Semi-nonparametric Varying Coefficients Models for Imaging Genetics

Ting Li[1], Yang Yu[2], Xiao Wang[3], J.S. Marron[2] and Hongtu Zhu[4]

[1]*School of Statistics and Data Science, Shanghai University of Finance and Economics.*

[2]*Department of Statistics, University of North Carolina at Chapel Hill.*

[3]*Department of Statistics, Purdue University.*

[4]*Department of Biostatistics, University of North Carolina at Chapel Hill.*

Supplementary Material

The supplementary material contains tuning parameter selection, additional simulation results, additional real data analysis, and details of all the proofs.

S1 Auxillary Properties

In this section, we present auxiliary properties of the eigenfunctions, which greatly facilitate the theoretical analysis.

Given that these eigenfunctions $\{\varphi_{kk'}\}$ constitute an orthonormal basis of the $\mathcal{L}_2$ space, this foundation allows us to derive significant properties and the explicit Fourier expansions for our newly defined kernel $\widetilde{K}_{\mathbf{u}}$ and $W_\lambda \varphi_{kk'}$ using this basis.

Proposition S1. *For any* $\mathbf{u} \in \mathcal{U}$ *and* $k, k' = 1, 2, \ldots,$ *we have*

$$\widetilde{K}_{\mathbf{u}} = \sum_{k=1}^{\infty} \sum_{k'=1}^{\infty} \frac{\varphi_{kk'}(\mathbf{u})}{1 + \lambda/\tau_{kk'}} \varphi_{kk'} \quad \text{and} \quad W_\lambda \varphi_{kk'} = \frac{\lambda}{\lambda + \tau_{kk'}} \varphi_{kk'}. \tag{S1.1}$$

Furthermore, $\mathbb{E}_{\mathbf{u}}[\widetilde{K}_{\mathbf{u}} \circledast \widetilde{K}_{\mathbf{u}}] + W_\lambda = id$*, where* $\circledast$ *denotes the outer product between operators on* $\mathcal{H}$ *and id is an identity operator on* $\mathcal{H}$*.*

From the second-order Fréchet derivative of $\ell_{n,m,\lambda}(\mu)$, we can find that $\mathbb{E}_{\mathbf{u}}[\widetilde{K}_{\mathbf{u}} \circledast \widetilde{K}_{\mathbf{u}}] + W_\lambda$ is the expectation of the Hessian of the loss function, greatly facilitating the theoretical analysis.

Furthermore, we introduce several definitions and notations that are essential for our analysis. We define the functional

$$J(\mathcal{F}, \delta) := \int_0^\delta \sqrt{\log(1 + \mathcal{N}(\mathcal{F}, \|\cdot\|_{\sup}, \epsilon))} \mathrm{d}\epsilon + \delta \sqrt{\log(1 + \mathcal{N}(\mathcal{F}, \|\cdot\|_{\sup}, \delta)^2)}, \tag{S1.2}$$

where $\mathcal{N}(\mathcal{F}, \|\cdot\|_{\sup}, \epsilon)$ represents the ϵ-covering number of the function space $\mathcal{F}$ with respect to the supremum norm. The ϵ-covering number is a measure of the complexity of a function space, which is the minimal number of balls of radius ϵ required to cover the entire space $\mathcal{F}$ (Van der Vaart, 2000).

We then define

$$\mathcal{Q}_1 = \{f \in \mathcal{H}_z \otimes \mathcal{H}_s : f(\mathbf{z}, s) = \mathbf{z}^T \boldsymbol{\gamma}(s) \text{ for } \mathbf{z} \in \mathcal{Z},\ \boldsymbol{\gamma} \in \mathbb{R}^p,\ \|f\|_{\sup} \leq 1,$$

$$\|f\|_{\mathcal{H}_z \otimes \mathcal{H}_s} \leq (\lambda d(\lambda))^{-1/2}\},$$

$$\mathcal{Q}_2 = \left\{f \in \mathcal{H}_x \otimes \mathcal{H}_s : \|f\|_{\sup} \leq 1,\ \|f\|_{\mathcal{H}_x \otimes \mathcal{H}_s} \leq (\lambda d(\lambda))^{-1/2}\right\},$$

$$\mathcal{Q} = \{f = f_1 + f_2 : f_1 \in \mathcal{Q}_1,\ f_2 \in \mathcal{Q}_2,\ \|f\|_{\sup} \leq 1/2\}.$$

S2 Tuning Parameter Selection

S2.1 Selection of Smoothing Parameters

Choosing appropriate tuning parameters is crucial for effectively fitting the SVC model (1.1). This includes the regularization parameters $\boldsymbol{\lambda} = \lambda(\theta_1^{-1}, \ldots, \theta_{p+1}^{-1})^T$ and the kernel parameters of $K_x(\cdot,\cdot)$ and $K_s(\cdot,\cdot)$. The regularization parameters $\boldsymbol{\lambda}$ control the trade-off between the fit of the model and the variability of the estimated $\widehat{\boldsymbol{\gamma}}(\cdot)$ and $\widehat{h}(\cdot,\cdot)$. The GCV method is used to select the regularization parameters, while the selection of kernel parameters is discussed in Section S2.2.

The GCV method seeks $\boldsymbol{\lambda}$ by minimizing

$$V(\boldsymbol{\lambda}) = \frac{nm\mathbf{Y}^T(\mathbf{K} + nm\lambda\mathbf{I})^{-2}\mathbf{Y}}{[\mathrm{tr}(\mathbf{K} + nm\lambda\mathbf{I})^{-1}]^2}. \tag{S2.3}$$

Minimizing the function $V(\boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ using grid search on a manually specified subset of the parameter space is computationally expensive. Even when using more efficient search techniques like golden section and bisection search, minimizing the function can still be computationally expensive, especially when the number of coefficient functions is large. To overcome this issue, we modify the BFGS algorithm to optimize the criterion scores with multiple regularization parameters.

The BFGS algorithm is one of the most popular quasi-Newton methods, which requires in each iteration only the evaluation of the gradient of the objective function. The Hessian matrix of second derivatives is approximated by a symmetric positive definite matrix $\boldsymbol{B}$. Starting from an initial value $\boldsymbol{B}_0$, we update $\boldsymbol{B}$ in each iteration by adding information about the curvature of the objective function obtained in the previous iteration. In fact, since we only need the inverse of $\boldsymbol{B}$ (denoted by $\boldsymbol{H}$) in the algorithm, we can start with $\boldsymbol{H}_0$ and update $\boldsymbol{H}$ instead in each iteration. This trick not only saves more computational cost per iteration but also increases numerical stability of the algorithm in practice.

In the SVC model (1.1), the regularization parameters $\boldsymbol{\lambda}$ are composed of two parts: a main regularization parameter λ for the entire function $\mu(\mathbf{x}, \mathbf{z}, s)$ and subsidiary regularization parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_{p+1})$ for each of the component functions $\boldsymbol{\gamma}(s) = (\gamma_1(s), \ldots, \gamma_p(s))$ and $h(\mathbf{x}, s)$. To avoid dealing with a constrained optimization problem ($\boldsymbol{\theta} \geq 0$), we use the log-transformation of $\boldsymbol{\theta}$, denoted by $\boldsymbol{\eta}$, and re-write $V(\boldsymbol{\lambda})$ as $V(\lambda, \boldsymbol{\eta})$. The proposed BFGS algorithm minimizes $V(\boldsymbol{\lambda})$ with respect to λ and $\boldsymbol{\eta}$ by alternatively performing the following two steps until convergence: (i) fix $\boldsymbol{\eta}$ and minimize the objective function with respect to λ and (ii) update $\boldsymbol{\eta}$ by performing an iteration of the BFGS algorithm. We summarize the algorithm in Algorithm 1. Specifically, the initial value of $\eta_0 = (1, 1, \ldots, 1)$, and the initial value of $\mathbf{H}_0$ is the identity matrix, the tolerance level $\tilde{C} = 10^{-5}$ and the step size $C = 10^{-4}$.

The convergence conditions in Algorithm 1 are directly inherited from Gu and Wahba (1991). In addition, we follow their approach of choosing a good starting value. We describe this procedure in the following algorithm.

S2.2 Selection of Kernel Parameters

We examine the kernel parameters of certain kernel functions. For example, the γ-th order polynomial kernel $K(\mathbf{x}_1, \mathbf{x}_2) = (\mathbf{x}_1^T\mathbf{x}_2 + \rho)^\gamma$ specifies an RKHS on $\mathbb{R}^p$ spanned by all monomials of the components of $\mathbf{x}$ with orders up to γ. Both γ and ρ are kernel parameters. The parameter γ controls the complexity of the generated RKHS. A larger γ allows higher-order basis functions and therefore leads to a more complex function space. The parameter ρ, on the other hand, controls the influence of monomials with order lower than

