SUPPLEMENTARY MATERIAL

A Variable Selection

Rank	fFASM	
	Name	Frequency
1	Employment (million people)	160
2	Adult literacy rate (age 15 and above) (%)	56
3	Total fertility rate	54
4	Unemployment rate (% of total labor force)	21
5	Gross national income per capita (current USD)	17
Rank	Lasso	
	Name	Frequency
1	Employment (million people)	166
2	Total fertility rate	57
3	Adult literacy rate (age 15 and above, %)	37
4	Labor force (people)	16
5	Unemployment rate (% of total labor force)	7
Rank	grLasso	
	Name	Frequency
1	Employment (million people)	187
2	Prevalence of undernourishment (% of total population)	7
3	Adult literacy rate (age 15 and above) (%)	3
4	Gross national income per capita (current USD)	2
5	Per capita health expenditure (current USD)	1

Table 1: Top 5 variables with highest selection frequency.

Rank	fFASM	
	Name	Frequency
1	Cement	198
2	Medium Tractors	13
3	Engines	12
3	Aluminum Materials	12
5	Chemical Pesticides (Active Ingredients)	4
Rank	Lasso	
	Name	Frequency
1	Cement	199
2	Medium Tractors	48
3	Aluminum Materials	19
4	Computers	18
4	Engines	18
Rank	grLasso	
	Name	Frequency
1	Cement	137
2	Aluminum Materials	97
3	Mechanized Paper	68
4	Hydropower Generation	30
5	Phosphate Rock	20

Table 2: Top 5 variables with highest selection frequency.

B Regularity Condition

In this section, we introduce some regularity conditions (Kong et al. 2016). Without loss of generality, we assume that $\{X^{(g)}(\cdot), g = 1, \ldots, G\}$ have been centred to have zero mean. With $W_i^{(g)}(t_{ij}) = X_i^{(g)}(t_{ij}) + \epsilon_i^{(g)}(t_{ij})$, for definiteness, we consider the local linear smoother for each set of subjects using bandwidths $\{h_i^{(g)}, g = 1, \ldots, G\}$, and denote the smoothed trajectories by $\widehat{X}_i^{(g)}(\cdot)$.

Condition (B1) consists of regularity assumptions for functional data. Condition (B2) is standard for local linear smoother, (B3)–(B4) concern how the functional predictors are sampled and smoothed.

(B1) For $g = 1, \ldots, G$, for any $C > 0$ there exists an $\epsilon > 0$ such that

$$\sup_{t\in\mathcal{T}}\left[\mathbb{E}\left\{\left|X^{(g)}(t)\right|^C\right\}\right] < \infty, \quad \sup_{s,t\in\mathcal{T}}\left(\mathbb{E}\left[\left\{|s-t|^{-\epsilon}\left|X^{(g)}(s) - X^{(g)}(t)\right|\right\}^C\right]\right) < \infty.$$

For each integer $r \geq 1$, $\left(\tau_k^{(g)}\right)^{-r} E\left(a_k^{(g)}\right)^{2r}$ is bounded uniformly in k.

(B2) For $g = 1, \ldots, G$, $X^{(g)}(\cdot)$ is twice continuously differentiable on $\mathcal{T}$ with probability 1 , and $\int \mathbb{E}\left\{X^{(g)\prime\prime}(t)\right\}^4 dt < \infty$, where $X^{(g)\prime\prime}(\cdot)$ denotes the second derivative of $X^{(g)}(\cdot)$. The following condition concerns the design on which $X_i^{(g)}(\cdot)$ is observed and the local linear smoother $\widehat{X}_i^{(g)}(\cdot)$. When a function is said to be smooth, we mean that it is continuously differentiable to an adequate order.

(B3) For $g = 1, \ldots, G$, $\left\{t_{ij}^{(g)}, j = 1, \ldots, n_i^{(g)}\right\}$ are considered deterministic and ordered increasingly for $i = 1, \ldots, n$. There exist densities $p_i^{(g)}$ uniformly smooth over $\mathcal{T}$, satisfying $\int_{\mathcal{T}} p_i^{(g)}(t)dt = 1$ and $0 < c_1 < \inf_i\left\{\inf_{t\in T} p_i^{(g)}(t)\right\} < \sup_i\left\{\sup_{t\in T} p_i^{(g)}(t)\right\} < c_2 < \infty$ that generate $t_{ij}^{(g)}$ according to $t_{ij}^{(g)} = P_i^{(g)-1}\{j/\left(n_i^{(g)} + 1\right)\}$, where $P_i^{(g)-1}$ is the inverse of $P_i^{(g)}(t) = \int_{-\infty}^t p_i^{(g)}(s)ds$. For each $g = 1, \ldots, G$, there exist a common sequence of bandwidths $h^{(g)}$ such that $0 < c_1 < \inf_i h_i^{(g)}/h^{(g)} < \sup_i h_i^{(g)}/h^{(g)} < c_2 < \infty$,

where $h_i^{(g)}$ is the bandwidth for $\widehat{X}_i^{(g)}$. The kernel density function is smooth and compactly supported.

Let $\mathcal{T} = [a_0, b_0], t_{i0}^{(g)} = a_0, t_{i,n_i^{(g)}+1}^{(g)} = b_0$, let $\Delta_i^{(g)} = \sup\left\{t_{i,j+1}^{(g)} - t_{i,j}^{(g)}, j = 0, \ldots, n_i^{(g)}\right\}$ and $n^{(g)} = n^{(g)}(n) = \inf_{i=1,\ldots,n} n_i^{(g)}$. The condition below is to let the smooth estimate $\widehat{X}_i^{(g)}$ serve as well as the true $X_i^{(g)}$ in the asymptotic analysis, denoting $0 < \lim a_n/b_n < \infty$ by $a_n \sim b_n$.

(B4) For $g = 1, \ldots, G$, $\sup_i \Delta_i^{(g)} = O\left(\left(n^{(g)}\right)^{-1}\right)$, $h^{(g)} \sim \left(n^{(g)}\right)^{-1/5}, n^{(g)} n^{-5/4} \to \infty$.

C Proof of Lemma

C.1 Proof of Lemma 1

Proof: To prove Lemma 1, we need to verify that under the given conditions, the pre-specified basis functions $\phi_{0j}^{(g)}(\cdot)$ are indeed the eigenfunctions of the covariance operator of $X_i^{(g)}(\cdot)$, and they correspond to the ordered eigenvalues. Consequently, the projection of $X_i^{(g)}(\cdot)$ onto these functions recovers the coefficients $a_{0,ij}^{(g)}$.

Let $K^{(g)}(s,t)$ denote the covariance kernel of the process $X_i^{(g)}(\cdot)$. Since the process is centered (i.e., $\mathbb{E}[X_i^{(g)}(t)] = 0$), the covariance is defined as:

$$K^{(g)}(s,t) = \mathrm{Cov}(X_i^{(g)}(s), X_i^{(g)}(t)) = \mathbb{E}\left[X_i^{(g)}(s) X_i^{(g)}(t)\right]. \tag{C.1}$$

Substituting the expansion $X_i^{(g)}(t) = \sum_{j=1}^{m_g} a_{0,ij}^{(g)} \phi_{0j}^{(g)}(t) + e_{0i}^{(g)}(t)$ into the expectation:

$$K^{(g)}(s,t) = \mathbb{E}\left[\left(\sum_{j=1}^{m_g} a_{0,ij}^{(g)} \phi_{0j}^{(g)}(s) + e_{0i}^{(g)}(s)\right)\left(\sum_{k=1}^{m_g} a_{0,ik}^{(g)} \phi_{0k}^{(g)}(t) + e_{0i}^{(g)}(t)\right)\right].$$

We apply the assumptions given in the Lemma:

- $\mathbb{E}(a_{0,ij}^{(g)} e_{0i}^{(g)}(t)) = 0$ implies cross-terms between $\boldsymbol{a}$ and $\boldsymbol{e}$ vanish.

- $\mathbb{E}(a_{0,ij}^{(g)} a_{0,ik}^{(g)}) = 0$ for $j \neq k$ implies cross-terms between distinct coefficients vanish.
- Let $\lambda_j^{(g)} = \mathrm{Var}(a_{0,ij}^{(g)}) = \mathbb{E}[(a_{0,ij}^{(g)})^2]$.
- Let $R^{(g)}(s,t) = \mathrm{Cov}(e_{0i}^{(g)}(s), e_{0i}^{(g)}(t))$ be the covariance of the residual.

The covariance kernel simplifies to:

$$K^{(g)}(s,t) = \sum_{j=1}^{m_g} \lambda_j^{(g)} \phi_{0j}^{(g)}(s) \phi_{0j}^{(g)}(t) + R^{(g)}(s,t). \tag{C.2}$$

By definition, $\psi(t)$ is an eigenfunction of the covariance operator $K^{(g)}$ with eigenvalue ν if $\int_{\mathcal{T}} K^{(g)}(s,t)\psi(t)dt = \nu\psi(s)$. We test the function $\phi_{0k}^{(g)}(t)$ for some $k \in \{1, \ldots, m_g\}$:

$$\begin{aligned}
\int_{\mathcal{T}} K^{(g)}(s,t)\phi_{0k}^{(g)}(t)dt &= \int_{\mathcal{T}} \left(\sum_{j=1}^{m_g} \lambda_j^{(g)} \phi_{0j}^{(g)}(s) \phi_{0j}^{(g)}(t) + R^{(g)}(s,t) \right) \phi_{0k}^{(g)}(t)dt \\
&= \sum_{j=1}^{m_g} \lambda_j^{(g)} \phi_{0j}^{(g)}(s) \int_{\mathcal{T}} \phi_{0j}^{(g)}(t)\phi_{0k}^{(g)}(t)dt + \int_{\mathcal{T}} R^{(g)}(s,t)\phi_{0k}^{(g)}(t)dt.
\end{aligned}$$

Using the orthonormality of the basis functions, $\int_{\mathcal{T}} \phi_{0j}^{(g)}(t)\phi_{0k}^{(g)}(t)dt = \delta_{jk}$. For the residual term, we use the assumption that $e_{0i}^{(g)}(\cdot)$ is orthogonal to the basis functions $\phi_{0k}^{(g)}(\cdot)$. This implies $\int_{\mathcal{T}} e_{0i}^{(g)}(t)\phi_{0k}^{(g)}(t)dt = 0$ almost surely, and consequently:

$$\int_{\mathcal{T}} R^{(g)}(s,t)\phi_{0k}^{(g)}(t)dt = \mathbb{E}\left[e_{0i}^{(g)}(s) \int_{\mathcal{T}} e_{0i}^{(g)}(t)\phi_{0k}^{(g)}(t)dt \right] = 0.$$

Thus, the equation reduces to:

$$\int_{\mathcal{T}} K^{(g)}(s,t)\phi_{0k}^{(g)}(t)dt = \lambda_k^{(g)} \phi_{0k}^{(g)}(s). \tag{C.3}$$

This confirms that $\phi_{0k}^{(g)}(\cdot)$ is an eigenfunction of $K^{(g)}(s,t)$ with eigenvalue $\lambda_k^{(g)} = \mathrm{Var}(a_{0,ik}^{(g)})$.

The eigenvalues of $K^{(g)}$ consist of $\{\lambda_1^{(g)}, \ldots, \lambda_{m_g}^{(g)}\}$ derived from the signal part, and the eigenvalues of $R^{(g)}$, say $\{\nu_l\}$. The lemma assumes:

1. $\lambda_1^{(g)} > \lambda_2^{(g)} > \cdots > \lambda_{m_g}^{(g)}$.

2. All eigenvalues of $R^{(g)}$ are less than $\lambda_{m_g}^{(g)}$.

These conditions ensure strictly ordered eigenvalues: $\lambda_1^{(g)} > \cdots > \lambda_{m_g}^{(g)} > \max_l(\nu_l)$. Therefore, the j-th eigenfunction $\gamma_j^{(g)}(\cdot)$ corresponds uniquely to $\phi_{0j}^{(g)}(\cdot)$ for $j = 1, \ldots, m_g$.

