P}}_{(M^*)})\varepsilon|}{R_n(\boldsymbol{\omega})} + \sup_{\boldsymbol{\omega}} \frac{|tr(\widehat{\mathbf{Q}}(\boldsymbol{\omega})\boldsymbol{\Omega})|}{R_n(\boldsymbol{\omega})}\\
&\equiv \Xi_1 + \Xi_2 + \Xi_3 + \Xi_4.
\end{aligned}$$

Now define $\rho = \max_{1\leq m\leq M} \max_{1\leq i\leq n} |\rho_{ii}^{(m)}|$. From Assumption 4 and Lemma

S3.2, we have $\max_{1\le m\le M}|tr(\widehat{\mathbf{K}}_{(m)})| = \max_{1\le m\le M}|tr(\mathbf{K}_{(m)})| + O_p(n^{-\frac{1}{2}}\widetilde{q})$ and

$$\begin{aligned}
\rho &\le cn^{-1}\max_{1\le m\le M}|tr(\widehat{\mathbf{P}}_{(m)})| \\
&\le cn^{-1}\max_{1\le m\le M}|tr(\overline{\mathbf{P}}_{(m)})| + cn^{-1}\max_{1\le m\le M}|tr(\overline{\mathbf{P}}_{(m)}\widehat{\mathbf{K}}_{(m)})| + cn^{-1}\max_{1\le m\le M}|tr(\widehat{\mathbf{K}}_{(m)})| \\
&\le cn^{-1}\max_{1\le m\le M} rank(\overline{\mathbf{P}}_{(m)}) + cn^{-1}\frac{1}{2}\max_{1\le m\le M}\big[\lambda_{\max}\big(\overline{\mathbf{P}}_{(m)}\widehat{\mathbf{K}}_{(m)} + \widehat{\mathbf{K}}_{(m)}^T\overline{\mathbf{P}}_{(m)}\big) \\
&\qquad \cdot rank\big(\overline{\mathbf{P}}_{(m)}\widehat{\mathbf{K}}_{(m)} + \widehat{\mathbf{K}}_{(m)}^T\overline{\mathbf{P}}_{(m)}\big)\big] + cn^{-1}\max_{1\le m\le M}|tr(\widehat{\mathbf{K}}_{(m)})| \\
&\le cn^{-1}\widetilde{p} + cn^{-1}\cdot 2\widetilde{p}\cdot\lambda_{\max}(\overline{\mathbf{P}}_{(m)})\lambda_{\max}(\widehat{\mathbf{K}}_{(m)}) + cn^{-1}\max_{1\le m\le M}|tr(\widehat{\mathbf{K}}_{(m)})| \\
&= cn^{-1}\widetilde{p} + cn^{-1}\widetilde{p}\cdot O_p(1) + cn^{-1}\cdot O_p(h^{-\widetilde{q}} + n^{-\frac{1}{2}}\widetilde{q}) \\
&= O_p(n^{-1}\widetilde{p} + n^{-1}h^{-\widetilde{q}} + n^{-\frac{3}{2}}\widetilde{q}).
\end{aligned} \tag{S5.22}$$

It follows from Lemma S3.2, Assumptions 3–4, and Eqs.(S4.5) and (S5.22) that

$$\begin{aligned}
\Xi_1 &\le \eta_n^{-1}\sup_{\boldsymbol{\omega}}\lambda_{\max}(\widehat{\mathbf{Q}}(\boldsymbol{\omega}))\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\|^2 \\
&\le \eta_n^{-1}\rho\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\|^2 \\
&\le \eta_n^{-1}\rho\cdot\big[1+\lambda_{\max}(\widehat{\mathbf{P}}_{(M^*)})\big]^2\cdot\|\boldsymbol{\mu}\|^2 \\
&= \eta_n^{-1}\cdot O_p(n^{-1}\widetilde{p} + n^{-1}h^{-\widetilde{q}} + n^{-\frac{3}{2}}\widetilde{q})\cdot O_p(1)\cdot O_p(n) \\
&= O_p(\eta_n^{-1}\widetilde{p} + \eta_n^{-1}h^{-\widetilde{q}} + n^{-\frac{1}{2}}\eta_n^{-1}\widetilde{q}).
\end{aligned}$$