Algorithm 1 Smoothing parameters selection

```
Require: The response vector $\mathbf{Y}$, the matrices $\mathbf{K}_\nu$, $\nu = 1, \dots, p+1$, and the starting
    values $\boldsymbol{\eta}_0$ and $\mathbf{H}_0$, a pre-determined tolerance level $\tilde{C}$, a constant $C$:
 1: Initialization: Set $\Delta\boldsymbol{\eta} = 0$, $\boldsymbol{\eta}_- = \boldsymbol{\eta}_0$, $\mathbf{H}_- = \mathbf{H} = \mathbf{H}_0$, $V_- = \infty$ and $\mathbf{g}_- = \mathbf{0}$.
 2: for $V_- - V < \tilde{C}(1+V)$ and $\|\mathbf{g}\|_\infty < \sqrt{\tilde{C}}(1+V)$, or $\|\mathbf{g}\|_\infty < \tilde{C}$. do
 3:     For the current $\boldsymbol{\eta} = \boldsymbol{\eta}_- + \Delta\boldsymbol{\eta}$, compute $\mathbf{K} = \sum_{\nu=1}^{p+1} \theta_\nu \mathbf{K}_\nu$.
 4:     Fix the current $\boldsymbol{\eta}$ and minimize $V(\lambda, \boldsymbol{\eta})$ with respect to $\lambda$.
 5:     if $V > V_- + C\mathbf{g}_-^T \Delta\boldsymbol{\eta}$ then
 6:         $\Delta\boldsymbol{\eta} = \Delta\boldsymbol{\eta}/2$, go back to step 3 .
 7:     end if
 8:     Set $\mathbf{g}_- = \mathbf{g}$ and update the gradient $\mathbf{g} = \partial V(\lambda, \boldsymbol{\eta})/\partial \boldsymbol{\eta}$ where $\lambda$ is set to be the
    minimizer obtained in the previous step. Calculate $\Delta\mathbf{g} = \mathbf{g} - \mathbf{g}_-$.
 9:     if $\Delta\boldsymbol{\eta} \neq \mathbf{0}$ then
10:         $\mathbf{H}_- = \mathbf{H}$ and update $\mathbf{H}$ following the rule
                $$\mathbf{H} = \left(\mathbf{I} - \frac{\Delta\boldsymbol{\eta}\Delta\mathbf{g}^T}{\Delta\boldsymbol{\eta}^T\Delta\mathbf{g}}\right) \mathbf{H}_- \left(\mathbf{I} - \frac{\Delta\mathbf{g}\Delta\boldsymbol{\eta}^T}{\Delta\boldsymbol{\eta}^T\Delta\mathbf{g}}\right) + \frac{\Delta\mathbf{g}\Delta\mathbf{g}^T}{\Delta\boldsymbol{\eta}^T\Delta\mathbf{g}}. \tag{S2.4}$$
11:         Calculate the increment $\Delta\boldsymbol{\eta} = -\mathbf{H}^{-1}\mathbf{g}$.
12:     end if
13:     Set $\boldsymbol{\eta}_- = \boldsymbol{\eta}$, $V_- = V$
14: end for
15: Calculate $\mathbf{K} = \sum_{\nu=1}^{p+1} \theta_\nu \mathbf{K}_\nu$ with the optimal $\boldsymbol{\theta}$ and minimize $V(\lambda, \boldsymbol{\eta})$ to obtain the
    optimal $\lambda$.
Ensure: The optimal $\boldsymbol{\theta}$ and optimal $\lambda$.
```

```
Algorithm 2 Starting value
Require: Set $\tilde{\theta}_\nu = (\mathrm{tr}(\mathbf{K}_\nu))^{-1}$ for $\nu = 1, \ldots, p+1$ and then fit model (1.1) with $\lambda$
    chosen by minimizing $V(\lambda, \tilde{\boldsymbol{\theta}})$. Calculate the estimate of the parameter matrix $\mathbf{C}$.
Ensure: Set the starting values of Algorithm 1 to be $\theta_\nu = \log(\tilde{\theta}_\nu^2 \mathbf{c}^T \mathbf{K}_\nu \mathbf{c})$ for all $\nu$.
```

γ on the approximation of the function. Another example is the Gaussian kernel. The kernel function is defined as $K(\mathbf{x}_1, \mathbf{x}_2) = \exp[-\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2/\rho]$ or $K(\mathbf{x}_1, \mathbf{x}_2) = \exp[-\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2/(2\sigma_\rho^2)]$, where ρ and σ_ρ are, respectively, the kernel parameter and the spread parameter.

We focus on the selection of the two spread parameters, σ_x and σ_s, of $K_x(\cdot, \cdot)$ and $K_s(\cdot, \cdot)$ below. As discussed in Chaudhuri and Marron (2000) and Wang et al. (2003), there is a similarity between the spread parameters of the Gaussian kernel in the curve estimation problem and the aperture scale of the Gaussian function in the scale space theory. For the SVC model, each design point $\mathbf{u}_{ij} = (\mathbf{x}_i, \mathbf{z}_i, s_j)$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$ can be considered to be a single light point in the scale space, whose distribution can be expressed as $\mathit{\Delta}(\mathbf{u} - \mathbf{u}_{ij})$, where $\mathit{\Delta}$ is the Dirac delta function. The whole training data set has the light density function $\mathcal{P}(\mathbf{u}) = \sum_{i=1}^n \sum_{j=1}^m y_{ij} \mathit{\Delta}(\mathbf{u} - \mathbf{u}_{ij})$. Convoluting with the Gaussian function $g(\mathbf{u}, \rho) = (\pi\rho)^{-(p+q+1)/2} \exp(-\|\mathbf{u}\|_2^2/\rho)$, we obtain the ρ-indexed image in scale space

$$\mathcal{I}(\mathbf{u},\rho)=\mathcal{P}(\mathbf{u})*g(\mathbf{u},\rho)=(\pi\rho)^{-(p+q+1)/2}\sum_{i=1}^{n}\sum_{j=1}^{m}y_{ij}K(\mathbf{u},\mathbf{u}_{ij}),\qquad\text{(S2.5)}$$

where $*$ represents convolution. This is very similar to estimation equation (2.4), which motivates the idea to investigate the influence of the spread parameters on the SVC estimates through a scale space point of view.

Following the scale space theory, there is an appropriate stable region of spread parameters in $(0,\infty)\times(0,\infty)$, in which any pair of (σ_x,σ_s), accompanied by suitable estimates of $\boldsymbol{\lambda}$, will give a good estimate of unknown parameters in our SVC model. Denote $\text{MSE}_\mu=(nm)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\{\mathbf{z}_i^T\widehat{\boldsymbol{\gamma}}(s_j)+\widehat{h}(\mathbf{x}_i,s_j)-y_{ij}\}^2$. Based on the changing rule of generalization performance with spread parameters, we can obtain suitable spread parameters by minimizing the MSE_μ. We present a simulation example in the supplementary material for illustration of the changing rule, but we propose the following algorithm to select the appropriate kernel parameters on the standard deviation scale.

Although the algorithm seems simple, scale space theory provides the theoretical basis that we can obtain suitable spread parameters.

Algorithm 3 Kernel parameters selection

```
Require: Initialization. Set MSE to be a large value $L$ and $\sigma_s = \sigma_x = \sigma_0$, where $\sigma_0$ is
    usually a very small value (e.g., $\sigma_0 = 0.0001$). Set $\Delta\sigma_s = \min_{j,j'=1,\ldots,m} \|s_j - s_{j'}\|_2$
    and $\Delta\sigma_x = \min_{i,i'=1,\ldots,n} \|\mathbf{x}_i - \mathbf{x}_{i'}\|_2$.
1: Train the SVC model and calculate $\text{MSE}_\mu$.
2: If $\text{MSE}_\mu \leq \text{MSE}$, set $\text{MSE} = \text{MSE}_\mu$ and $\sigma_x = \sigma_x + \Delta\sigma_x$ and go to step 1. Otherwise,
    continue.
3: Set $\sigma_s = \sigma_s + \Delta\sigma_s$, train the SVC model and calculate $\text{MSE}_\mu$.
4: If $\text{MSE}_\mu \leq \text{MSE}$, set $\text{MSE} = \text{MSE}_\mu$ and go to step 3. Otherwise, continue.
Ensure: Stop and output the current $(\sigma_x, \sigma_s)$.
```

S3 Additional Simulation Results

In this section, we present additional simulation results to investigate the finite sample performance of the proposed method for the sensitivity analysis of the kernel parameters.

To depict the changing rule of MSE with kernel parameters ρ_x ad ρ_s, we simulate an example based on the true functions $\gamma_1(s) = 10s^3 - 15s^2 + 5s + 1$, $\gamma_2(s) = 10s^6 - 30s^5 + 25s^4 - 5s^2 + 5/21 + \sin(6\pi s)$, and $h(\mathbf{x}, s) = 2\cos(2\pi(x_1 - s)) + s \cdot \sin(2\pi(x_1 + x_2))$, an equally spaced design $s_j = j - 1/(m-1)$ for $j = 1, \ldots, m = 10$, and training data $z_{i1} = 1$, $z_{i2} \sim N(0, 1)$, $(z_{i1}, z_{i2})^T \sim U[0, 1]^2$, and $y_{ij} = \mathbf{z}_i^T \boldsymbol{\gamma}(s_j) + h(\mathbf{x}_i, s_j) + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ with $\sigma_\epsilon = 0.1$ for $i = 1, \ldots, n = 50$ and $j = 1, \ldots, 10$. We let both of the two spread

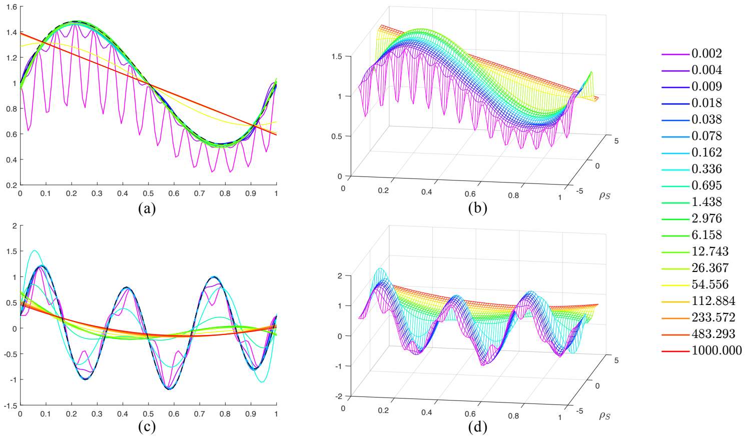

Figure S3.1: Illustrations of the influence of the kernel parameter ρ_s on $\widehat{\gamma}$. One can observe that when the kernel parameters are taken in a certain region, the estimation performances remain unchanged.

parameters vary in a wide range and fit the SVC model. Figure S3.1 displays the estimation results. The left column shows the true underlying functions as the dashed lines and a family of SVC estimates as the colored solid lines. The right column shows the corresponding empirical scale space surfaces. The top row corresponds to $\gamma_1(s)$, while the bottom row corresponds to $\gamma_2(s)$. It can be observed that when the kernel parameter ρ_s is taken in a certain region, the estimates are very similar.