Finally, we calculate the functional score $a_{ij}^{(g)}$:

$$\begin{aligned} a_{ij}^{(g)} &:= \int_{\mathcal{T}} X_i^{(g)}(t)\gamma_j^{(g)}(t)dt = \int_{\mathcal{T}} X_i^{(g)}(t)\phi_{0j}^{(g)}(t)dt \\ &= \int_{\mathcal{T}} \left(\sum_{k=1}^{m_g} a_{0,ik}^{(g)}\phi_{0k}^{(g)}(t) + e_{0i}^{(g)}(t) \right) \phi_{0j}^{(g)}(t)dt \\ &= \sum_{k=1}^{m_g} a_{0,ik}^{(g)}\delta_{kj} + 0 = a_{0,ij}^{(g)}. \end{aligned}$$

This completes the proof.

C.2 Proof of Lemma 2

Proof: The proof consists of two steps: establishing the rotation between bases within each group and deriving the joint covariance structure.

For the g-th group, let $\Sigma_0^{(g)} = \mathrm{Cov}(\boldsymbol{a}_{0,i}^{(g)})$. By spectral decomposition, $\Sigma_0^{(g)} = P^{(g)}\Lambda^{(g)}{P^{(g)}}^\top$, where $P^{(g)}$ is orthogonal and $\Lambda^{(g)}$ is diagonal. The covariance kernel of the signal part is

$$K^{(g)}(s,t) = \boldsymbol{\phi}_0^{(g)}(s)^\top \Sigma_0^{(g)} \boldsymbol{\phi}_0^{(g)}(t) = \left({P^{(g)}}^\top \boldsymbol{\phi}_0^{(g)}(s) \right)^\top \Lambda^{(g)} \left({P^{(g)}}^\top \boldsymbol{\phi}_0^{(g)}(t) \right).$$

This implies the K-L eigenfunctions are $\boldsymbol{\gamma}^{(g)}(t) = {P^{(g)}}^\top \boldsymbol{\phi}_0^{(g)}(t)$, or equivalently $\boldsymbol{\phi}_0^{(g)}(t) = P^{(g)}\boldsymbol{\gamma}^{(g)}(t)$. By equating the expansions $X_i^{(g)}(t) = \boldsymbol{a}_{0,i}^{(g)\top}\boldsymbol{\phi}_0^{(g)}(t) = \boldsymbol{a}_i^{(g)\top}\boldsymbol{\gamma}^{(g)}(t)$, we obtain the coefficient relationship:

$$\boldsymbol{a}_{0,i}^{(g)} = P^{(g)}\boldsymbol{a}_i^{(g)}. \tag{C.4}$$

Define the block-diagonal matrix $\mathbb{P} = \mathrm{diag}(P^{(1)}, \ldots, P^{(G)})$. Using (C.4), the joint coefficient

vectors satisfy $\boldsymbol{a}_{0,i} = \mathbb{P}\boldsymbol{a}_i$. Since $\mathbb{P}$ is orthogonal, $\boldsymbol{a}_i = \mathbb{P}^\top \boldsymbol{a}_{0,i}$. The covariance of the functional scores is then derived as:

$$\mathrm{Cov}(\boldsymbol{a}_i) = \mathrm{Cov}(\mathbb{P}^\top \boldsymbol{a}_{0,i}) = \mathbb{P}^\top \mathrm{Cov}(\boldsymbol{a}_{0,i})\mathbb{P}.$$

This completes the proof.

C.3 Proof of Lemma 3

Proof: Instead of relying on element-wise Taylor expansions, we analyze the perturbation of the covariance matrix through the lens of spectral operators. Let the population covariance matrix of the functional scores be $\boldsymbol{\Sigma}_C(\omega) = C_{\boldsymbol{a}}(\omega) = \mathbf{B}\mathbf{B}^\top + \omega\Lambda_{\boldsymbol{u}}$. We define the base unperturbed matrix as $\boldsymbol{\Sigma}_B = \mathbf{B}\mathbf{B}^\top$, which has rank K. The idiosyncratic component $\omega\Lambda_{\boldsymbol{u}}$ is thus treated strictly as a perturbation matrix $\mathbf{E} = \omega\Lambda_{\boldsymbol{u}}$.

Let $\Lambda_{C,K}(\omega)$ and $\Lambda_{B,K}$ be the $K \times K$ diagonal matrices containing the top K eigenvalues of $\boldsymbol{\Sigma}_C(\omega)$ and $\boldsymbol{\Sigma}_B$, respectively. Let $\mathbf{V}_C(\omega) = (\boldsymbol{\xi}_1^{(C)}(\omega), \ldots, \boldsymbol{\xi}_K^{(C)}(\omega))$ and $\mathbf{V}_B = (\boldsymbol{\xi}_1^{(B)}, \ldots, \boldsymbol{\xi}_K^{(B)})$ be their corresponding eigenvector matrices.

First, by Weyl's inequality, the perturbation of the eigenvalues is bounded by the spectral norm of the perturbation matrix:

$$\max_{1\le j\le K} |\lambda_j^{(C)}(\omega) - \lambda_j^{(B)}| \le \|\mathbf{E}\|_2 = \omega\lambda_{\max}(\Lambda_{\boldsymbol{u}}). \tag{C.5}$$

Consequently, we can bound the difference between the inverse square roots of the eigen-

value matrices:

$$\begin{aligned}\|\Lambda_{C,K}^{-1/2}(\omega)-\Lambda_{B,K}^{-1/2}\|_2 &= \max_{1\le j\le K}\left|\frac{1}{\sqrt{\lambda_j^{(C)}(\omega)}}-\frac{1}{\sqrt{\lambda_j^{(B)}}}\right| \\ &= \max_{1\le j\le K}\frac{|\lambda_j^{(B)}-\lambda_j^{(C)}(\omega)|}{\sqrt{\lambda_j^{(B)}\lambda_j^{(C)}(\omega)}\left(\sqrt{\lambda_j^{(B)}}+\sqrt{\lambda_j^{(C)}(\omega)}\right)} \\ &\le \frac{\omega\lambda_{\max}(\Lambda_{\boldsymbol{u}})}{2(\lambda_K^{(B)})^{3/2}} = O(\omega).\end{aligned}$$

Second, let $\mathbf{P}_C(\omega)=\mathbf{V}_C(\omega)\mathbf{V}_C(\omega)^\top$ and $\mathbf{P}_B=\mathbf{V}_B\mathbf{V}_B^\top$ be the projection matrices onto the top K eigenspaces. Since $\boldsymbol{\Sigma}_B$ has rank K, its $(K+1)$-th eigenvalue is exactly 0. The principal eigengap is $\delta_K=\lambda_K^{(B)}-0=\lambda_K^{(B)}$. By the Davis-Kahan $\sin\Theta$ theorem, the distance between the projection matrices satisfies:

$$\|\mathbf{P}_C(\omega)-\mathbf{P}_B\|_2 \le \frac{\|\mathbf{E}\|_2}{\delta_K} = \frac{\omega\lambda_{\max}(\Lambda_{\boldsymbol{u}})}{\lambda_K^{(B)}} = O(\omega). \tag{C.6}$$

Under standard eigendecomposition alignment, this implies the eigenvector rotation is also bounded as $\|\mathbf{V}_C(\omega)-\mathbf{V}_B\|_2\le O(\omega)$.

Now, we evaluate the difference in the estimated factors. The factor vectors are defined as $\boldsymbol{f}_i^\top(\omega)=\boldsymbol{a}_i^\top\mathbf{V}_C(\omega)\Lambda_{C,K}^{-1/2}(\omega)$ and the ideal factor $\boldsymbol{f}_i^\top=(\boldsymbol{a}_i^\top-\boldsymbol{u}_i^\top)\mathbf{V}_B\Lambda_{B,K}^{-1/2}$. Decomposing the difference gives:

$$\begin{aligned}\boldsymbol{f}_i^\top(\omega)-\boldsymbol{f}_i^\top &= \boldsymbol{a}_i^\top\left[\mathbf{V}_C(\omega)\Lambda_{C,K}^{-1/2}(\omega)-\mathbf{V}_B\Lambda_{B,K}^{-1/2}\right]+\boldsymbol{u}_i^\top\mathbf{V}_B\Lambda_{B,K}^{-1/2} \\ &= \boldsymbol{a}_i^\top\left[\mathbf{V}_C(\omega)\left(\Lambda_{C,K}^{-1/2}(\omega)-\Lambda_{B,K}^{-1/2}\right)+\left(\mathbf{V}_C(\omega)-\mathbf{V}_B\right)\Lambda_{B,K}^{-1/2}\right]+\boldsymbol{u}_i^\top\mathbf{V}_B\Lambda_{B,K}^{-1/2}.\end{aligned}$$

Taking the L_2-norm and applying the triangle inequality along with our Weyl and Davis-Kahan bounds:

$$\begin{aligned}\|\boldsymbol{f}_i^\top(\omega)-\boldsymbol{f}_i^\top\|_2 &\le \|\boldsymbol{a}_i\|_2\left(\underbrace{\|\mathbf{V}_C(\omega)\|_2}_{=1}\|\Lambda_{C,K}^{-1/2}(\omega)-\Lambda_{B,K}^{-1/2}\|_2+\|\mathbf{V}_C(\omega)-\mathbf{V}_B\|_2\|\Lambda_{B,K}^{-1/2}\|_2\right) \\ &\quad+\|\boldsymbol{u}_i^\top\mathbf{V}_B\Lambda_{B,K}^{-1/2}\|_2 \\ &\le \|\boldsymbol{a}_i\|_2\Big(1\cdot O(\omega)+O(\omega)\cdot(\lambda_K^{(B)})^{-1/2}\Big)+\|\boldsymbol{u}_i^\top\mathbf{V}_B\Lambda_{B,K}^{-1/2}\|_2.\end{aligned}$$

Squaring both sides yields:

$$\|\boldsymbol{f}_i^\top(\omega) - \boldsymbol{f}_i^\top\|_2^2 = O_P\left(\omega^2\|\boldsymbol{a}_i\|_2^2 + \|\boldsymbol{u}_i^\top \mathbf{B}\Lambda_B^{-1}\|_2^2\right). \tag{C.7}$$

Similarly, for the idiosyncratic components, recall that $\boldsymbol{a}_i = \mathbf{B}\boldsymbol{f}_i + \boldsymbol{u}_i$, and $\mathbf{P}_B\boldsymbol{a}_i = \mathbf{B}\boldsymbol{f}_i + \mathbf{P}_B\boldsymbol{u}_i$. The ideal idiosyncratic component satisfies $\boldsymbol{u}_i = \boldsymbol{a}_i - \mathbf{B}\boldsymbol{f}_i = (\mathbf{I}_p - \mathbf{P}_B)\boldsymbol{a}_i + \mathbf{P}_B\boldsymbol{u}_i$. The ω-dependent component is constructed as $\boldsymbol{u}_i(\omega) = (\mathbf{I}_p - \mathbf{P}_C(\omega))\boldsymbol{a}_i$. Thus, their difference is:

$$\boldsymbol{u}_i^\top(\omega) - \boldsymbol{u}_i^\top = \boldsymbol{a}_i^\top\left(\mathbf{P}_B - \mathbf{P}_C(\omega)\right) - \boldsymbol{u}_i^\top\mathbf{P}_B.$$

Taking the L_2-norm and applying the Davis-Kahan bound from (C.6):

$$\begin{aligned}\|\boldsymbol{u}_i^\top(\omega) - \boldsymbol{u}_i^\top\|_2 &\leq \|\mathbf{P}_B - \mathbf{P}_C(\omega)\|_2\|\boldsymbol{a}_i\|_2 + \|\boldsymbol{u}_i^\top\mathbf{P}_B\|_2 \\ &\leq O(\omega)\|\boldsymbol{a}_i\|_2 + \|\boldsymbol{u}_i^\top\mathbf{P}_B\|_2.\end{aligned}$$

Noting that $\mathbf{P}_B = \mathbf{B}(\mathbf{B}^\top\mathbf{B})^{-1}\mathbf{B}^\top$ is the exact projection onto the factor loading space, squaring this term gives the final bound:

$$\|\boldsymbol{u}_i^\top(\omega) - \boldsymbol{u}_i^\top\|_2^2 = O_P\left(\omega^2\|\boldsymbol{a}_i\|_2^2 + \|\boldsymbol{u}_i^\top\mathbf{B}\Lambda_B^{-1}\mathbf{B}^\top\|_2^2\right). \tag{C.8}$$

This completes the proof.