Using Lemma S3.2, Assumptions 3–4, and Eqs.(S4.5) and (S5.22), we obtain that

$$\begin{aligned}
\Xi_2 &\leq 2\eta_n^{-1}\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\|\cdot\sup_{\boldsymbol{\omega}}\|\widehat{\mathbf{Q}}(\boldsymbol{\omega})(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\varepsilon\|\\
&\leq 2\eta_n^{-1}\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\|\cdot\sup_{\boldsymbol{\omega}}\lambda_{\max}(\widehat{\mathbf{Q}}(\boldsymbol{\omega}))\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\varepsilon\|\\
&\leq 2\eta_n^{-1}\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\boldsymbol{\mu}\|\cdot\rho\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\varepsilon\|\\
&\leq 2\eta_n^{-1}\cdot\big(1+\lambda_{\max}(\widehat{\mathbf{P}}_{(M^*)})\big)\cdot\|\boldsymbol{\mu}\|\cdot\rho\cdot\big(1+\lambda_{\max}(\widehat{\mathbf{P}}_{(M^*)})\big)\cdot\|\varepsilon\|\\
&= 2\eta_n^{-1}\cdot O_p(1)\cdot O_p(n^{\frac{1}{2}})\cdot O_p(n^{-1}\widetilde{p}+n^{-1}h^{-\widetilde{q}}+n^{-\frac{3}{2}}\widetilde{q})\cdot O_p(1)\cdot O_p(n^{\frac{1}{2}})\\
&= O_p(\eta_n^{-1}\widetilde{p}+\eta_n^{-1}h^{-\widetilde{q}}+n^{-\frac{1}{2}}\eta_n^{-1}\widetilde{q}).
\end{aligned}$$

Using Lemma S3.2, Assumptions 3–4, and Eqs.(S4.5) and (S5.22), we have

$$\begin{aligned}
\Xi_3 &\leq \eta_n^{-1}\cdot\sup_{\boldsymbol{\omega}}\lambda_{\max}(\widehat{\mathbf{Q}}(\boldsymbol{\omega}))\cdot\|(\mathbf{I}-\widehat{\mathbf{P}}_{(M^*)})\varepsilon\|^2\\
&\leq \eta_n^{-1}\cdot\rho\big[1+\lambda_{\max}(\widehat{\mathbf{P}}_{(M^*)})\big]^2\cdot\|\varepsilon\|^2\\
&= \eta_n^{-1}\cdot O_p(n^{-1}\widetilde{p}+n^{-1}h^{-\widetilde{q}}+n^{-\frac{3}{2}}\widetilde{q})\cdot O_p(1)\cdot O_p(n)\\
&= O_p(\eta_n^{-1}\widetilde{p}+\eta_n^{-1}h^{-\widetilde{q}}).
\end{aligned}$$

Using Assumptions 3–4, and Eqs.(S4.5) and (S5.22), we have

$$\begin{aligned}
\Xi_4 &\leq \eta_n^{-1}\cdot n\sup_{\boldsymbol{\omega}}\lambda_{\max}(\widehat{\mathbf{Q}}(\boldsymbol{\omega}))\cdot\lambda_{\max}(\boldsymbol{\Omega})\\
&\leq \eta_n^{-1}\cdot n\cdot\rho\cdot\lambda_{\max}(\boldsymbol{\Omega})\\
&= \eta_n^{-1}\cdot n\cdot O_p(n^{-1}\widetilde{p}+n^{-1}h^{-\widetilde{q}}+n^{-\frac{3}{2}}\widetilde{q})\cdot O_p(1)\\
&= O_p(\eta_n^{-1}\widetilde{p}+\eta_n^{-1}h^{-\widetilde{q}}+n^{-\frac{1}{2}}\eta_n^{-1}\widetilde{q}).
\end{aligned}$$

Finally, it follows from Assumptions 3–4 that $\Xi_1 = o_p(1)$, $\Xi_2 = o_p(1)$, $\Xi_3 = o_p(1)$ and $\Xi_4 = o_p(1)$. Therefore, we have verified Eq.(S5.21) and this completes the proof.

□

Remark 1. In our current development, we focus on the setting of densely observed functional data with noise. Through our theoretical derivation, we find that the estimation error from FPCA mainly influences asymptotic optimality via the convergence rate of $\lambda_{\max}\big(\mathbf{P}_{(m)} - \widehat{\mathbf{P}}_{(m)}\big)$. When $\{X_i(t)\}$ are sparsely or irregularly observed, similar to cases encountered in longitudinal studies, we can similarly justify the optimality for this sparse and irregular setting. This is contingent on FPCA for sparsely or irregularly observed data yielding appropriate convergence rates for the estimator of transformed FPC scores $\{\widehat{\xi}_{il} - \xi_{il}\}$.

Bibliography

[1] Li, K.-C. (1987). Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set. *The Annals of Statistics*, 958–975.

[2] Speckman, P. (1988). Kernel smoothing in partial linear models. *Journal of the Royal Statistical Society: Series B (Methodological) 50*(3), 413–436.

[3] Wan, A. T., X. Zhang, and G. Zou (2010). Least squares model averaging by mallows criterion. *Journal of Econometrics 156*(2), 277–283.

[4] Wong, R. K., Y. Li, and Z. Zhu (2019). Partially linear functional additive models for multivariate functional data. *Journal of the American Statistical Association 114*(525), 406–418.