Based on $R = 50$ independent replications, we calculate the Mean

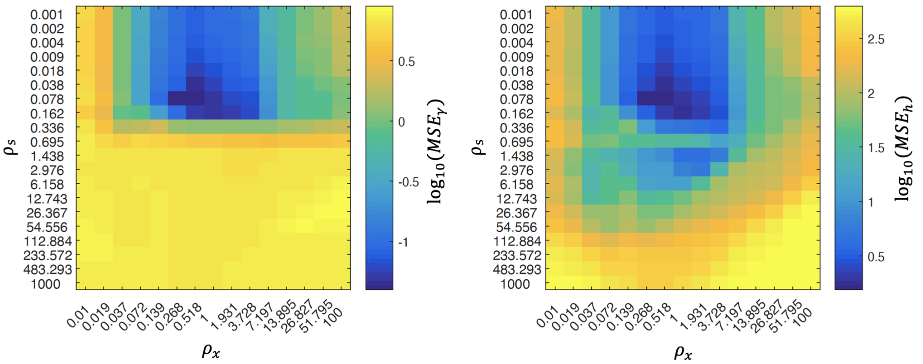

Figure S3.2: Displays of the $\log_{10}$(MSE) of $\widehat{\boldsymbol{\gamma}}$ (left) and $\widehat{h}$ (right) over (ρ_x, ρ_s). The stable regions of kernel parameters are shown in dark blue colors.

Squared Errors (MSE) of $\widehat{\boldsymbol{\gamma}}$, $\widehat{h}$ and μ, defined by

$$\text{MSE}_\gamma := (Rmp)^{-1} \sum_{r=1}^{R} \sum_{j=1}^{m} \sum_{\nu=1}^{p} \{\widehat{\gamma}_\nu^{(r)}(s_j) - \gamma_\nu(s_j)\}^2, \tag{S3.1}$$

$$\text{MSE}_h := (Rmp)^{-1} \sum_{r=1}^{R} \sum_{i=1}^{n} \sum_{j=1}^{m} \{\widehat{h}^{(r)}(\mathbf{x}_i, s_j) - h(\mathbf{x}_i, s_j)\}^2, \tag{S3.2}$$

$$\text{MSE}_\mu := (Rmp)^{-1} \sum_{r=1}^{R} \sum_{i=1}^{n} \sum_{j=1}^{m} \{\mathbf{z}_i^T \widehat{\boldsymbol{\gamma}}(s_j) + \widehat{h}^{(r)}(\mathbf{x}_i, s_j) - y_{ij}\}^2, \tag{S3.3}$$

where the superscript (r) indicates the estimates from the r-th run. The values of $\log_{10}(\text{MSE}_\gamma)$ and $\log_{10}(\text{MSE}_h)$ over (ρ_x, ρ_s) are shown in Figure S3.2. Any spread parameter pair (ρ_x, ρ_s) in the dashed box is appropriate for training the SVC model. We term this region as the stable region.

To investigate the influence of the spread parameters on the SVC estimates through a scale space point of view, we plot a family of GCV-tuned SVC estimates $\widehat{\boldsymbol{\gamma}}(s; \rho_s)$, indexed by ρ_s, and overlay them in Figures S3.1(a) and S3.1(c). Figures S3.1(b) and S3.1(d) present the empirical scale space surfaces (Chaudhuri and Marron, 2000) of the same family of estimates, arranged one behind the other in an increasing order of ρ_s. We can clearly observe the existence of a certain range of ρ_s, within which the estimates are stable. This corresponds to the stable region shown in Figure S3.2.

Figure S3.1 also demonstrates that as the kernel parameter increases, the estimates become more simplified and structures disappear monotonically. The relationship between this phenomenon and scale space is discussed in Chaudhuri and Marron (2000) for the kernel smoothers, such as the Priestley-Chao estimate and the Gasser-Müller estimate. As noted in Section 3.1 of Härdle (1990), the general form of the kernel smoother is $\widehat{f}(x) = \sum_{i=1}^{n} W_{hi}(x)y_i$, where $\sum_{i=1}^{n} W_{hi}(x) = 1$ and the explicit form depends on the particular method. One major difference between the SVC estimate and the kernel smoother is that the former is not a weighted average of $\{y_i : 1 \leq i \leq n\}$ as is the latter. As a result, we can observe that the SVC estimates are shrunk to zero for small kernel parameters as illustrated in Figures S3.1. To see why, notice that the SVC estimate can be written in the form

$$\widehat{f}(x) = \sum_{i=1}^{n} W'_{\rho i}(x)y_i = \sum_{i=1}^{n+1} W'_{\rho i}(x)y_i \tag{S3.4}$$

where $y_{n+1} := 0$ and $\sum_{i=1}^{n} W'_{\rho i} = C$. Without loss of generality, we assume $C < 1$ and $W'_{\rho,n+1} = 1 - C$. This indicates that the SVC estimate can be viewed as a weighted average of $\{y_i : 1 \leq i \leq n\}$ and 0, which explains the shrinkage effect.

S4 Additional Real Data Analysis

In this section, we present additional results about the 27 significant blocks for both the left and right hippocampi.

Figure S4.3 presents the median positions and p-values of the common 27 blocks and the range and p-values of the top 10 significant blocks for the left and right hippocampi. It shows that the well-known block 19q13.32 region on the 19th chromosome is identified to be important for both the left and right hippocampi. Among the 27 blocks, 21 blocks are associated with cognitive performance such as age-related cognitive decline, language ability, cognitive empathy and AD, 20 blocks are associated with education attainment, 16 blocks contain SNPs associated with PHF tau protein, 15 blocks are associated with brain measurement such as brain morphology, brain shape and brain volume, 12 blocks are associated with neurofibrillary tangles, 9 blocks are associated with Amyloid-beta, and 8 blocks are associated with memory performance such as logical memory, memory decline and immediate memory. For example, rs199852994 on chromosome 4, rs1925531 on chromosome 10, rs58935614 on chromosome 11, rs8054299 on

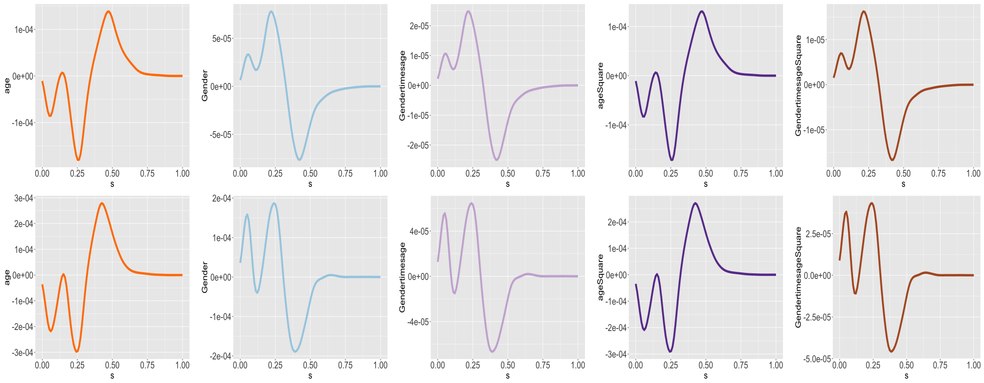

Figure S4.1: Age, Gender, Age·Gender, Age^2, and Age^2·Gender effects on the left (the first row) and right (the second row) hippocampi.

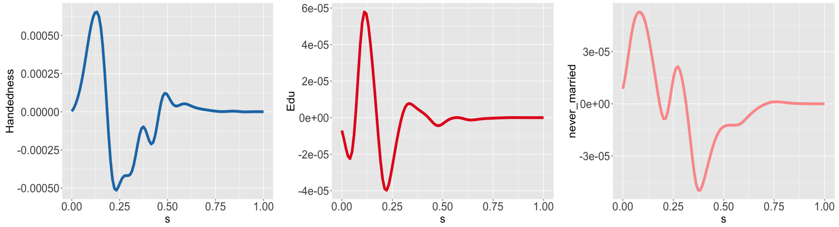

Figure S4.2: Estimates of Handedness for the left hippocampus, and estimates of Education and Never Married for the right hippocampus (from left to right).

chromosome 16, rs7519289 on chromosome 1, rs114545261 on chromosome 6, and rs12225836 on chromosome 11 have been found to be associated with education attainment (Okbay et al., 2022).