C.4 Proof of Lemma 4

Proof: Under Assumption 1, 2 and Assumption B1-B4 in Section B, using Lemma 2(a) in Kong et al. (2016). $i = 1, \ldots, n; k_1, k_2 = 1, \ldots, s_n; j = 1, \ldots, G$

$$\widehat{\boldsymbol{A}}^\top \widehat{\boldsymbol{A}} = (\sum_{i=1}^{n} \boldsymbol{a}_i \boldsymbol{a}_i^\top)_{Gs_n \times Gs_n} = \sum_{i=1}^{n} \begin{pmatrix} \boldsymbol{a}_i^{(1)} \boldsymbol{a}_i^{(1)\top} & \ldots & \boldsymbol{a}_i^{(1)} \boldsymbol{a}_i^{(G)\top} \\ \ldots & \ldots & \ldots \\ \boldsymbol{a}_i^{(G)} \boldsymbol{a}_i^{(1)\top} & \ldots & \boldsymbol{a}_i^{(G)} \boldsymbol{a}_i^{(G)\top} \end{pmatrix}$$

$$= \sum_{i=1}^{n} \begin{pmatrix} (a_{ik_1}^{(1)} a_{ik_2}^{(1)})_{s_n \times s_n} & \ldots & (a_{ik_1}^{(1)} a_{ik_2}^{(G)})_{s_n \times s_n} \\ \ldots & \ldots & \ldots \\ (a_{ik_1}^{(G)} a_{ik_2}^{(1)})_{s_n \times s_n} & \ldots & (a_{ik_1}^{(G)} a_{ik_2}^{(G)})_{s_n \times s_n} \end{pmatrix}$$

$$= \begin{pmatrix} (\sum_{i=1}^{n} a_{ik_1}^{(1)} a_{ik_2}^{(1)})_{s_n \times s_n} & \ldots & (\sum_{i=1}^{n} a_{ik_1}^{(1)} a_{ik_2}^{(G)})_{s_n \times s_n} \\ \ldots & \ldots & \ldots \\ (\sum_{i=1}^{n} a_{ik_1}^{(G)} a_{ik_2}^{(1)})_{s_n \times s_n} & \ldots & (\sum_{i=1}^{n} a_{ik_1}^{(G)} a_{ik_2}^{(G)})_{s_n \times s_n} \end{pmatrix},$$

then the analysis for $\{(\boldsymbol{A}^\top \boldsymbol{A})_{ij} - E(\boldsymbol{A}^\top \boldsymbol{A})_{ij}\}/n$ is is equivalent to analyzing the following

.

$$n^{-1} \sum_{i=1}^{n} [\widehat{a}_{ik_1}^{(j)} \widehat{a}_{ik_2}^{(j)} - E(a_{ik_1}^{(j)} a_{ik_2}^{(j)})] = n^{-1} \sum_{i=1}^{n} [(\widehat{a}_{ik_1}^{(j)} \widehat{a}_{ik_2}^{(j)} - a_{ik_1}^{(j)} a_{ik_2}^{(j)}) + (a_{ik_1}^{(j)} a_{ik_2}^{(j)} - E(a_{ik_1}^{(j)} a_{ik_2}^{(j)}))]$$

$$= O_p(k_1^{a/2+1} n^{-1/2} + k_2^{a/2+1} n^{-1/2} + n^{-1/2})$$

$$\leq O_p(s_n^{a/2+1} n^{-1/2})$$

$$\|n^{-1} \widehat{\boldsymbol{A}}^\top \widehat{\boldsymbol{A}} - \boldsymbol{\Sigma}_A\|_{\max} = \max_{j,k} \left| n^{-1} \sum_{i=1}^{n} [\widehat{a}_{ik_1}^{(j)} \widehat{a}_{ik_2}^{(j)} - E(a_{ik_1}^{(j)} a_{ik_2}^{(j)})] \right|$$

$$= O_p(s_n^{a/2+1} n^{-1/2}).$$

Under the fixed-dimension regime or appropriate growth rate of s_n, this implies the spectral norm of the perturbation $\boldsymbol{E} = n^{-1} \widehat{\boldsymbol{A}}^\top \widehat{\boldsymbol{A}} - \boldsymbol{\Sigma}_A$ satisfies $\|\boldsymbol{E}\|_2 = O_p(n^{-1/2})$.

This completes the proof.

D Proof of Theorem

D.1 Proof of Theorem 1

Proof: To establish the bound for factor estimation, we treat the empirical covariance matrix as a perturbed version of the population covariance. Let $\mathbf{\Sigma} = \mathbb{E}(\boldsymbol{a}_i \boldsymbol{a}_i^\top)$ and $\widehat{\mathbf{\Sigma}} = \frac{1}{n}\widehat{\boldsymbol{A}}^\top \widehat{\boldsymbol{A}}$. From Lemma 4, we have the concentration result:

$$\boldsymbol{E} = \widehat{\mathbf{\Sigma}} - \mathbf{\Sigma}, \quad \text{with } \|\boldsymbol{E}\|_2 = O_P(n^{-1/2}).$$

Let $\mathbf{V}_K = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_K)$ be the matrix of the top K eigenvectors of $\mathbf{\Sigma}$, and $\widehat{\mathbf{V}}_K = (\widehat{\boldsymbol{\xi}}_1, \ldots, \widehat{\boldsymbol{\xi}}_K)$ be its empirical counterpart. Instead of expanding the loss function, we invoke the Davis-Kahan Theorem. Let $\delta_n = \lambda_K - \lambda_{K+1}$ denote the principal eigengap of $\mathbf{\Sigma}$. The distance between the estimated and population eigenspaces is bounded as:

$$\|\sin\Theta(\widehat{\mathbf{V}}_K, \mathbf{V}_K)\|_2 \leq \frac{\|\boldsymbol{E}\|_2}{\delta_n} = O_P\left(\frac{1}{\delta_n\sqrt{n}}\right). \tag{D.1}$$

The factor estimation error can be decomposed using the projection matrices $\widehat{\mathbf{P}}_K = \widehat{\mathbf{V}}_K\widehat{\mathbf{V}}_K^\top$ and $\mathbf{P}_K = \mathbf{V}_K\mathbf{V}_K^\top$. Under Assumption 1, the spectral norm of the difference in projections is equivalent to the sine of the canonical angles: $\|\widehat{\mathbf{P}}_K - \mathbf{P}_K\|_2 = \|\sin\Theta(\widehat{\mathbf{V}}_K, \mathbf{V}_K)\|_2$.

For the estimated factors $\widehat{\boldsymbol{f}}_i$ and the ideal factors $\boldsymbol{f}_i(\omega)$, the estimation error is dominated by the rotation of the singular subspace. By combining (D.1) with the results from Lemma 3, we obtain:

$$\begin{aligned}\|\widehat{\boldsymbol{f}}_i^\top - \boldsymbol{f}_i^\top(\omega)\|_2^2 &\leq C \cdot \|\widehat{\mathbf{P}}_K - \mathbf{P}_K\|_2^2 \cdot \|\boldsymbol{a}_i\|_2^2 \\ &= O_P\left(\frac{1}{n\delta_n^2}\|\boldsymbol{a}_i\|_2^2\right).\end{aligned}$$

Similarly, for the idiosyncratic components $\widehat{\boldsymbol{u}}_i$:

$$\|\widehat{\boldsymbol{u}}_i^\top - \boldsymbol{u}_i^\top(\omega)\|_2^2 \leq O_P\left(\frac{1}{n\delta_n^2}\|\boldsymbol{a}_i\|_2^2\right).$$

Since δ_n is assumed to be a positive constant (or bounded away from zero) under the factor model identification conditions, this simplifies to $O_P(n^{-1}\|\boldsymbol{a}_i\|_2^2)$, which completes the proof.

D.2 Useful Lemmas

In this subsection, we summarize several technical lemmas required for our main theoretical analysis. These results cover matrix perturbation bounds, properties of the loss function Hessian, and general convergence guarantees for convex optimization.

The first lemma provides a perturbation bound for the inverse of a matrix sum, which is adapted from Fan et al. (2020).

Lemma D.1. *(Lemma B.1 in Fan et al. (2020)) Suppose* $\mathbf{A} \in \mathbb{R}^{q\times r}$ *and* $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{r\times r}$*. If* $\left\|\mathbf{C}\mathbf{B}^{-1}\right\| < 1$*, where* $\|\cdot\|$ *denotes an induced norm, then*

$$\left\|\mathbf{A}\left[(\mathbf{B}+\mathbf{C})^{-1} - \mathbf{B}^{-1}\right]\right\| \leq \frac{\left\|\mathbf{A}\mathbf{B}^{-1}\right\| \cdot \left\|\mathbf{C}\mathbf{B}^{-1}\right\|}{1 - \left\|\mathbf{C}\mathbf{B}^{-1}\right\|}.$$

The following lemma establishes key properties of the Hessian matrix of the loss function $L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta})$ (denoted as $L(\boldsymbol{\theta})$ for brevity) under Assumptions 3–5.

Lemma D.2. *(Lemma C.1 of Fan et al. (2020)) Let Assumptions 3–5 hold. Consider* $L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta})$ *as a function of* $\boldsymbol{\theta}$*. Under Assumption 5, we have* $\|\boldsymbol{W}\|_{\max} \leq M_0/2$*. Then, the following bounds hold:*

(i) Lipschitz continuity of the Hessian restricted to subspace S*:*

$$\left\|\nabla^2_{\cdot S} L(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta}) - \nabla^2_{\cdot S} L\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right\|_\infty \leq M\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right\|_2$$

for all $\boldsymbol{\theta}$ *satisfying* $\operatorname{supp}(\boldsymbol{\theta}) \subseteq S$*.*

(ii) Infinity-norm bound on the inverse Hessian:

$$\left\|\left(\nabla^2_{SS} L\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_\infty \leq \frac{1}{2\kappa_\infty}.$$

(iii) Spectral-norm bound on the inverse Hessian:

$$\left\|\left(\nabla^2_{SS} L\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_2 \leq \frac{1}{2\kappa_2}.$$

(iv) Irrepresentable condition bound:

$$\left\|\nabla^2_{S_2S} L\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\left(\nabla^2_{SS} L\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_\infty \leq 1 - \tau.$$

Building on Lemma D.2, the next result ensures the invertibility of the Hessian within a local neighborhood of the true parameter $\boldsymbol{\theta}^*$.

Lemma D.3. *Under conditions (i)–(iii) in Lemma D.2, for any $\boldsymbol{\theta}$ in the neighborhood $B_S\left(\boldsymbol{\theta}^*, \min\left\{A, \frac{\kappa_\infty}{M}\right\}\right)$, we have:*

$$\left\|\left(\nabla^2_{SS} L_n(\boldsymbol{\theta})\right)^{-1}\right\|_2 < \kappa_2^{-1} \quad \textit{and} \quad \left\|\left(\nabla^2_{SS} L_n(\boldsymbol{\theta})\right)^{-1}\right\|_\infty < \kappa_\infty^{-1}.$$

The following two lemmas provide general theoretical guarantee for regularized M-estimators, establishing error bounds and conditions for unique global minimizers.