S5 Auxillary Results

Let $\Delta\mu = (\Delta g, \Delta h)$ and $\Delta\mu_j = (\Delta g_j, \Delta h_j)$ for $j = 1, 2$. The first- and second-order Fréchet derivatives of $\ell_{n,m,\lambda}(\mu)$ with respect to μ are, respec-

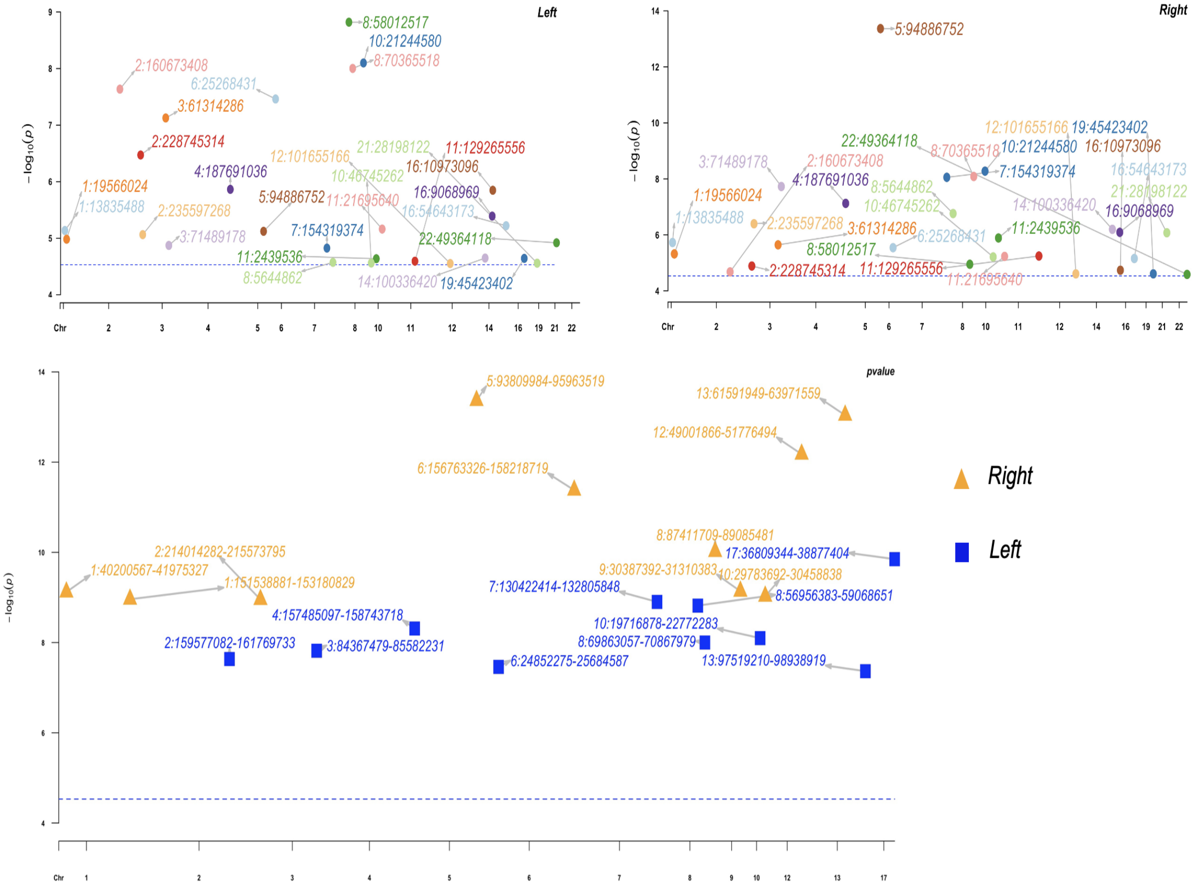

Figure S4.3: Manhattan plots of 27 significant common blocks for the left and right hippocampi (the first row) and of top 10 significant blocks for the left(blue rectangle) and right(orange triangle) hippocampi (the second row).

tively, given by

$$\begin{aligned} D\ell_{n,m,\lambda}(\mu)\Delta\mu &= -\frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}[y_i(s_j)-\mu(u_{ij})]\langle \widetilde{K}_{\mathbf{u_{ij}}}, \Delta\mu\rangle_{\widetilde{\mathcal{H}}} + \langle W_\lambda \mu, \Delta\mu\rangle_{\widetilde{\mathcal{H}}}, \\ D^2\ell_{n,m,\lambda}(\mu)\Delta\mu_1\Delta\mu_2 &= \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\langle \widetilde{K}_{\mathbf{u_{ij}}}, \Delta\mu_1\rangle_{\widetilde{\mathcal{H}}}\langle \widetilde{K}_{\mathbf{u_{ij}}}, \Delta\mu_2\rangle_{\widetilde{\mathcal{H}}} + \langle W_\lambda \Delta\mu_1, \Delta\mu_2\rangle_{\widetilde{\mathcal{H}}}. \end{aligned}$$

Denote $T = (Y, X, Z, S) \in \mathcal{T}$ as the data vector. Let $\psi_{n,m}(T; f)$ be a function over $\mathcal{T}\times\mathcal{H}$. Define $H_{n,m}(f) = (nm)^{-1/2}\sum_{i=1}^{n}\sum_{j=1}^{m}[\psi_{n,m}(T_{ij}; f)\widetilde{K}_{\mathbf{u_{ij}}} - E\{\psi_{n,m}(T_{ij}; f)\widetilde{K}_{\mathbf{u_{ij}}}\}]$. The following lemma proves a concentration inequality as a preliminary step in obtaining the convergence rate.

Lemma 1. *Suppose that Assumptions 1 and 2 hold, the function* $\psi_{n,m}(T; f)$ *satisfies* $\psi_{n,m}(T_{ij}; 0) = 0$, *and*

$$|\psi_{n,m}(T; f_1) - \psi_{n,m}(T; f_2)| \le C_\varphi^{-1} d(\lambda)^{-1/2}\|f_1 - f_2\|_{sup} \quad \text{for any } f_1, f_2 \in \mathcal{Q}\ .$$

Then as $n, m \to \infty$, *we have*

$$\sup_{f\in\mathcal{Q}}\|H_{n,m}(f)\|_{\widetilde{\mathcal{H}}} = O_p(\sqrt{\log\log\big(nmJ(\mathcal{Q},1)\big)}(J(\mathcal{Q},1) + (nm)^{-1/2})).$$

Proof. For any $g, f \in \mathcal{Q}$, it follows directly from (S6.4) that

$$\|(\psi_{n,m}(T; f) - \psi_{n,m}(T; g))\widetilde{K}_{\mathbf{u}}\|_{\widetilde{\mathcal{H}}} \le C_\varphi^{-1} d(\lambda)^{-1/2} C_\varphi d(\lambda)^{1/2}\|f - g\|_{\sup} = \|f - g\|_{\sup}.$$

By Theorem 3.5 of Pinelis (1994), for any $t > 0$,

$$P(\|H_{n,m}(f) - H_{n,m}(g)\|_{\widetilde{\mathcal{H}}} \ge t) \le \exp\Big(-\frac{t^2}{8\|f - g\|_{\sup}^2}\Big).$$

Then by Lemma 8.1 of Kosorok (2007), we have $\big\|\|H_{n,m}(f) - H_{n,m}(g)\|_{\widetilde{\mathcal{H}}}\big\|_{\psi_2} \leq 8\|f-g\|$, where $\|\cdot\|_{\psi_2}$ denotes the Orlicz norm associated with $\psi_2(s) = \exp(s^2)-1$. It follows by Theorem 8.4 of Kosorok (2007) that for arbitrary $\delta > 0$, there exists a universal positive constant $C > 0$ that

$$\begin{aligned}
&\Big\| \sup_{g,f\in\mathcal{Q},\|f-g\|_{\sup}\leq\delta} \|H_{n,m}(f) - H_{n,m}(g)\|_{\widetilde{\mathcal{H}}} \Big\|_{\psi_2} \\
\leq\ & C\Big(\int_0^\delta \sqrt{\log(1+\mathcal{N}(\mathcal{F},\|\cdot\|_{\sup},\epsilon))}\mathrm{d}\epsilon + \delta\sqrt{\log(1+\mathcal{N}(\mathcal{F},\|\cdot\|_{\sup},\delta)^2)}\Big) = CJ(\mathcal{Q},\delta).
\end{aligned}$$

Therefore, by Lemma 8.1 of Kosorok (2007),

$$P(\sup_{f\in\mathcal{Q},\|f\|_{\sup}\leq\delta} \|H_{n,m}(f)\|_{\widetilde{\mathcal{H}}} \geq t) \leq 2\exp\Big(-\frac{t^2}{C^2J^2(\mathcal{Q},\delta)}\Big).$$

Let $\delta = 1, \epsilon^{-1} = \sqrt{nm}J(\mathcal{Q},1), Q_\epsilon = -\log\epsilon - 1$ and $T_{nm} = \sqrt{\log\log(nmJ(\mathcal{Q},1))}$.

Then we have

$$\begin{aligned}
&P\Big(\sup_{f\in\mathcal{Q}} \frac{\sqrt{nm}\|H_{n,m}(f)\|_{\widetilde{\mathcal{H}}}}{\sqrt{nm}J(\mathcal{Q},1)+1} \geq T_{nm}\Big) \\
\leq\ & P\Big(\sup_{\|f\|_{\sup}\leq J(\mathcal{Q},1)^{-1}(nm)^{-1/2}} \frac{\sqrt{nm}\|H_{n,m}(f)\|_{\widetilde{\mathcal{H}}}}{\sqrt{nm}J(\mathcal{Q},1)+1} \geq T_{nm}\Big) \\
&+\sum_{l=1}^{Q_\epsilon} P\Big(\sup_{J(\mathcal{Q},1)^{-1}(nm)^{-1/2}\exp(l+1)\leq\|f\|_{\sup}\leq J(\mathcal{Q},1)^{-1}(nm)^{-1/2}\exp(l+1)} \frac{\sqrt{nm}\|H_{n,m}(f)\|_{\widetilde{\mathcal{H}}}}{\sqrt{nm}J(\mathcal{Q},1)+1} \geq T_{nm}\Big) \\
\leq\ & P\Big(\sup_{\|f\|_{\sup}\leq J(\mathcal{Q},1)^{-1}(nm)^{-1/2}} \sqrt{nm}\|H_{n,m}(f)\|_{\widetilde{\mathcal{H}}} \geq T_{nm}\Big) \\
&+\sum_{l=1}^{Q_\epsilon} P\Big(\sup_{\|f\|_{\sup}\leq J(\mathcal{Q},1)^{-1}(nm)^{-1/2}\exp(l+1)} \sqrt{nm}\|H_{n,m}(f)\|_{\widetilde{\mathcal{H}}} \geq T_{nm}(1+\exp(l))\Big) \\
\leq\ & 2\exp(-T_{nm}^2/C^2) + \sum_{l=1}^{Q_\epsilon} 2\exp(-T_{nm}^2/(C^2\exp(2))) \\
\leq\ & 2(Q_\epsilon+2)\exp\Big(-\frac{T_n^2}{C^2\exp(2)}\Big) \to 0.
\end{aligned}$$

This completes the proof. □

S6 Proofs

Proof of Proposition S1. Assuming that $\widetilde{K}_{\mathbf{u}} = \sum_{k,k'} a_{kk'}\varphi_{kk'}$, we have

$$\begin{aligned} a_{kk'} &= \langle \widetilde{K}_{\mathbf{u}}, \varphi_{kk'}\rangle_{\mathcal{L}_2} = \langle \widetilde{K}_{\mathbf{u}}, \varphi_{kk'}\rangle_{\widetilde{\mathcal{H}}} - \lambda\langle \widetilde{K}_{\mathbf{u}}, \varphi_{kk'}\rangle_{\mathcal{H}} \\ &= \varphi_{kk'}(\mathbf{u}) - \lambda a_{kk'}/\tau_{kk'} \end{aligned}$$

Solving for $a_{kk'}$, it is easy to have $a_{kk'} = \varphi_{kk'}(\mathbf{u})(1+\lambda/\tau_{kk'})^{-1}$ and the expansion of $\widetilde{K}_{\mathbf{u}}$ then follows.