Lemma D.4. *Suppose $\lambda \geq 0$, $\mathcal{M}$ is a Euclidean space, and $\boldsymbol{\theta}_0 \in \mathcal{M}$. Let $L(\boldsymbol{\theta}) \in C^2(\mathcal{M})$ and $R(\boldsymbol{\theta})$ be convex functions. Assume there exist $\kappa, A > 0$ such that $\nabla^2 L(\boldsymbol{\theta}) \succeq \kappa I$ whenever $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2 \leq A$. If the condition*

$$\|\nabla L\left(\boldsymbol{\theta}_0\right)\|_2 + \lambda \inf_{\mathbf{h} \in \partial R(\boldsymbol{\theta}_0)} \|\mathbf{h}\|_2 < \frac{1}{2}\kappa A$$

holds, then the regularized objective $L_\lambda(\boldsymbol{\theta}) = L(\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta})$ has a unique minimizer $\widehat{\boldsymbol{\theta}}$, which satisfies

$$\left\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right\|_2 \leq \frac{2}{\kappa}\left(\|\nabla L\left(\boldsymbol{\theta}_0\right)\|_2 + \lambda \inf_{\mathbf{h} \in \partial R(\boldsymbol{\theta}_0)} \|\mathbf{h}\|_2\right).$$

Lemma D.5. *Suppose $L(\boldsymbol{\theta}) \in C^2(\mathbb{R}^p)$ is convex. Let $R(\boldsymbol{\theta})$ be a convex decomposable regularizer such that $R(\boldsymbol{\alpha}+\boldsymbol{\beta}) = R(\boldsymbol{\alpha})+R(\boldsymbol{\beta})$ for $\boldsymbol{\alpha} \in \mathcal{M}$ and $\boldsymbol{\beta} \in \mathcal{M}^{\perp}$, where $\mathcal{M}$ is a linear subspace of $\mathbb{R}^p$. Assume there exists a dual norm $R^*(\cdot)$ such that $|\langle \boldsymbol{\alpha}, \boldsymbol{\beta} \rangle| \leq R(\boldsymbol{\alpha}) R^*(\boldsymbol{\beta})$ for $\boldsymbol{\alpha} \in \mathcal{M}^{\perp}$ and $\boldsymbol{\beta} \in \mathbb{R}^p$.*

Let $\widehat{\boldsymbol{\theta}} \in \operatorname{argmin}_{\boldsymbol{\theta} \in \mathcal{M}} L_\lambda(\boldsymbol{\theta})$. If the strict dual feasibility condition $R^*(\nabla L(\widehat{\boldsymbol{\theta}})) < \lambda$ holds and the restricted Hessian satisfies $\boldsymbol{\theta}^\top \nabla^2 L(\widehat{\boldsymbol{\theta}}) \boldsymbol{\theta} > 0$ for all non-zero $\boldsymbol{\theta} \in \mathcal{M}$, then $\widehat{\boldsymbol{\theta}}$ is the unique global minimizer of $L_\lambda(\boldsymbol{\theta})$.

Finally, we restate the error bounds and sign consistency results for the optimization problem (D.2) derived in Fan et al. (2020).

Lemma D.6. *(Theorem B.1 of Fan et al. (2020)) Consider the optimization problem* (D.2)*. If Lemma D.2 is satisfied, then the following results hold:*

1. ***Error bounds**: If the regularization parameter λ satisfies*

$$\frac{7}{\tau} \left\| \nabla L_n(\boldsymbol{\theta}^*) \right\|_\infty < \lambda < \frac{\kappa_2}{4\sqrt{|S|}} \min\left\{ A, \frac{\kappa_\infty \tau}{3M} \right\},$$

then $\operatorname{supp}(\widehat{\boldsymbol{\theta}}) \subseteq S$, and the estimator satisfies:

$$\begin{aligned}
&\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_\infty \leq \frac{3}{5\kappa_\infty} \left(\left\| \nabla_S L_n(\boldsymbol{\theta}^*) \right\|_\infty + \lambda \right), \\
&\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_2 \leq \frac{2}{\kappa_2} \left(\left\| \nabla_S L_n(\boldsymbol{\theta}^*) \right\|_2 + \lambda \sqrt{|S_1|} \right), \\
&\left\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \right\|_1 \leq \min\left\{ \frac{3}{5\kappa_\infty} \left(\left\| \nabla_S L_n(\boldsymbol{\theta}^*) \right\|_1 + \lambda |S_1| \right), \frac{2\sqrt{|S|}}{\kappa_2} \left(\left\| \nabla_S L_n(\boldsymbol{\theta}^*) \right\|_2 + \lambda \sqrt{|S_1|} \right) \right\}.
\end{aligned}$$

2. ***Sign consistency**: Suppose there exists a constant $C \geq 5$ such that*

$$\min\left\{ \left| \mathbf{H}^*_{1j} \right| : \mathbf{H}^*_{1j} \neq 0, j \in [p_1] \right\} > \frac{C}{\kappa_\infty \tau} \left\| \nabla L_n(\boldsymbol{\theta}^*) \right\|_\infty$$

and

$$\left\| \nabla L_n(\boldsymbol{\theta}^*) \right\|_\infty < \frac{\kappa_2 \tau}{7C\sqrt{|S|}} \min\left\{ A, \frac{\kappa_\infty \tau}{3M} \right\}.$$

Then, by choosing $\lambda \in \left(\frac{7}{\tau}\left\|\nabla L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty, \frac{1}{\tau}\left(\frac{5C}{3}-1\right)\left\|\nabla L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty\right)$*, the estimator achieves sign consistency, i.e.,* $\operatorname{sign}(\widehat{\mathbf{H}}) = \operatorname{sign}(\mathbf{H}^*)$.

Recall that

$$\widehat{\boldsymbol{\theta}} = \begin{pmatrix} \widehat{\mathbf{H}}_1 \\ \widehat{\boldsymbol{\gamma}} \end{pmatrix} = \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ L_n(\mathbf{y}, \widehat{\mathbf{W}}\boldsymbol{\theta}) + \lambda \left\|\boldsymbol{\theta}_{[p_1]}\right\|_1 \right\}. \tag{D.2}$$

Also, Assumption 5 tells us $\mathbf{V}_0$ is nonsingular, and so is $\mathbf{V} = \begin{pmatrix} \mathbf{I}_{p_1} & \mathbf{0}_{p_1\times K} \\ \mathbf{0}_{K\times p_1} & \mathbf{V}_0 \end{pmatrix}$.

D.3 Proof of Theorem 2

Proof: Define $\overline{\mathbf{W}} = \left(1, \bar{\boldsymbol{u}}_t^\top, \bar{\boldsymbol{f}}_t^\top\right) = \widehat{\mathbf{W}}\mathbf{V} = \left(1, \widehat{\boldsymbol{u}}_t^\top, \widehat{\boldsymbol{f}}_t^\top\right)\mathbf{V}$, $\bar{\boldsymbol{\theta}} = \mathbf{V}^{-1}\widehat{\boldsymbol{\theta}}$, $\widehat{\mathbf{B}}_0 = \left(\mathbf{0}_K^\top, \widehat{\mathbf{B}}^\top\right)^\top$, $\widehat{\boldsymbol{\theta}}^* = \begin{pmatrix} \mathbf{H}_1^* \\ \widehat{\mathbf{B}}_0^\top \mathbf{H}_1^* \end{pmatrix}$ and $\bar{\boldsymbol{\theta}}^* = \mathbf{V}^{-1}\widehat{\boldsymbol{\theta}}^*$. We easily see that $\widehat{\mathbf{H}} = \widehat{\boldsymbol{\theta}}_{[p_1]} = \bar{\boldsymbol{\theta}}_{[p_1]}$ and

$$\bar{\boldsymbol{\theta}} = \operatorname{argmin}_{\boldsymbol{\theta}} \left\{ L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta}) + \lambda \left\|\boldsymbol{\theta}_{[p_1]}\right\|_1 \right\}.$$

Then it follows that $\operatorname{supp}(\widehat{\mathbf{H}}_1) = \operatorname{supp}\left(\bar{\boldsymbol{\theta}}_{[p_1]}\right)$ and $\left\|\widehat{\mathbf{H}}_1 - \mathbf{H}_1^*\right\| = \left\|\bar{\boldsymbol{\theta}}_{[p_1]} - \bar{\boldsymbol{\theta}}_{[p_1]}^*\right\| \le \left\|\bar{\boldsymbol{\theta}} - \bar{\boldsymbol{\theta}}^*\right\|$ for any norm $\|\cdot\|$.

Consequently, Theorem 2 is reduced to studying $\bar{\boldsymbol{\theta}}$ and the loss function $L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta})$. Lemma D.2 below implies that all the regularity conditions (with $A = \infty$) in Lemma D.6 are satisfied.

Let w_{tj} and $\bar{w}_{tj}$ be the (t,j)-th element of $\mathbf{W}$ and $\overline{\mathbf{W}}$, respectively. Observe that $L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^n \left[-y_t \overline{\mathbf{w}}_t^\top \boldsymbol{\theta} + b\left(\overline{\mathbf{w}}_t^\top \boldsymbol{\theta}\right)\right]$, $\nabla L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^n \left[-y_t + b'\left(\overline{\mathbf{w}}_t^\top \boldsymbol{\theta}\right)\right]\overline{\mathbf{w}}_t$ and $\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^* = \mathbf{A}\mathbf{H}^* + \beta_0^* \mathbf{1}_n$. Hence $\left\|\nabla L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right\|_\infty = \varepsilon^*$ and consequently, $\left\|\nabla_S L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right\|_\infty \le \varepsilon^*$, $\left\|\nabla_S L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right\|_2 \le \varepsilon^*\sqrt{|S|}$ and $\left\|\nabla_S L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right\|_1 \le \varepsilon^*|S|$. In addition, $\lambda > 7\varepsilon^*/\tau \ge \varepsilon^*$.

Based on these estimates, all the results follow from Theorem 2 and some simple algebra.

Here we present Lemma D.2 used above and its proof.

(i) Based on the fact that $\mathbf{W}\boldsymbol{\theta}^* = \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^* = \mathbf{A}\mathbf{H}^* + \beta_0^*\mathbf{1}_n$, we have $\nabla^2 L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*) = \frac{1}{n}\sum_{t=1}^n b''\left(\overline{\mathbf{w}}_t^\top \bar{\boldsymbol{\theta}}^*\right)\mathbf{w}_t\mathbf{w}_t^\top$ and $\nabla^2 L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right) = \frac{1}{n}\sum_{t=1}^n b''\left(\overline{\mathbf{w}}_t^\top \bar{\boldsymbol{\theta}}^*\right)\overline{\mathbf{w}}_t\overline{\mathbf{w}}_t^\top$. For any $j, k \in [p_1 + K]$ and $\mathrm{supp}(\boldsymbol{\theta}) \subseteq S$,

$$\begin{aligned}\left|\nabla^2_{jk} L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta}) - \nabla^2_{jk} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right| &\le \frac{1}{n}\sum_{t=1}^n \left|b''\left(\overline{\mathbf{w}}_t^\top \boldsymbol{\theta}\right) - b''\left(\overline{\mathbf{w}}_t^\top \bar{\boldsymbol{\theta}}^*\right)\right| \cdot |\overline{\mathbf{w}}_{tj}\overline{\mathbf{w}}_{tk}| \\ &\le \frac{1}{n}\sum_{t=1}^n M_3 \left|\overline{\mathbf{w}}_t^\top\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right)\right| \cdot \|\overline{\mathbf{W}}\|_{\max}^2. \qquad \text{(D.3)}\end{aligned}$$

By the Cauchy-Schwarz inequality and $\|\overline{\mathbf{W}}\|_{\max} \le \|\mathbf{W}\|_{\max} + \|\overline{\mathbf{W}} - \mathbf{W}\|_{\max} \le M_0$, we obtain that for $t \in [n]$, $\left|\overline{\mathbf{w}}_t^\top\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right)\right| = \left|\overline{\mathbf{w}}_{tS}^\top\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right)_S\right| \le \|\overline{\mathbf{w}}_{tS}\|_2 \left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right\|_2 \le \sqrt{|S|}M_0\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right\|_2$. Plugging this result back to (D.3), we get

$$\begin{aligned}&\left|\nabla^2_{jk} L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta}) - \nabla^2_{jk} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right| \le \sqrt{|S|}M_3M_0^3\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right\|_2, \quad \forall j, k \in [p+K], \\ &\left\|\nabla^2_{\cdot S} L_n(\mathbf{y}, \overline{\mathbf{W}}\boldsymbol{\theta}) - \nabla^2_{\cdot S} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right\|_\infty \le |S|^{3/2}M_3M_0^3\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right\|_2 = M\left\|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^*\right\|_2.\end{aligned}$$

(ii) Now we come to the second claim. For any $k \in [p+K]$,

$$\begin{aligned}\left\|\nabla^2_{kS} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right) - \nabla^2_{kS} L_n\left(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*\right)\right\|_\infty &\le \frac{1}{n}\sum_{t=1}^n b''\left(\mathbf{x}_t^\top\boldsymbol{\beta}^*\right)\left\|\bar{w}_{tk}\overline{\mathbf{w}}_{tS}^\top - w_{tk}\mathbf{w}_{tS}^\top\right\|_\infty \\ &\le \frac{M_2\sqrt{|S|}}{n}\sum_{t=1}^n \left\|\bar{w}_{tk}\overline{\mathbf{w}}_{tS}^\top - w_{tk}\mathbf{w}_{tS}^\top\right\|_2.\end{aligned}$$

Also, by $\|\mathbf{W}\|_{\max} \le M_0/2$ and $\|\overline{\mathbf{W}}\|_{\max} \le M_0$, we have

$$\begin{aligned}\left\|\bar{w}_{tk}\overline{\mathbf{w}}_{tS}^\top - w_{tk}\mathbf{w}_{tS}^\top\right\|_2 &\le |w_{tk}| \cdot \left\|\left(\overline{\mathbf{w}}_{tS} - \mathbf{w}_{tS}\right)^\top\right\|_2 + |\bar{w}_{tk} - w_{tk}| \cdot \left\|\overline{\mathbf{w}}_{tS}^\top\right\|_2 \\ &\le \|\mathbf{W}\|_{\max}\left\|\overline{\mathbf{w}}_{tS} - \mathbf{w}_{tS}\right\|_2 + |\bar{w}_{tk} - w_{tk}| \cdot \sqrt{|S|}\|\overline{\mathbf{W}}\|_{\max} \\ &\le \frac{M_0}{2}\left\|\overline{\mathbf{w}}_{tS} - \mathbf{w}_{tS}\right\|_2 + M_0\sqrt{|S|} \cdot |\bar{w}_{tk} - w_{tk}|.\end{aligned}$$

Define $\delta = \max_{j\in[p+K]}\left(\frac{1}{n}\sum_{t=1}^n |\bar{w}_{tj} - w_{tj}|^2\right)^{1/2}$. By Jensen's inequality, $\forall J \subseteq [p+K]$,

$$\frac{1}{n}\sum_{t=1}^n \|\overline{\mathbf{w}}_{tJ} - \mathbf{w}_{tJ}\|_2 \le \left(\frac{1}{n}\sum_{t=1}^n \|\overline{\mathbf{w}}_{tJ} - \mathbf{w}_{tJ}\|_2^2\right)^{1/2} \le \left(\frac{|J|}{n}\max_{j\in[p+K]}\sum_{t=1}^n |\bar{w}_{tj} - w_{tj}|^2\right)^{1/2} \le \sqrt{|J|}\delta.$$

As a result,

$$\left\|\nabla^2_{\cdot S}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)-\nabla^2_{\cdot S}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right\|_\infty=\max_{k\in[p+K]}\left\|\nabla^2_{kS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)-\nabla^2_{kS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right\|_\infty$$
$$\leq\frac{3}{2}M_0M_2|S|\delta. \tag{D.4}$$

Let $\alpha=\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\left[\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)-\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right]\right\|_\infty$. Then

$$\alpha\leq\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty\left\|\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)-\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right\|_\infty$$
$$\leq\frac{3}{8\kappa_\infty}M_0M_2|S|\delta\leq\frac{1}{2}. \tag{D.5}$$

Lemma D.1 yields

$$\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}-\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty\leq\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty\frac{\alpha}{1-\alpha}$$
$$\leq\frac{1}{4\kappa_\infty}\cdot\frac{\alpha}{1-\frac{1}{2}}\leq\frac{3}{16\kappa^2_\infty}M_0M_2|S|\delta.$$

We also have a cruder bound $\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}-\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty\leq\frac{1}{4\kappa_\infty}$, which leads to

$$\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_\infty\leq\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty+\frac{1}{4\kappa_\infty}\leq\frac{1}{2\kappa_\infty}. \tag{D.6}$$

(iii) The third argument follows from (D.6) easily. Since $\|\mathbf{A}\|_2\leq\|\mathbf{A}\|_\infty$ holds for any symmetric matrix $\mathbf{A}$, we have $\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}-\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_2\leq\frac{1}{4\kappa_\infty}\leq\frac{1}{4\kappa_2}$ and thus $\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_2\leq\frac{1}{2\kappa_2}$.

(iv) Finally we prove the last inequality. On the one hand,

$$\left\|\nabla^2_{S_2S}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}-\nabla^2_{S_2S}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty$$
$$\leq\left\|\nabla^2_{S_2S}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)-\nabla^2_{S_2S}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right\|_\infty\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_\infty$$
$$+\left\|\nabla^2_{S_2S}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\left[\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}-\left(\nabla^2_{SS}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right)^{-1}\right]\right\|_\infty.$$

From claim (ii) and (D.4) it is easy to see that

$$\left\|\nabla^2_{S_2S}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)-\nabla^2_{S_2S}L_n\left(\mathbf{y},\mathbf{W}\boldsymbol{\theta}^*\right)\right\|_\infty\left\|\left(\nabla^2_{SS}L_n\left(\mathbf{y},\overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_\infty\leq\frac{1}{4\kappa_\infty}3M_0M_2|S|\delta.$$

On the other hand, we can take $\mathbf{A} = \nabla^2_{S_2S} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)$, $\mathbf{B} = \nabla^2_{SS} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)$ and $\mathbf{C} = \nabla^2_{SS} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right) - \nabla^2_{SS} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)$. By Assumption 4, $\left\|\mathbf{A}\mathbf{B}^{-1}\right\|_\infty \leq 1 - 2\tau \leq 1$. Lemma D.1 forces that

$$\begin{aligned}
&\left\|\nabla^2_{S_2S} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)\left[\left(\nabla^2_{SS} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1} - \left(\nabla^2_{SS} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)\right)^{-1}\right]\right\|_\infty \\
&= \left\|\mathbf{A}\left[(\mathbf{B}+\mathbf{C})^{-1} - \mathbf{B}^{-1}\right]\right\|_\infty \leq \left\|\mathbf{A}\mathbf{B}^{-1}\right\|_\infty \frac{\|\mathbf{C}\mathbf{B}^{-1}\|_\infty}{1 - \|\mathbf{C}\mathbf{B}^{-1}\|_\infty} \leq \frac{\|\mathbf{C}\|_\infty \|\mathbf{B}^{-1}\|_\infty}{1 - \|\mathbf{C}\|_\infty \|\mathbf{B}^{-1}\|_\infty}.
\end{aligned}$$

We have shown above in (D.5) that $\|\mathbf{C}\|_\infty \|\mathbf{B}^{-1}\|_\infty \leq \frac{3}{8\kappa_\infty} M_0 M_2 |S| \delta \leq 1/2$. As a result,

$$\left\|\nabla^2_{S_2S} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)\left[\left(\nabla^2_{SS} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1} - \left(\nabla^2_{SS} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)\right)^{-1}\right]\right\|_\infty \leq \frac{3}{4\kappa_\infty} M_0 M_2 |S| \delta.$$

By combining these estimates, we have

$$\begin{aligned}
&\left\|\nabla^2_{S_2S} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\left(\nabla^2_{SS} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1} - \nabla^2_{S_2S} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)\left(\nabla^2_{SS} L_n(\mathbf{y}, \mathbf{W}\boldsymbol{\theta}^*)\right)^{-1}\right\|_\infty \\
&\leq \frac{3}{2\kappa_\infty} M_0 M_2 |S| \delta \leq \tau.
\end{aligned}$$

Therefore $\left\|\nabla^2_{S_2S} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\left(\nabla^2_{SS} L_n\left(\mathbf{y}, \overline{\mathbf{W}}\bar{\boldsymbol{\theta}}^*\right)\right)^{-1}\right\|_\infty \leq (1-2\tau) + \tau = 1 - \tau$.

Finally, we use the following lemma to prove the theorem.

We need the following two lemmas to prove Lemma D.6. Define $B_S(\boldsymbol{\theta}^*, r) = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_2 \leq r, \operatorname{supp}(\boldsymbol{\theta}) \subseteq S\}$ for $r > 0$. We first introduce two useful lemmas.

Now we start the proof. First we study the restricted problem

$$\bar{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \mathcal{M}} \{L_n(\boldsymbol{\theta}) + \lambda R(\boldsymbol{\theta})\},$$

where $\mathcal{M} = \{\boldsymbol{\theta} : \boldsymbol{\theta} \in \mathbb{R}^{p+K}, \boldsymbol{\theta}_{S_2} = \mathbf{0}\}$ is the oracle parameter set. Take $R(\boldsymbol{\theta}) = \left\|\boldsymbol{\theta}_{[p]}\right\|_1$ and $R^*(\boldsymbol{\theta}) = \|\boldsymbol{\theta}_{S_2}\|_\infty$. Let $A_1 = \min\left\{A, \frac{\kappa_\infty \tau}{3M}\right\}$ and hence $A_1 \leq \min\left\{A, \frac{\kappa_\infty}{M}\right\}$. Lemma D.3 shows that $\left\|\left(\nabla^2_{SS} L_n(\boldsymbol{\theta})\right)^{-1}\right\|_2 < \kappa_2^{-1}$ and $\left\|\left(\nabla^2_{SS} L_n(\boldsymbol{\theta})\right)^{-1}\right\|_\infty < \kappa_\infty^{-1}$ over $B_S(\boldsymbol{\theta}^*, A_1)$.

Since $\operatorname{supp}(\boldsymbol{\theta}^*) \subseteq S$, any $\mathbf{h} \in \partial R(\boldsymbol{\theta}^*)$ satisfies $\|\mathbf{h}\|_2 \leq \sqrt{|S_1|}$. Therefore

$$\|\nabla_S L_n(\boldsymbol{\theta}^*)\|_2 + \lambda \|\mathbf{h}\|_2 \leq \frac{1}{2}\kappa_2 A_1 \leq \frac{1}{2}\kappa_2 A.$$

Then Lemma D.4 implies that $\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2 \leq \frac{2}{\kappa_2}\left(\left\|\nabla_S L\left(\boldsymbol{\theta}^*\right)\right\|_2+\lambda\sqrt{\left|S_1\right|}\right) \leq A_1$.

Second, we study the L_∞ bound. On the one hand, the optimality condition yields $\nabla_S L_n(\bar{\boldsymbol{\theta}}) \in -\lambda\partial\left\|\bar{\boldsymbol{\theta}}_S\right\|_\infty$ and hence $\left\|\nabla_S L_n(\bar{\boldsymbol{\theta}})\right\|_\infty \leq \lambda$. On the other hand, by letting $\boldsymbol{\theta}_t=(1-t)\boldsymbol{\theta}^*+t\bar{\boldsymbol{\theta}}$ $(0 \leq t \leq 1)$ we have

$$\begin{aligned}\nabla_S L_n(\bar{\boldsymbol{\theta}})-\nabla_S L_n\left(\boldsymbol{\theta}^*\right) &= \int_0^1 \nabla_{SS}^2 L_n\left(\boldsymbol{\theta}_t\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S \mathrm{d}t \\ &= \nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S+\int_0^1\left[\nabla_{SS}^2 L_n\left(\bar{\boldsymbol{\theta}}_t\right)_S-\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right]\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S \mathrm{d}t.\end{aligned}$$

Hence

$$\begin{aligned}&\left\|\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S-\left(\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right)^{-1}\left[\nabla_S L_n(\bar{\boldsymbol{\theta}})-\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right]\right\|_\infty \\ &\leq \int_0^1\left\|\left(\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right)^{-1}\left[\nabla_{SS}^2 L_n\left(\bar{\boldsymbol{\theta}}_t\right)-\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right]\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S\right\|_\infty \mathrm{d}t \\ &\leq\left\|\left(\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty \sup_{t\in[0,1]}\left\|\nabla_{SS}^2 L_n\left(\bar{\boldsymbol{\theta}}_t\right)-\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty\end{aligned}$$

By (i) (ii) (iii) in Lemma D.2, we obtain that

$$\left\|\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S-\left(\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right)^{-1}\left[\nabla_S L_n(\bar{\boldsymbol{\theta}})-\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right]\right\|_\infty \leq \frac{M}{2\kappa_\infty}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty.$$

By $\bar{\boldsymbol{\theta}} \in B_S\left(\boldsymbol{\theta}^*, A_1\right)$ we have

$$\begin{aligned}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty &\leq\left\|\left(\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_\infty\left\|\nabla_S L_n(\bar{\boldsymbol{\theta}})-\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty+\frac{M}{2\kappa_\infty}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty \\ &\leq \frac{1}{2\kappa_\infty}\left(\lambda+\left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty\right)+\frac{1}{6}\left\|\overline{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty.\end{aligned}$$

Therefore,

$$\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty \leq \frac{3}{5\kappa_\infty}\left(\left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty+\lambda\right). \tag{D.7}$$

Third, we study the L_1 bound. The bound on $\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1$ can be obtained in a similar way. Using the fact that $\|\cdot\|_1=\|\cdot\|_\infty$ for symmetric matrices,

$$\begin{aligned}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1 &\leq\left\|\left(\nabla_{SS}^2 L_n\left(\boldsymbol{\theta}^*\right)\right)^{-1}\right\|_1\left\|\nabla_S L_n(\bar{\boldsymbol{\theta}})-\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_1+\frac{M}{2\kappa_\infty}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1 \\ &\leq \frac{1}{2\kappa_\infty}\left(\lambda\left|S_1\right|+\left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_1\right)+\frac{1}{6}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1.\end{aligned}$$

Hence $\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1 \leq \frac{3}{5\kappa_\infty}\left(\left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_1+\lambda\left|S_1\right|\right)$. Since $\operatorname{supp}(\bar{\boldsymbol{\theta}}) \subseteq S$, we also have

$$\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_1 \leq \sqrt{|S|}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2 \leq \frac{2\sqrt{|S|}}{\kappa_2}\left(\left\|\nabla_S L\left(\boldsymbol{\theta}^*\right)\right\|_2+\lambda\sqrt{|S_1|}\right).$$

This gives another L_1 bound.