To obtain the form of $W_\lambda\varphi_{\ell\ell'}$, notice that, for any $\mu, \tilde{\mu} \in \mathcal{H}$,

$$\langle W_\lambda\mu, \tilde{\mu}\rangle_{\widetilde{\mathcal{H}}} = \lambda\langle \mu, \tilde{\mu}\rangle_{\mathcal{H}}, \quad \langle W_\lambda\mu, \tilde{\mu}\rangle_{\widetilde{\mathcal{H}}} = \langle W_\lambda\mu, \tilde{\mu}\rangle_{\mathcal{L}_2} + \lambda\langle W_\lambda\mu, \tilde{\mu}\rangle_{\mathcal{H}}.$$

Combining the above two equations gives us

$$\langle W_\lambda\mu, \tilde{\mu}\rangle_{\mathcal{L}_2} = \lambda\langle (\mathrm{id} - W_\lambda)\mu, \tilde{\mu}\rangle_{\mathcal{H}}. \tag{S6.1}$$

Assuming that $W_\lambda\varphi_{\ell\ell'} = \sum_{k,k'} b_{kk'}\varphi_{kk'}$, we have

$$b_{\ell\ell'} = \langle W_\lambda\varphi_{\ell\ell'}, \varphi_{\ell\ell'}\rangle_{\mathcal{L}_2} = \lambda\langle (\mathrm{id} - W_\lambda)\varphi_{\ell\ell'}, \varphi_{\ell\ell'}\rangle_{\mathcal{H}} = \lambda/\tau_{\ell\ell'} - \lambda b_{\ell\ell'}/\tau_{\ell\ell'}.$$

Solving for $w_{\ell\ell'}$ gives us $b_{\ell\ell'} = \lambda/(\lambda+\tau_{\ell\ell'})$. For an index pair $(k,k') \neq (\ell,\ell')$, we can similarly obtain

$$b_{kk'} = \langle W_\lambda\varphi_{\ell\ell'}, \varphi_{kk'}\rangle_{\mathcal{L}_2} = \lambda\langle (\mathrm{id} - W_\lambda)\varphi_{\ell\ell'}, \varphi_{kk'}\rangle_{\mathcal{H}} = 0.$$

Therefore $W_\lambda \varphi_{\ell\ell'} = \frac{\lambda}{\lambda+\tau_{\ell\ell'}} \varphi_{\ell\ell'}$ holds.

Meanwhile, for any arbitrarily chosen $\mu, \tilde{\mu} \in \mathcal{H}$, we have

$$\begin{aligned}\langle (\mathbb{E}_{\mathbf{u}}[\widetilde{K}_{\mathbf{u}} \circledast \widetilde{K}_{\mathbf{u}}] + W_\lambda)\mu, \tilde{\mu}\rangle_{\widetilde{\mathcal{H}}} &= \langle \mathbb{E}_{\mathbf{u}}[\widetilde{K}_{\mathbf{u}} \circledast \widetilde{K}_{\mathbf{u}}]\mu, \tilde{\mu}\rangle_{\widetilde{\mathcal{H}}} + \langle W_\lambda \mu, \tilde{\mu}\rangle_{\widetilde{\mathcal{H}}} \\ &= \mathbb{E}_{\mathbf{u}}[\mu(\mathbf{u})\tilde{\mu}(\mathbf{u})] + \lambda \langle \mu, \tilde{\mu}\rangle_{\mathcal{H}} \\ &= \langle \mu, \tilde{\mu}\rangle_{\widetilde{\mathcal{H}}}\end{aligned}$$

Proof of Theorem 1. Let $\langle \cdot, \cdot \rangle$ denote the inner product on $\mathcal{H}$ and $\mathbf{u}_{ij} = (\mathbf{x}_i, \mathbf{z}_i, s_j)$. We can decompose μ into a sum of two orthogonal functions, one lying in $\operatorname{span}\{K(\mathbf{u}_{ij}, \cdot) : i = 1, \dots, n;\ j = 1, \dots, m\}$ and the other one, $v(\mathbf{u})$, lying in the orthogonal complement:

$$\mu(\mathbf{u}) = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} K(\mathbf{u}_{ij}, \mathbf{u}) + v(\mathbf{u}). \tag{S6.2}$$

At each training data point $\mathbf{u}_{i'j'}$, we have

$$\mu(\mathbf{u}_{i'j'}) = \left\langle \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} K(\mathbf{u}_{ij}, \cdot), K(\mathbf{u}_{i'j'}, \cdot) \right\rangle_{\mathcal{H}} = \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} K(\mathbf{u}_{ij}, \mathbf{u}_{i'j'}) \tag{S6.3}$$

which is independent of v. Therefore the empirical risk term in (2.3) is independent of v as well.

For the penalty term, we can write

$$\begin{aligned}\|P^\nu \mu\|_{\mathcal{H}}^2 &= \left\| \sum_{i=1}^{n} \sum_{j=1}^{m} \theta_\nu c_{ij} K_\nu((z_{i\nu}, s_j), \cdot) + P^\nu v \right\|_{\mathcal{H}}^2 = \left\| \sum_{i=1}^{n} \sum_{j=1}^{m} \theta_\nu c_{ij} K_\nu((z_{i\nu}, s_j), \cdot) \right\|_{\mathcal{H}}^2 + \|P^\nu v\|_{\mathcal{H}}^2 \\ &\geq \left\| \sum_{i=1}^{n} \sum_{j=1}^{m} \theta_\nu c_{ij} K_\nu((z_{i\nu}, s_j), \cdot) \right\|_{\mathcal{H}}^2,\end{aligned}$$

which implies that the penalty term is minimized when $v = 0$. Therefore, any minimizer of (2.3) must have the form of (2.4).

Proof of Proposition 1. By the reproducing property,

$$\|\widetilde{K}_{\mathbf{u}}\|_{\widetilde{\mathcal{H}}}^2 = \langle \widetilde{K}_{\mathbf{u}}, \widetilde{K}_{\mathbf{u}} \rangle_{\widetilde{\mathcal{H}}} = \widetilde{K}(\mathbf{u}, \mathbf{u}) = \sum_{kk'} \frac{\varphi_{kk'}^2(\mathbf{u})}{(1 + \lambda/\tau_{kk'})} \leq C_{\varphi}^2 d(\lambda). \tag{S6.4}$$

Hence, for all $\mu \in \mathcal{H}$,

$$|\mu(\mathbf{u})| = |\langle \mu, \widetilde{K}_{\mathbf{u}} \rangle_{\widetilde{\mathcal{H}}}| \leq \|\mu\|_{\widetilde{\mathcal{H}}} \|\widetilde{K}_{\mathbf{u}}\|_{\widetilde{\mathcal{H}}} \leq C_{\varphi} d(\lambda)^{1/2} \|\mu\|_{\widetilde{\mathcal{H}}} \tag{S6.5}$$

where the first inequality holds by Cauchy-Schwarz inequality. Taking suprema on both sides yields $\|\mu\|_{\sup} \leq C_{\varphi} d(\lambda)^{1/2} \|\mu\|_{\widetilde{\mathcal{H}}}$.

Proof of Lemma 2. Denote $\tilde{\mu} = \hat{\mu}_{n,m,\lambda} - \mu_0$. By Taylor's expansion, we have

$$\ell_{n,m,\lambda}(\mu_0 + \tilde{\mu}) - \ell_{n,m,\lambda}(\mu_0) = S_{n,m,\lambda}(\mu_0)\tilde{\mu} + \frac{1}{2} DS_{n,m,\lambda}(\mu_0)\tilde{\mu}\tilde{\mu} \leq 0, \tag{S6.6}$$

We will study the rates of the above two terms respectively.

Recall that $S_{n,m,\lambda}(\mu_0) = -\frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m \widetilde{K}_{\mathbf{u_{ij}}} \epsilon_i(s_j) + W_\lambda \mu_0$, direct calculations lead to

$$\begin{aligned} \|W_\lambda \mu_0\|_{\widetilde{\mathcal{H}}} &= \sup_{\|\mu\|_{\widetilde{\mathcal{H}}}=1} |\langle W_\lambda \mu_0, \mu \rangle_{\widetilde{\mathcal{H}}}| = \sup_{\|\mu\|_{\widetilde{\mathcal{H}}}=1} \lambda |\langle \mu_0, \mu \rangle_{\mathcal{H}}| \leq \sup_{\|\mu\|_{\widetilde{\mathcal{H}}}=1} \sqrt{\lambda} \|\mu_0\|_{\mathcal{H}} \sqrt{\lambda} \|\mu\|_{\mathcal{H}} \\ &\leq \sqrt{\lambda} \|\mu_0\|_{\mathcal{H}} = O(\sqrt{\lambda}). \end{aligned} \tag{S6.7}$$

Meanwhile, by the definition of $\widetilde{K}_{\mathbf{u_{ij}}}$ in Proposition S1 and Assumption 2, it is easy to find that

$$\begin{aligned} & E\big(\|\sum_{i=1}^{n}\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j)\|_{\widetilde{\mathcal{H}}}^2\big) \\ = \;& nE\big(\|\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j)\|_{\widetilde{\mathcal{H}}}^2\big)+nm(m-1)E\big[\langle\widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j),\widetilde{K}_{\mathbf{u_{ij'}}}\epsilon_i(s_{j'})\rangle\big] \\ = \;& O_p(nmd(\lambda))+O_p(nm(m-1)). \end{aligned} \tag{S6.8}$$

This follows from the fact that $E\big(\|\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j)\|_{\widetilde{\mathcal{H}}}^2\big)<nmd(\lambda)$, and for the second term in (S6.8), one can verify by taking conditional expectation,

$$\begin{aligned} E\big[\langle\widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j),\widetilde{K}_{\mathbf{u_{ij'}}}\epsilon_i(s_{j'})\rangle\big] &= \sum_{k,k'}\frac{E[\varphi_{kk'}(\mathbf{u}_{ij})\varphi_{kk'}(\mathbf{u}_{ij'})\epsilon_i(s_j)\epsilon_i(s_{j'})]}{(1+\lambda/\tau_{kk'})^2} \\ \leq \;& \sum_{k,k'}E\Big\{\int\varphi_{kk'}(\mathbf{u})\epsilon_i(s)ds\int\varphi_{kk'}(\mathbf{u}')\epsilon_i(s')ds'\Big\}<\infty, \end{aligned}$$

where the last inequality follows from Assumption 2 and the condition $\inf_s P(s)\geq c_0>0$. Hence, combining (S6.7) and (S6.8), we can have

$$\|S_{n,m,\lambda}(\mu_0)\|_{\widetilde{\mathcal{H}}}^2=O_p\Big(\frac{d(\lambda)}{nm}+\frac{1}{n}+\lambda\Big). \tag{S6.9}$$

Let $\psi_{n,m,\lambda}(T_{ij},f)=d(\lambda)^{-1/2}\widetilde{K}_{\mathbf{u_{ij}}}f=C_{\varphi}^{-1}d(\lambda)^{-1/2}f(\mathbf{u_{ij}})$, then we have

$$|\psi_{n,m}(T;f_1)-\psi_{n,m}(T;f_2)|\leq C^{-1}d(\lambda)^{-1/2}\|f_1-f_2\|_{\sup}\quad\text{for any } f_1,f_2\in\mathcal{Q}\ .$$

Denote $f^*=d(\lambda)^{-1/2}f$, then it is easy to verify that $f^*\in\mathcal{Q}$. According to Lemma 1, it is easy to find that

$$\begin{aligned} & \big|[DS_{n,m,\lambda}(\mu_0)-E\{DS_{n,m,\lambda}(\mu_0)\}]\tilde{\mu}^*\tilde{\mu}^*\big| \\ \leq \;& (nm)^{-1/2}d(\lambda)^{1/2}\sqrt{\log\log\big(nmJ(\mathcal{F},1)\big)}(J(\mathcal{F},1)+(nm)^{-1/2}))\|\tilde{\mu}^*\|_{\widetilde{\mathcal{H}}}. \end{aligned}$$

Meanwhile, for any $f_1, f_2 \in \mathcal{Q}$, by Proposition S1, we have

$$\langle E\{DS_{n,m,\lambda}(\mu_0)\}f_1, f_2\rangle_{\widetilde{\mathcal{H}}} = \langle f_1, f_2\rangle_{\widetilde{\mathcal{H}}}.$$

Hence, we can directly find that

$$\begin{aligned} & DS_{n,m,\lambda}(\mu_0)\tilde{\mu}\tilde{\mu} \\ \leq\ & \|\tilde{\mu}\|_{\widetilde{\mathcal{H}}}^2 + (nm)^{-1/2}d(\lambda)\sqrt{\log\log\big(nmJ(\mathcal{F},1)\big)}(J(\mathcal{F},1) + (nm)^{-1/2}))\|\tilde{\mu}\|_{\widetilde{\mathcal{H}}} \\ =\ & \|\tilde{\mu}\|_{\widetilde{\mathcal{H}}}^2 + o_p(1)\|\tilde{\mu}\|_{\widetilde{\mathcal{H}}} \qquad \text{(S6.10)} \end{aligned}$$

where the last inequality follows from the condition that $\sqrt{\log\log\big(nmJ(\mathcal{F},1)\big)}J(\mathcal{F},1) = o_p((nm)^{1/2}d(\lambda)^{-1})$.

Plugging (S6.7), (S6.9) and (S6.10) into (S6.10), it is easy to derive that

$$(1 + o_p(1))\|\tilde{\mu}\|_{\widetilde{\mathcal{H}}} \leq \sqrt{d(\lambda)/(nm)} + n^{-1/2} + \lambda^{1/2},$$

leading to $\|\tilde{\mu}\|_{\widetilde{\mathcal{H}}} = \|\widehat{\mu} - \mu_0\|_{\widetilde{\mathcal{H}}} = O_p(\sqrt{d(\lambda)/(nm)} + n^{-1/2} + \lambda^{1/2})$.

Proof of Theorem 3. By Taylor's expansion and $S_{n,m,\lambda}(\hat{\mu})$ and Proposition S1, we have

$$\begin{aligned} & \|S_{n,m,\lambda}(\tilde{\mu} + \mu_0) - S_\lambda(\tilde{\mu} + \mu_0) - \big(S_{n,m,\lambda}(\mu_0) - S_\lambda(\mu_0)\big)\|_{\widetilde{\mathcal{H}}} \\ =\ & \|S_\lambda(\tilde{\mu} + \mu_0) + S_{n,m,\lambda}(\mu_0) - S_\lambda(\mu_0)\|_{\widetilde{\mathcal{H}}} = \|\tilde{\mu} + S_{n,m,\lambda}(\mu_0)\|_{\widetilde{\mathcal{H}}}, \end{aligned}$$

where $S_\lambda(\mu) = E(S_{n,m,\lambda}(\mu))$. It can be easily verify that

$$\begin{aligned}
& \|S_{n,m,\lambda}(\tilde\mu + \mu_0) - S_\lambda(\tilde\mu + \mu_0) - \big(S_{n,m,\lambda}(\mu_0) - S_\lambda(\mu_0)\big)\|_{\widetilde{\mathcal{H}}} \\
= \; & \|S_{n,m}(\tilde\mu + \mu_0) - S(\tilde\mu + \mu_0) - \big(S_{n,m}(\mu_0) - S(\mu_0)\big)\|_{\widetilde{\mathcal{H}}} \\
= \; & \|DS_{n,m}(\mu_0)\tilde\mu - E(DS_{n,m}(\mu_0))\tilde\mu\|_{\widetilde{\mathcal{H}}} \\
= \; & \Big\|\frac{1}{nm}\sum_{i=1}\sum_{j=1}[\psi(T_{ij};\tilde\mu)\widetilde{K}_{\mathbf{u_{ij}}} - E\{\psi(T_{ij};\tilde\mu)\widetilde{K}_{\mathbf{u_{ij}}}\}]\Big\|_{\widetilde{\mathcal{H}}},
\end{aligned}$$

where $\psi(T_{ij};\tilde\mu) = C_\varphi^{-1}d(\lambda)^{-1/2}\tilde\mu(\mathbf{u_{ij}})$. Denote $r_n^2 = d(\lambda)/(nm) + n^{-1} + \lambda$ and $\tilde\mu^* = r_n^{-1}d(\lambda)^{-1/2}\tilde\mu$, then $\|\tilde\mu^*\|_{\sup} \le 1$ and $\|\tilde\mu^*\|_{\mathcal{H}}^2 \le = r_n^{-2}d(\lambda)^{-1}\lambda^{-1}\lambda J(\tilde\mu,\tilde\mu) \le Cr_n^{-2}d(\lambda)^{-1}\lambda^{-1}r_n^2 = Cd(\lambda)^{-1}\lambda^{-1}$ by observing that $\lambda J(\tilde\mu,\tilde\mu) \le \|\tilde\mu\|_{\widetilde{\mathcal{H}}} = O_p(r_n)$. Hence $\tilde\mu^* \in \mathcal{Q}$ and $|\psi(T_{ij};\tilde\mu_1^*) - \psi(T_{ij};\tilde\mu_2^*)| \le C_\varphi^{-1}d(\lambda)^{-1/2}\|\tilde\mu_1^* - \tilde\mu_2^*\|_{\sup}$. By Lemma 1 and using an argument similar to the proof of S6.10, it is easy to derive that there exist constants $C, C' > 0$ such that

$$\begin{aligned}
& \|S_{n,m,\lambda}(\tilde\mu + \mu_0) - S_\lambda(\tilde\mu + \mu_0) - \big(S_{n,m,\lambda}(\mu_0) - S_\lambda(\mu_0)\big)\|_{\widetilde{\mathcal{H}}} \\
\le \; & C(nm)^{-1/2}d(\lambda)\sqrt{\log\log\big(nmJ(\mathcal{Q},1)\big)}(J(\mathcal{Q},1) + (nm)^{-1/2}))r_n \\
\le \; & C'(nm)^{-1/2}d(\lambda)\sqrt{\log\log\big(nmJ(\mathcal{Q},1)\big)}J(\mathcal{Q},1)r_n.
\end{aligned}$$

This completes the proof.