By Lemma D.5, to derive $\widehat{\boldsymbol{\theta}}=\bar{\boldsymbol{\theta}}$, it remains to show that $\left\|\nabla_{S_2} L_n(\bar{\boldsymbol{\theta}})\right\|_\infty<\lambda$. Using Taylor expansion, we have

$$\begin{aligned}\nabla_{S_2} L_n(\bar{\boldsymbol{\theta}})-\nabla_{S_2} L_n\left(\boldsymbol{\theta}^*\right) &= \int_0^1 \nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}_t\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S \mathrm{d}t \\ &= \nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}^*\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)+\int_0^1\left[\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}_t\right)-\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}^*\right)\right]\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)_S \mathrm{d}t. \end{aligned} \tag{D.8}$$

On the one hand, the first term in (D.8) follows

$$\begin{aligned}\left\|\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}^*\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|_\infty &= \left\|\left[\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}^*\right)\left(\nabla^2_{S S} L_n\left(\boldsymbol{\theta}^*\right)\right)^{-1}\right]\left[\nabla^2_{S S} L_n\left(\boldsymbol{\theta}^*\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right]\right\|_\infty \\ &\leq (1-\tau)\left\|\nabla^2_{S S} L_n\left(\boldsymbol{\theta}^*\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|_\infty.\end{aligned}$$

By the Taylor expansion, triangle's inequality, (i) in Lemma D.2 and the fact that $\bar{\boldsymbol{\theta}} \in B_S\left(\boldsymbol{\theta}^*, A_1\right)$,

$$\begin{aligned}\left\|\nabla^2_{S S} L_n\left(\boldsymbol{\theta}^*\right)\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|_\infty &\leq \left\|\nabla_S L_n(\bar{\boldsymbol{\theta}})-\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty \\ &\quad+\int_0^1\left\|\left[\nabla^2_{S S} L_n\left(\bar{\boldsymbol{\theta}}_t\right)-\nabla^2_{S S} L_n\left(\boldsymbol{\theta}^*\right)\right]\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right)\right\|_\infty \mathrm{d}t \\ &\leq\left\|\nabla_S L_n(\bar{\boldsymbol{\theta}})\right\|_\infty+\left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty+M\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_2\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty \\ &\leq \lambda+\left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty+\frac{\kappa_\infty \tau}{3}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty.\end{aligned}$$

On the other hand, we bound the second term in (D.8). Note that $\boldsymbol{\theta}_t \in B_S\left(\boldsymbol{\theta}^*, A_1\right)$ for all $t \in[0,1]$. (i) in Lemma D.2 yields

$$\begin{aligned}&\left\|\int_0^1\left[\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}_t\right)-\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}^*\right)\right]\left(\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right) \mathrm{d}t\right\|_\infty \\ &\leq \sup_{t \in[0,1]}\left\|\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}_t\right)-\nabla^2_{S_2 S} L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty \leq \frac{\kappa_\infty \tau}{3}\left\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\right\|_\infty.\end{aligned}$$

As a result,

$$\begin{aligned}\left\|\nabla_{S_2} L_n(\bar{\boldsymbol{\theta}})\right\|_\infty &\leq \left\|\nabla_{S_2} L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty + (1-\tau)\left(\lambda + \left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty + \frac{\kappa_\infty \tau}{3}\left\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty\right) \\ &\quad + \frac{\kappa_\infty \tau}{3}\left\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty \\ &\leq \lambda - \tau\left(\lambda - \frac{2\kappa_\infty}{3}\left\|\bar{\boldsymbol{\theta}} - \boldsymbol{\theta}^*\right\|_\infty - \frac{2}{\tau}\left\|\nabla L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty\right).\end{aligned}$$

Recall the L_∞ bound in (D.7). By plugging in this estimate, and using the assumptions $0 < \tau < 1$ and $\lambda > \frac{20}{3\tau}\left\|\nabla L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty$, we derive that

$$\begin{aligned}\left\|\nabla_{S_2} L_n(\bar{\boldsymbol{\theta}})\right\|_\infty &\leq \lambda - \tau\left(\lambda - \frac{2}{5}\left(\left\|\nabla_S L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty + \lambda\right) - \frac{2}{\tau}\left\|\nabla L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty\right) \\ &\leq \lambda - \tau\left(\frac{3}{5}\lambda - \frac{4}{\tau}\left\|\nabla L_n\left(\boldsymbol{\theta}^*\right)\right\|_\infty\right) < \lambda.\end{aligned}$$

This implies $\widehat{\boldsymbol{\theta}} = \bar{\boldsymbol{\theta}}$ and translates all the bounds for $\bar{\boldsymbol{\theta}}$ to the ones for $\widehat{\boldsymbol{\theta}}$. The proposition on sign consistency follows from elementary computation, thus we omit its proof.

E Analysis of Truncation Parameter m_g

Following the assumptions in Lemma 2, we note $\mathrm{Var}(\boldsymbol{a}_{0,i}^{(g)}) = \Sigma_0^{(g)}$. When $\widehat{m}_g > m_g$, we deal with the covariance matrix

$$\mathrm{Var}((\boldsymbol{a}_{0,i}^{(g)\top}, \tilde{\boldsymbol{a}}_{0,i}^{(g)\top})^\top) = \begin{pmatrix} \Sigma_0^{(g)} & \mathbf{0} \\ \mathbf{0} & \tilde{\Lambda}_0^{(g)} \end{pmatrix}.$$

For convenience, we assume we are using $\tilde{\boldsymbol{a}}_i = (\boldsymbol{a}_i^{(1)\top}, \tilde{\boldsymbol{a}}_i^{(1)\top}, \boldsymbol{a}_i^{(2)\top}, \ldots, \boldsymbol{a}_i^{(G)\top})^\top$ that only

$\widehat{m}_1 > m_1$ and others equal. As $\tilde{\boldsymbol{a}}_i$ are calculated from K-L expansion

$$\mathrm{Var}(\tilde{\boldsymbol{a}}_i) = \begin{pmatrix} \Lambda_{m_1} & \mathbf{0} & \Sigma_{m_1 m_2} & \cdots & \Sigma_{m_1 m_G} \\ \mathbf{0} & \tilde{\Lambda}_0^{(1)} & \mathbf{0} & \cdots & \mathbf{0} \\ \Sigma_{m_2 m_1} & \mathbf{0} & \Lambda_{m_2} & \cdots & \Sigma_{m_2 m_G} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Sigma_{m_G m_1} & \mathbf{0} & \Sigma_{m_G m_2} & \cdots & \Lambda_{m_G} \end{pmatrix},$$

If $\xi^\star = (\xi_1^{\star\top}, \xi_2^{\star\top})^\top$ is an one of the first K eigenvector of $\mathrm{Var}(\boldsymbol{a}_i)$, the $(\xi_1^{\star\top}, \mathbf{0}, \xi_2^{\star\top})^\top$ is an eigenvector of $\mathrm{Var}(\tilde{\boldsymbol{a}}_i)$ with the same eigenvalue. The position of $\mathbf{0}$ in $(\xi_1^{\star\top}, \mathbf{0}, \xi_2^{\star\top})^\top$ corresponds to the position of $\tilde{\Lambda}_0^{(1)}$ in $\mathrm{Var}(\tilde{\boldsymbol{a}}_i)$. As we use $\widehat{\mathbf{F}} = \mathbf{A}\widehat{B}\,\mathrm{diag}\left(\lambda_1^{-1} \ldots, \lambda_K^{-1}\right)$ for estimation, $\tilde{\boldsymbol{a}}_i^{(g)}$ will not effect $\widehat{\mathbf{F}}$ since the elements of B corresponding to $\tilde{\boldsymbol{a}}_i^{(g)}$ is 0 in the population distribution when $\widehat{K} = K$.

Another situation worth for discussion is when ω is relatively small, the calculation results tend to underestimate m_g, such as $\widehat{m}_g = K$. Below, we provide an example to illustrate that when $\omega \approx 0$, underestimating m_g will not significantly affect the estimation results of $\widehat{f}_i$.

Setting $G = 2$, the covariance matrix we are faced to is

$$\mathrm{Var}(\boldsymbol{a}_i) = \begin{pmatrix} \Lambda_{m_1} & \Sigma_{m_1 m_2} \\ \Sigma_{m_2 m_1} & \Lambda_{m_2} \end{pmatrix}.$$

Note the j-th eigenvalue of $B^{(g)} B^{(g)\top}$ as $\tau_j^{(g)}$, with $j = 1, \ldots, K$. As $\Lambda_{m_g} = P^{(g)\top} B^{(g)} B^{(g)\top} P^{(g)} +$

ωI, we have $\Lambda_{m_g} = \mathbf{Diag}\{\tau_1^{(g)} + \omega, \ldots, \tau_K^{(g)} + \omega, \omega, \ldots, \omega\}$. We rewrite $\mathrm{Var}(\boldsymbol{a}_i)$ as

$$\mathrm{Var}(\boldsymbol{a}_i) = \begin{pmatrix} \Lambda_{m_{1(K)}} & 0 & \Sigma_{m_{1(K)}m_{2(K)}} & \Sigma_{m_{1(K)}m_{2(C)}} \\ 0 & \Lambda_{m_{1(C)}} & \Sigma_{m_{1(C)}m_{2(K)}} & \Sigma_{m_{1(C)}m_{2(C)}} \\ \Sigma_{m_{2(K)}m_{1(K)}} & \Sigma_{m_{2(K)}m_{1(C)}} & \Lambda_{m_{2(K)}} & 0 \\ \Sigma_{m_{2(C)}m_{1(K)}} & \Sigma_{m_{2(C)}m_{1(C)}} & 0 & \Lambda_{m_{2(C)}} \end{pmatrix} \tag{E.1}$$

$$\approx \begin{pmatrix} \Lambda_{m_{1(K)}} & 0 & \Sigma_{m_{1(K)}m_{2(K)}} & 0 \\ 0 & 0 & 0 & 0 \\ \Sigma_{m_{2(K)}m_{1(K)}} & 0 & \Lambda_{m_{2(K)}} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{E.2}$$

with $\Lambda_{m_g(K)} = \mathbf{Diag}\{\tau_1^{(g)}+\omega, \ldots, \tau_K^{(g)}+\omega\}$, the first K elements and $\Lambda_{m_g(C)} = \mathbf{Diag}\{\omega, \ldots, \omega\}$, the remaining $m_g - K$ elements. So when $\omega \approx 0$, as long as $\widehat{m}_g \geq K$, underestimating m_g will not significantly affect the estimation results of $\widehat{f}_i$.