Proof of Theorem 4. Recall that $S_{n,m,\lambda}(\mu_0) = -\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j) + W_\lambda\mu_0$, and $(nm)^{1/2}(\hat\mu(\mathbf{u}_0) - \mu_0^*(\mathbf{u}_0)) = (nm)^{1/2}\langle\widetilde{K}_{\mathbf{u}_0}, \hat\mu - \mu_0^*\rangle_{\widetilde{\mathcal{H}}}$, where $\mu_0^* = \mu_0 - W_\lambda\mu_0$. By Theorem 3 and the condition $a_n^2 d(\lambda) = o_p((nm)^{-1}(d_2(\lambda) +$

$m)$), we have

$$\begin{aligned} & (nm)^{1/2}|\langle \widetilde{K}_{\mathbf{u}_0}, \hat{\mu} - \mu_0^* - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j)\rangle_{\widetilde{\mathcal{H}}}| \\ \leq \ & (nm)^{1/2}\|\widetilde{K}_{\mathbf{u}_0}\|_{\widetilde{\mathcal{H}}}\|\widehat{\mu} - \mu_0 + S_{n,m,\lambda}(\mu_0)\|_{\widetilde{\mathcal{H}}} = o_p([d_2(\lambda) + m]^{1/2}). \end{aligned}$$

Note that $\sigma_{\mathbf{u}_0}^2 = O_p(d_2(\lambda))$ and $r_{\mathbf{u}_0}^2 = O_p(1)$ according to Assumption 2, we only need to find the limiting distribution of

$$\sqrt{\frac{nm}{\sigma_{\mathbf{u}_0}^2 + mr_{\mathbf{u}_0}^2}}\langle \widetilde{K}_{\mathbf{u}_0}, \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}\epsilon_i(s_j)\rangle_{\widetilde{\mathcal{H}}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}(\mathbf{u}_0)\epsilon_i(s_j)/[(nm)\sigma_{\mathbf{u}_0}^2 + (nm^2)r_{\mathbf{u}_0}^2]^{1/2}.$$

Direct calculations lead to

$$\begin{aligned} & Var(\sum_{i=1}^{n}\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}(\mathbf{u}_0)\epsilon_i(s_j)) \\ = \ & (nm)\sigma_\epsilon^2 E(\widetilde{K}_{\mathbf{u_{ij}}}(\mathbf{u}_0)^2) + (nm(m-1))E[\widetilde{K}_{\mathbf{u_{ij}}}(\mathbf{u}_0)\epsilon_i(s_j)\widetilde{K}_{\mathbf{u_{ij'}}}(\mathbf{u}_0)\epsilon_i(s_{j'})] \\ = \ & (nm)\sigma_{\mathbf{u}_0}^2 + (nm(m-1))r_{\mathbf{u}_0}^2 \asymp (nm)\sigma_{\mathbf{u}_0}^2 + (nm^2)r_{\mathbf{u}_0}^2. \end{aligned}$$

Hence, according to the central limit theorem, we can obtain that

$$\sum_{i=1}^{n}\sum_{j=1}^{m}\widetilde{K}_{\mathbf{u_{ij}}}(\mathbf{u}_0)\epsilon_i(s_j)/[(nm)\sigma_{\mathbf{u}_0}^2 + (nm^2)r_{\mathbf{u}_0}^2]^{1/2} \xrightarrow{d} N(0,1).$$

Next we show that the bias converges to zero. It is easy to verify that

$$\sqrt{\frac{nm}{\sigma_{\mathbf{u}_0}^2 + mr_{\mathbf{u}_0}^2}}W_\lambda\mu_0(\mathbf{u}_0) = \sqrt{\frac{nm\lambda}{\sigma_{\mathbf{u}_0}^2 + mr_{\mathbf{u}_0}^2}}\sum_{k=1}^{\infty}\sum_{k'=1}^{\infty}\mu_{kk'}/(\tau_{kk'})^{1/2}\frac{(\lambda/\tau_{kk'})^{1/2}}{\lambda/\tau_{kk'}+1}\varphi_{kk'}(\mathbf{u}_0) \to 0.$$

By the condition $\frac{nm\lambda}{\sigma_{\mathbf{u}_0}^2 + mr_{\mathbf{u}_0}^2} \asymp (nm\lambda)/(d_2(\lambda)+m) = O_p(1)$, $\sum_{k=1}^{\infty}\sum_{k'=1}^{\infty}\mu_{kk'}/(\tau_{kk'})^{1/2} < \infty$, $\sup_{x\geq 0}\frac{x}{1+x^2} < \infty$ and the dominated convergence theorem, as $n, m \to 0$, we have

$$\sqrt{\frac{nm}{\sigma^2_{\mathbf{u}_0} + mr^2_{\mathbf{u}_0}}} W_\lambda \mu_0(\mathbf{u}_0) \to 0.$$

This completes the proof.

Proof of Theorem 5. (i) It is easy to derive that $S_\nu(\phi^0, \mathbf{\Sigma}_\epsilon, \boldsymbol{\rho}_\nu) \xrightarrow{d} \sum_\ell \lambda_\ell x_\ell^2$ as $n \to \infty$.

(ii) Under the alternative hypothesis such that for any sequence $c_n \to \infty$, $\tau_\nu \geq c_n \sum_\ell \lambda_\ell / \sum_\ell \lambda_\ell^2$,

$$S_\nu(\phi^0, \mathbf{\Sigma}, \boldsymbol{\rho}_\nu) = \frac{1}{2}\mathbf{Y}^T \tilde{V}^{-1}\mathbf{K}_\nu \tilde{V}^{-1}\mathbf{Y} + \frac{1}{2}\mathbf{Y}^T(V^{-1}\mathbf{K}_\nu V^{-1} - \tilde{V}^{-1}\mathbf{K}_\nu \tilde{V}^{-1})\mathbf{Y} \tag{S6.11}$$

where $\tilde{V} = V + \tau_\nu \mathbf{K}_\nu$.We can derive that

$$\frac{1}{2}\mathbf{Y}^T \tilde{V}^{-1}\mathbf{K}_\nu \tilde{V}^{-1}\mathbf{Y} \xrightarrow{d} \sum_\ell \tilde{\lambda}_\ell x_\ell^2$$

with $\{\tilde{\lambda}_\ell\}$ being eigenvalues of $\tilde{V}^{-1}\mathbf{K}_\nu/2$ and $x_\ell \sim N(0,1)$. The second term in (S6.11) can be rewritten as

$$\frac{1}{2}\mathbf{Y}^T \tilde{V}^{-1/2}(\tilde{V}^{1/2}V^{-1}\mathbf{K}_\nu V^{-1}\tilde{V}^{-1/2} - \tilde{V}^{-/2}\mathbf{K}_\nu \tilde{V}^{-1/2})\tilde{V}^{-1/2}\mathbf{Y} \xrightarrow{d} \sum_\ell \lambda^*_\ell x_\ell^2,$$

where $\{\lambda^*_\ell\}$ are eigenvalues of $(\tilde{V}V^{-1}\mathbf{K}_\nu V^{-1} - \tilde{V}^{-1}\mathbf{K}_\nu)/2$. Notice that $\tilde{V} = V + \tau_\nu \mathbf{K}_\nu$, we can derive that

$$\tilde{V}V^{-1}\mathbf{K}_\nu V^{-1} - \mathbf{K}_\nu \tilde{V}^{-1} = \mathbf{K}_\nu V^{-1} + \tau_\nu \mathbf{K}_\nu V^{-1}\mathbf{K}_\nu V^{-1} - \mathbf{K}_\nu \tilde{V}^{-1},$$

leading to

$$tr(\tilde{V}V^{-1}\mathbf{K}_\nu V^{-1} - \mathbf{K}_\nu \tilde{V}^{-1}) = tr(\mathbf{K}_\nu V^{-1}(1 + \tau_\nu \mathbf{K}_\nu V^{-1})) - tr(\mathbf{K}_\nu \tilde{V}^{-1})).$$

Hence, under the alternative,

$$S_\nu(\phi^0, \boldsymbol{\Sigma}, \boldsymbol{\rho}_\nu) \xrightarrow{d} \sum_\ell (\lambda_\ell + \tau_\nu \lambda_\ell^2) x_\ell^2.$$

It then can be obtained that when choose sufficient large $M > 0$, $\tau_\nu \geq M \sum_\ell \lambda_\ell / \sum_\ell \lambda_\ell^2$, $S_\nu(\phi^0, \boldsymbol{\Sigma}, \boldsymbol{\rho}_\nu)$ can be larger than the cutoff value the power approaches to one.

(iii) We now show that $S_\nu(\hat{\phi}, \widehat{\boldsymbol{\Sigma}}_\epsilon, \boldsymbol{\rho}_\nu) \xrightarrow{d} S_\nu(\phi^0, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\rho}_\nu)$. Recall that $V = \tau \mathbf{K} + \boldsymbol{\Sigma}_\epsilon = \sum_{\nu=1}^{p+1} \tau_\nu \mathbf{K}_\nu + \boldsymbol{\Sigma}_\epsilon$ with $\tau_\nu = 0$ under the null. Direct calculations lead to

$$S_\nu(\hat{\phi}, \widehat{\boldsymbol{\Sigma}}_\epsilon, \boldsymbol{\rho}_\nu) - S_\nu(\phi^0, \boldsymbol{\Sigma}, \boldsymbol{\rho}_\nu) = \frac{1}{2} \mathbf{Y}^T \big(\hat{V}^{-1} \mathbf{K}_\nu \hat{V}^{-1} - V^{-1} \mathbf{K}_\nu V^{-1} \big) \mathbf{Y}.$$

Denote $J_{nm} = ntr(K_s)$, it is easy to see that $tr(V^{-1}K_\nu/2) = O_p(ntr(K_s))$. It is enough to show that $\|\hat{V}^{-1}\mathbf{K}_\nu \hat{V}^{-1}/J_{nm} - V^{-1}\mathbf{K}_\nu V^{-1}/J_{nm}\|_s = o_p(1)$ to obtain $S_\nu(\hat{\phi}, \widehat{\boldsymbol{\Sigma}}_\epsilon, \boldsymbol{\rho}_\nu) \xrightarrow{d} S_\nu(\phi^0, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\rho}_\nu)$ by using $S_\nu(\phi^0, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\rho}_\nu) = O_p(J_{nm})$.

Let $\mathbf{K} = \mathbf{U}\mathbf{U}^\top$, $\mathbf{K}_\nu = \mathbf{U}_\nu \mathbf{U}_\nu^\top$, $A = V^{-1}\mathbf{U}_\nu/\sqrt{J_{nm}}$ and $\hat{A} = \hat{V}^{-1}\mathbf{U}_\nu/\sqrt{J_{nm}}$, then

$$\begin{aligned} & \| \frac{\hat{V}^{-1}\mathbf{K}_\nu \hat{V}^{-1}}{J_{nm}} - \frac{V^{-1}\mathbf{K}_\nu V^{-1}}{J_{nm}} \|_s = \| \frac{\hat{V}^{-1}\mathbf{U}_\nu \mathbf{U}_\nu^\top \hat{V}^{-1}}{J_{nm}} - \frac{V^{-1}\mathbf{U}_\nu \mathbf{U}_\nu^\top V^{-1}}{J_{nm}} \|_s \\ = \; & \|\hat{A}\hat{A}^\top - AA^\top\|_s \leq \|\hat{A}\|_s \|\hat{A}^\top - A^\top\|_s + \|A\|_s \|\hat{A}^\top - A^\top\|_s. \end{aligned}$$

We next show that $\|\hat{A}\|_s$ and $\|A\|_s$ are bounded and $\|\hat{A}^\top - A^\top\|_s = o_p(1)$, which implies that $\|\hat{V}^{-1}\mathbf{K}_\nu \hat{V}^{-1}/J_{nm} - V^{-1}\mathbf{K}_\nu V^{-1}/J_{nm}\|_s = o_p(1)$.

For $\nu = 1, \ldots, p$, there exists a universal constant $C > 0$ such that $tr(\mathbf{K}_\nu / J_{nm}) = tr(\mathbf{Z}_\nu \mathbf{Z}_\nu^\top / n) tr(\mathbf{K}_s)/tr(\mathbf{K}_s) < C$ because $E_Z(\mathbf{Z}\mathbf{Z}^\top)$ is positive definite. Similarly, for $\nu = p+1$, $tr(\mathbf{K}_\nu / J_{nm}) = tr(\mathbf{K}_x / n) tr(\mathbf{K}_s)/tr(\mathbf{K}_s) < C$ for some constant C. By simple calculations using Woodbury matrix identity, we have

$$\hat{A}^\top = \mathbf{U}_\nu^\top \hat{V}^{-1} = \frac{\mathbf{U}_\nu^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}}{\sqrt{J_{nm}}} - \hat{\tau}\Big(\frac{\mathbf{U}_\nu^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1} \mathbf{U}}{J_{nm}}\Big)\Big(\frac{I + \hat{\tau}\mathbf{U}^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\Big)^{-1}\Big(\frac{\mathbf{U}^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}}{\sqrt{J_{nm}}}\Big) \quad (S6.12)$$

Notice that $\|\mathbf{U}_\nu^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}/\sqrt{J_{nm}}\|_F \leq \|\widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}\|_F \sqrt{tr(\mathbf{U}_\nu^\top \mathbf{U}_\nu / J_{nm})} = \|\widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}\|_F \sqrt{tr(\mathbf{K}_\nu / J_{nm})} < C$ for some constant $C > 0$, where $\|\cdot\|_F$ denotes the Frobenius norm. It follows that $\|\hat{A}\|_s \leq \|\mathbf{U}_\nu^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}/\sqrt{J_{nm}}\|_s \leq \|\mathbf{U}_\nu^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}/\sqrt{J_{nm}}\|_F \leq C$. We can conclude that $\|\hat{A}\|_s$ is bounded. Similarly, we can have $\|A\|_s$ is bounded.

To show $\|\hat{A}^\top - A^\top\|_s = o_p(1)$, we next show that each term converges to the counterpart with τ_0 and Σ in place of $\hat{\tau}$ and $\widehat{\boldsymbol{\Sigma}}_\epsilon$. Specifically, according to the condition $\|\widehat{\boldsymbol{\Sigma}}_\epsilon^{-1} - \Sigma^{-1}\|_s = o_p(1)$, one can directly have

$$\begin{aligned}
\Big\|\frac{\mathbf{U}_\nu^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}}{\sqrt{J_{nm}}} - \frac{\mathbf{U}_\nu^\top \boldsymbol{\Sigma}_\epsilon^{-1}}{\sqrt{J_{nm}}}\Big\|_s &\leq \sqrt{tr(\mathbf{K}_\nu / J_{nm})}\|\widehat{\boldsymbol{\Sigma}}_\epsilon^{-1} - \boldsymbol{\Sigma}_\epsilon^{-1}\|_s = o_p(1), \\
\Big\|\frac{\mathbf{U}_\nu^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}\mathbf{U}}{J_{nm}} - \frac{\mathbf{U}_\nu^\top \boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\Big\|_s &\leq \sqrt{tr(\mathbf{K}_\nu / J_{nm})}\sqrt{tr(\mathbf{K}/J_{nm})}\|\widehat{\boldsymbol{\Sigma}}_\epsilon^{-1} - \boldsymbol{\Sigma}_\epsilon^{-1}\|_s = o_p(1).
\end{aligned}$$

Meanwhile, it can be seen that

$$\begin{aligned}
&\Big\|\Big(\frac{I + \hat{\tau}\mathbf{U}^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\Big)^{-1} - \Big(\frac{I + \tau_0 \mathbf{U}^\top \boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\Big)^{-1}\Big\|_s \\
\leq\ & \Big\|\Big(\frac{I + \hat{\tau}\mathbf{U}^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\Big)^{-1}\Big\|_s \Big\|\frac{\hat{\tau}\mathbf{U}^\top \widehat{\boldsymbol{\Sigma}}_\epsilon^{-1}\mathbf{U} - \tau_0 \mathbf{U}^\top \boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\Big\|_s \Big\|\Big(\frac{I + \tau_0 \mathbf{U}^\top \boldsymbol{\Sigma}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\Big)^{-1}\Big\|_s.
\end{aligned}$$

By observing that the eigenvalues of $(I+\hat{\tau}\mathbf{U}^\top\widehat{\mathbf{\Sigma}}_\epsilon^{-1}\mathbf{U})/J_{nm}$ and $(I+\tau_0\mathbf{U}^\top\mathbf{\Sigma}_\epsilon^{-1}\mathbf{U})/J_{nm}$ are positive, we have $\|((I+\hat{\tau}\mathbf{U}^\top\widehat{\mathbf{\Sigma}}_\epsilon^{-1}\mathbf{U})/J_{nm})^{-1}\|_s$ and $\|((I+\tau_0\mathbf{U}^\top\mathbf{\Sigma}_\epsilon^{-1}\mathbf{U})/J_{nm})^{-1}\|_s$ are bounded. Furthermore,

$$\begin{aligned} & \left\|\frac{\hat{\tau}\mathbf{U}^\top\widehat{\mathbf{\Sigma}}_\epsilon^{-1}\mathbf{U} - \tau_0\mathbf{U}^\top\mathbf{\Sigma}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\right\|_s \leq |\hat{\tau}-\tau_0|\left\|\frac{\mathbf{U}^\top\widehat{\mathbf{\Sigma}}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\right\|_s + |\tau_0|\left\|\frac{\mathbf{U}^\top\widehat{\mathbf{\Sigma}}_\epsilon^{-1}\mathbf{U} - \mathbf{U}^\top\mathbf{\Sigma}_\epsilon^{-1}\mathbf{U}}{J_{nm}}\right\|_s \\ \leq\ & |\hat{\tau}-\tau_0|\|\widehat{\mathbf{\Sigma}}_\epsilon^{-1}\|_s tr(\mathbf{K}/J_{nm}) + |\tau_0|\|\widehat{\mathbf{\Sigma}}_\epsilon^{-1} - \mathbf{\Sigma}_\epsilon^{-1}\|_s tr(\mathbf{K}/J_{nm}) = o_p(1). \end{aligned}$$

Hence, we obtain $\|\hat{A}^\top - A^\top\|_s = o_p(1)$, then $\|\hat{V}^{-1}\mathbf{K}_\nu\hat{V}^{-1}/J_{nm} - V^{-1}\mathbf{K}_\nu V^{-1}/J_{nm}\|_s = o_p(1)$. This completes the proof.

Bibliography

Chaudhuri, P. and J. S. Marron (2000). Scale space view of curve estimation. *The Annals of Statistics 28*(2), 408–428.

Gu, C. and G. Wahba (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *Journal on Scientific and Statistical Computing 12*(2), 383–398.

Härdle, W. (1990). *Applied nonparametric regression.* Cambridge university press.

Kosorok, M. R. (2007). *Introduction to empirical processes and semiparametric inference.* Springer Science & Business Media.

Okbay, A., Y. Wu, N. Wang, H. Jayashankar, M. Bennett, S. M. Nehzati, J. Sidorenko, H. Kweon, G. Goldman, and T. Gjorgjieva (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics 54*(4), 437–449.

Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability 22*(4), 1679–1706.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, Volume 3. Cambridge university press.

Wang, W., Z. Xu, W. Lu, and X. Zhang (2003). Determination of the spread parameter in the Gaussian kernel for classification and regression. *Neurocomputing 55*(3), 643–663.