F Complete Experimental Simulation Results

F.1 Selection Model Size

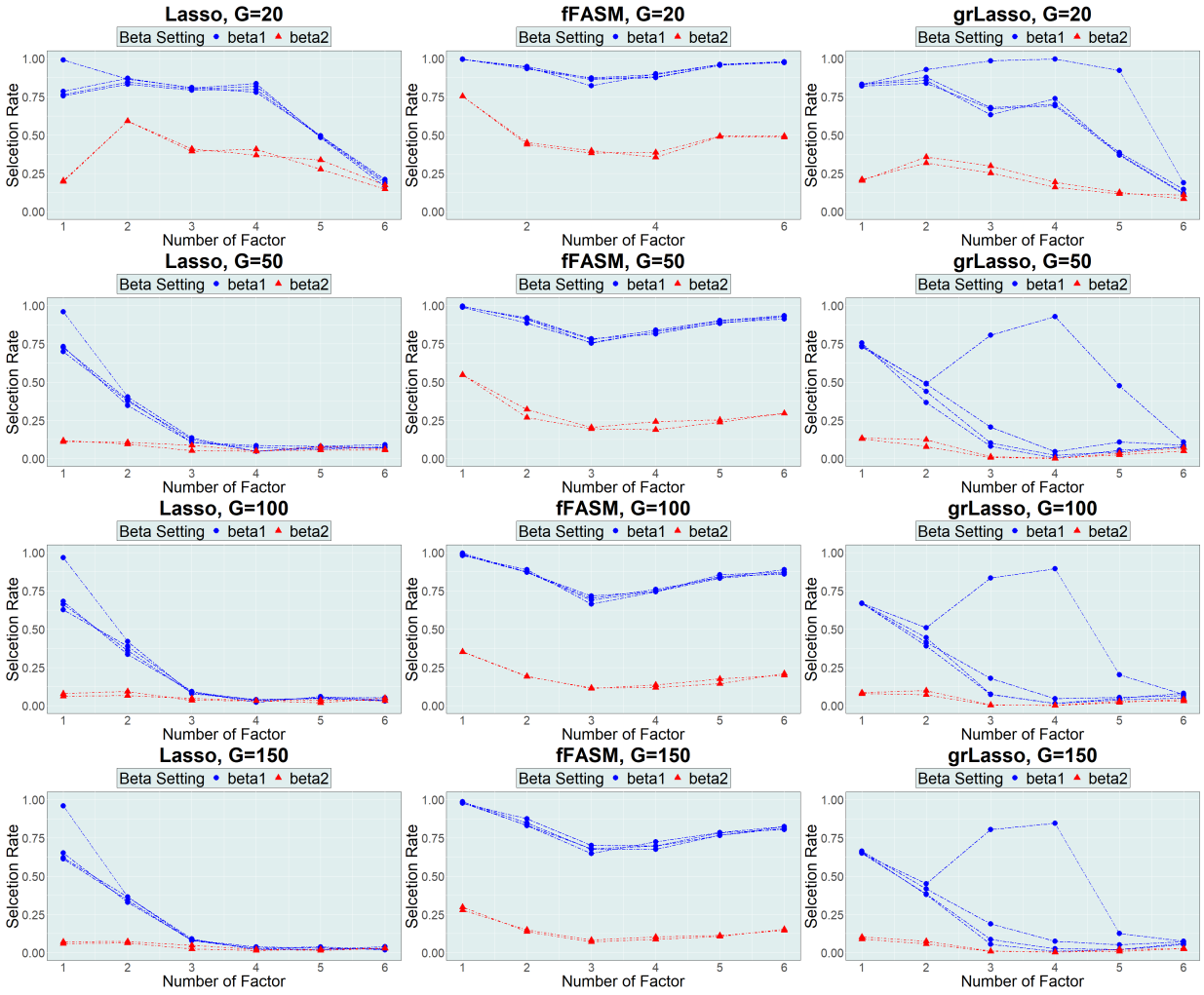

Figure 1: The β selection rate of the three methods vary with the number of factors K. Left, middle, right part are respectively Lasso, fFASM, grLasso; top, middle, bottom part are respectively different $G = 20, 50, 100, 150$.

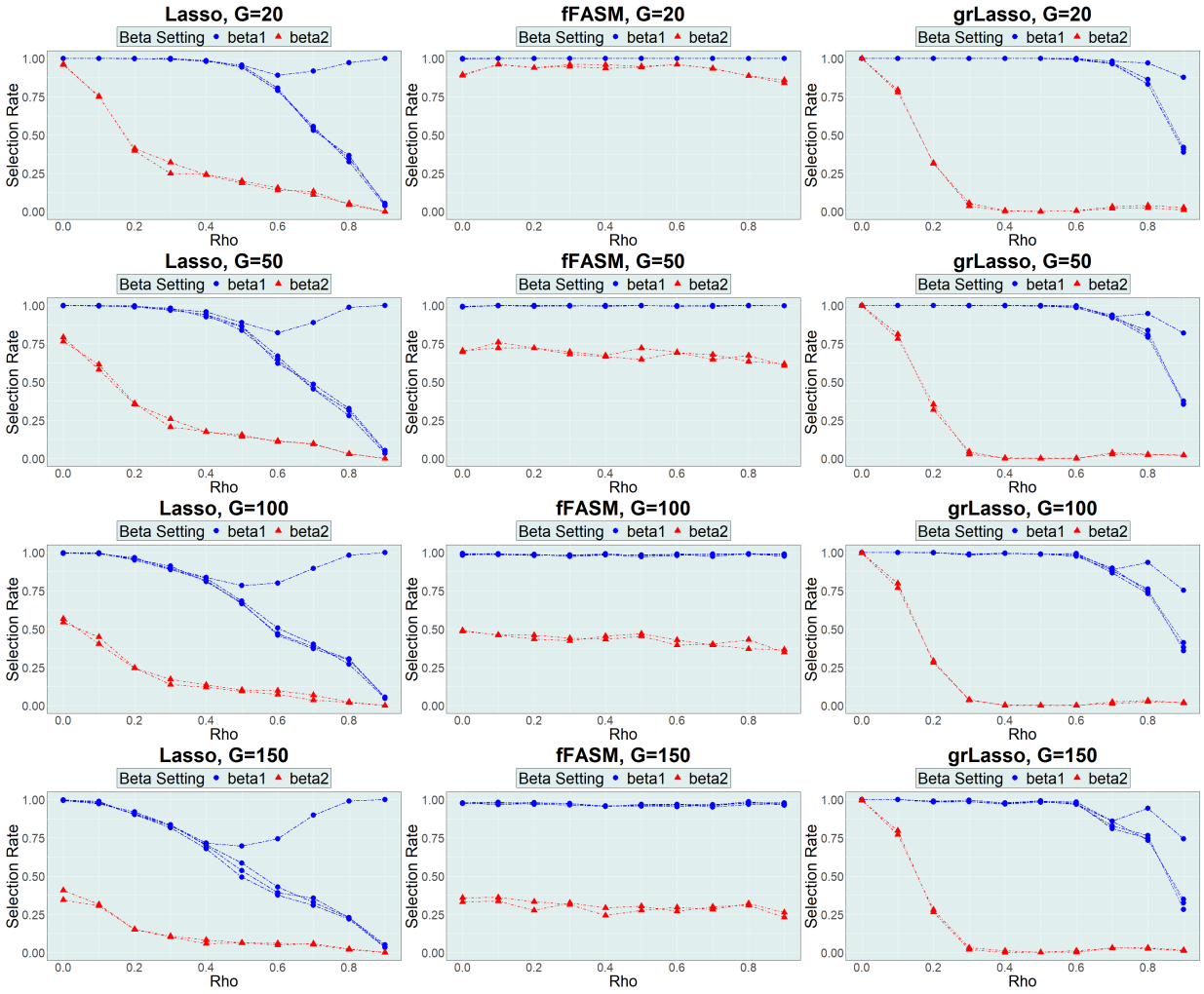

Figure 2: The β selection rate of the three methods vary with the correlation ρ. Left, middle, right part are respectively Lasso, fFASM, grLasso; top, middle, bottom part are respectively different $G = 20, 50, 100, 150$.

F.2 Convergence of Functional Factors

In this section, we present experimental simulations that demonstrate the convergence of the factor $\boldsymbol{f}_i$. Due to certain indeterminate issues, we make the assumption that $K = 4, G = 20, m_g = 10 (1 \leq g \leq G)$ and the elements of $\boldsymbol{f}_i \in \mathbb{R}^K$, $\boldsymbol{u}_i \in \mathbb{R}^{Gm_g}$ are generated from $N(0,1)$, and $N(0,\omega)$, respectively. $\boldsymbol{B} \in \mathbb{R}^{Gm \times K}$. We generate $\boldsymbol{B}$ for factor identifiability, and reconstruct the $\boldsymbol{B}$ matrix based on eigenvalue decreasing for an $N(0,1)$ randomly

generated matrix after performing the singular value decomposition (SVD) operation. Each functional covariate is generated similarly in Section 4 in the manuscript.

Our approach involves measuring the error in factor estimation using the absolute value (abs) metric, as well as employing the **H**-measure as discussed by Fan et al. (2013*a*). We conduct a series of experiments based on the same dataset, varying the settings for N and ω to observe their impact. The absolute measurement error for factors is defined as

$$abs(\widehat{\boldsymbol{f}}_i - \boldsymbol{f}_i) = \sqrt{\sum_{i=1}^{K} ||\widehat{f}_i| - |f_i||^2}.$$

The **H**-measure errors for factors is defined as

$$\|\widehat{\boldsymbol{f}}_i - \mathbf{H}\boldsymbol{f}_i\|_2,$$

where $\mathbf{H} = \frac{1}{T}\mathbf{V}^{-1}\widehat{\mathbf{F}}'\mathbf{F}\mathbf{B}'\mathbf{B}$, $\mathbf{V}$ denote the $\widehat{K} \times \widehat{K}$ diagonal matrix of the first $\widehat{K}$ largest eigenvalues of the sample covariance matrix in decreasing order. Recall that $\mathbf{F}' = (\mathbf{f}_1, \ldots, \mathbf{f}_T)$ and define a $\widehat{K} \times \widehat{K}$ matrix .

The results of our experiments are visualized through several figures, each representing different configurations of n and ω. These visualizations provide a clear depiction of how the estimation error evolves under various conditions, offering insights into the stability and reliability of the factor $\boldsymbol{f}_i$ as the parameters change.

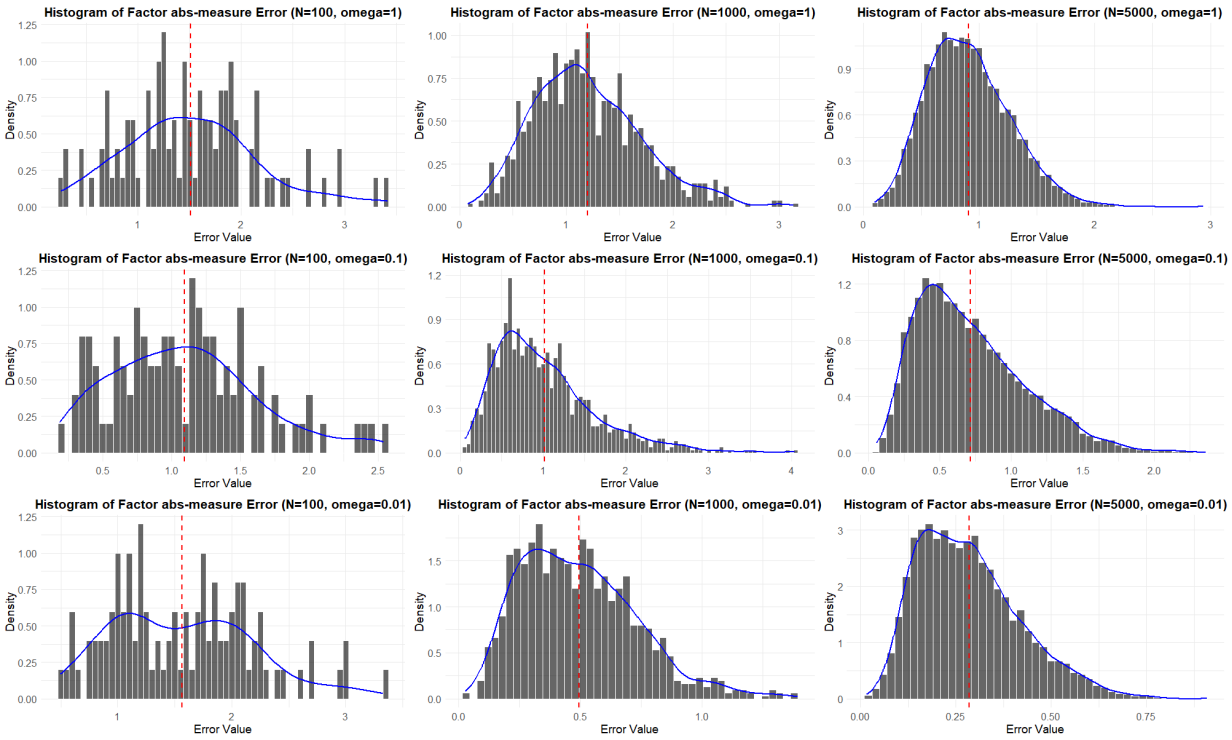

Figure 3: Histogram of absolute measurement errors for factors, with different values of N, ω.

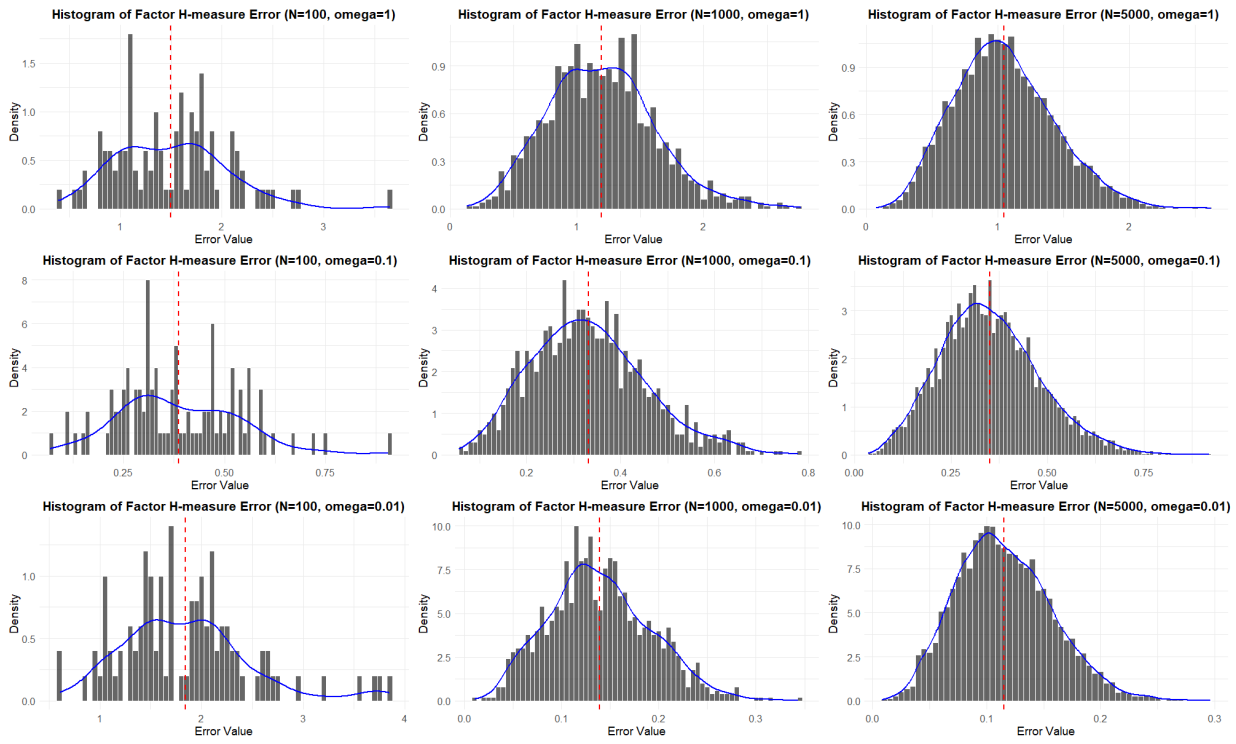

Figure 4: Histogram of **H**-measure errors for factors, with different values of N, ω.

As is shown in Figure 3 and 4 Both types of errors exhibit a pronounced convergence to zero as N approaches infinity and ω approaches zero, indicating a diminishing error magnitude with increasing sample size and decreasing parameter influence.

F.3 Guidelines for Truncation Parameter Selection

Selecting an appropriate truncation parameter (the number of retained functional principal components, m_g) is a fundamental step in functional data analysis. While our theoretical framework establishes that mild over-truncation ($\hat{m}_g > m_g$) does not asymptotically affect the selection consistency and convergence rate (as mathematically bounded by $O_P(n^{-1/2})$ via the Davis-Kahan theorem; see Lemma 4), choosing a robust $\hat{m}_g$ in finite samples remains crucial for practical implementations. To provide clearer guidance for practitioners, this section systematically compares traditional criteria with data-driven spectral methods

under varying noise levels.

We consider five representative criteria for selecting m_g: the traditional Fraction of Variance Explained (FVE) (Ramsay & Silverman 2005), two classical likelihood-based criteria (AIC and BIC), and two spectral data-driven methods—the Eigenvalue Ratio (ER) (Ahn & Horenstein 2013) and Information Criterion (IC) (Bai & Ng 2002).

Table 3 summarizes their mathematical formulations and required hyperparameters. Notably, to ensure strict numerical stability in our fFASM implementation, we adopt a modified ER method. By explicitly incorporating a small positive sequence $C_n = 0.01 \log(\min(n,p))/\min(n,p)$ into both the numerator and denominator, we completely eliminate the computational risk of division by zero when higher-order eigenvalues strictly approach zero, perfectly preserving the asymptotic properties of the original ER. λ_k denotes the k-th eigenvalue, $\mathcal{L}(K)$ is the pseudo-likelihood, V_K is the residual variance, n is the sample size, M is the number of dense observation grid points, $g(n,M)$ represents the penalty function, and C_n is a small positive constant added to ensure numerical stability (preventing division by zero).

Criterion	**Formula / Objective**	**Required Parameter**
FVE (e.g., Ramsay & Silverman 2005)	$\min\left\{K \le K_{\max} : \frac{\sum_{k=1}^{K}\lambda_k}{\sum_j \lambda_j} \ge \tau\right\}$	$K_{\max}, \tau$
AIC (e.g., Akaike 1974, Yao et al. 2005)	$\arg\min_{K \le K_{\max}} \{-2\log \mathcal{L}(K) + 2K\}$	$K_{\max}$
BIC (e.g., Schwarz 1978)	$\arg\min_{K \le K_{\max}} \{-2\log \mathcal{L}(K) + K\log(nM)\}$	$K_{\max}$
ER (e.g., Ahn & Horenstein 2013, Fan et al. 2020)	$\arg\max_{K \le K_{\max}} \frac{\lambda_K + C_n}{\lambda_{K+1} + C_n}$	$K_{\max}, C_n$
IC (e.g., Bai & Ng 2002, Fan et al. 2013*b*)	$\arg\min_{K \le K_{\max}} \{\log(V_K) + K \cdot g(n,M)\}$	$K_{\max}$

Table 3: Summary of truncation criteria and required hyperparameters.

To empirically evaluate the robustness of these criteria, we conduct a numerical experiment by generating functional trajectories contaminated with varying levels of independent

high-frequency white noise $\epsilon_i(t) \sim N(0, \sigma^2_{noise})$. The true underlying truncation parameter is set to $K = 5$, and the maximum search boundary is set to $K_{\max} = 15$.

σ_{noise}	K	**95% FVE**	**ER Method**	**IC Method**	**AIC**	**BIC**
0.1	5	4.90	4.60	4.90	6.55	6.55
0.5	5	5.00	5.00	5.00	15.00	15.00
1.0	5	16.60	5.00	5.00	15.00	15.00
2.0	5	30.60	4.70	4.70	15.00	15.00

Table 4: Estimated truncation parameter ($\hat{K}$) under different noise levels (σ_{noise}). The true m_g is 5, and the maximum search boundary $K_{\max}$ is set to 15.

As shown in Table F.3, with the noise of a small standard deviation, FVE, ER and IC seem reasonably reliable, while traditional likelihood-based criterion such as AIC and BIC may show a slightly over truncation. However, as the standard deviation of the noise increases, FVE seems subject to severe over truncation, and traditional likelihood-based criterion turn to fail dramatically (selecting the maximum bound 15), since their parameter penalty may be less sufficient to suppress the continuous spectrum of high-frequency noise. In contrast, spectral data-driven approaches like ER and IC perfectly capture the "eigengap" and remain highly robust (accurately estimating ~ 5) even under extreme noise contamination.

F.4 Computational Complexity and Scalability

In this section, we provide a detailed computational complexity analysis and empirical runtime evaluation of the proposed fFASM algorithm (Algorithm 1 in the main manuscript) to demonstrate its scalability for extremely high-dimensional functional data.

To formalize the complexity analysis, let n be the sample size, G be the number of functional covariates, M be the number of dense observation grid points per curve, and $p = \sum_{g=1}^{G} m_g$ be the total number of retained functional scores (where $p \ll G \times M$ since m_g is typically small). The computational cost of each step is evaluated as follows:

- **Step 1: Marginal FPCA.** A naive joint FPCA across all functional covariates would require the eigen-decomposition of a dense $(GM) \times (GM)$ covariance matrix. By standard matrix computation theory (Golub & Van Loan 2013), this yields a disastrous computational complexity of $O(n(GM)^2 + (GM)^3)$, which is intractable for thousands of curves. In contrast, the proposed fFASM performs FPCA *marginally* for each group. For a single covariate, a dense FPCA requires $O(nM^2 + M^3)$. Hence, for G covariates, the total complexity is strictly $O(G \cdot (nM^2 + M^3))$. This structure ensures that our method scales perfectly linearly with the number of functional covariates G, and parallel computing is highly feasible in this step.

- **Step 2: Latent Factor Extraction (fFAS).** By extracting factors from the dimension-reduced stacked score matrix $\widehat{\boldsymbol{A}} \in \mathbb{R}^{n \times p}$ rather than the raw dense curves, the complexity of a standard PCA (via Singular Value Decomposition) is drastically reduced to $O(\min(np^2, n^2 p))$ (Golub & Van Loan 2013). Because m_g is intrinsically small, p remains highly manageable even if G is exceptionally large.

- **Step 3: Augmented Variable Selection.** Solving the factor-augmented penalized regression using group coordinate descent algorithms (e.g., for Group SCAD or Group Lasso) is computationally highly efficient. Following the complexity analyses of coordinate descent methods (Friedman et al. 2010, Breheny & Huang 2015), this step requires approximately $O(n(p + \widehat{K}))$ operations per entire iteration loop over all

parameter groups. Operating on the drastically reduced functional subspace makes this convergence extremely fast.

To empirically examine the scalability of fFASM with respect to high-dimensional functional covariates, we conducted runtime comparisons for each step in the algorithm. Simulated data were generated with dense observations and varying numbers of functional covariates, ranging from $G = 100$ to $G = 2000$. As shown in Table 5, the total execution time of the fFASM framework grows linearly with G, confirming its computational efficiency for high-dimensional settings.

Variables (G)	**Step 1 (FPCA)**	**Step 2 (Factor)**	**Step 3 (Penalty)**	**Total Time (s)**
100	2.99	0.01	0.14	3.14
500	14.93	0.05	0.29	15.26
1000	31.58	0.09	0.50	32.17
2000	63.53	0.18	0.96	64.67

Table 5: Empirical time (in seconds) of fFASM steps across varying functional covariate dimensions (G).

F.5 Sensitivity to Penalty Choice and Practical Guidelines

Although the proposed fFASM framework primarily focuses on the Group Lasso penalty to establish consistent theoretical results, the framework is highly flexible and can seamlessly accommodate other penalty functions, such as Group SCAD and Group MCP. In this section, we discuss the theoretical distinctions among these penalties and provide practical guidelines based on empirical sensitivity experiments.

As established in Theorem 2 of the main manuscript, the selection consistency of fFASM under the Group Lasso (L_1) penalty relies on a generalized irrepresentable condition. The

primary advantage of the Group Lasso is its convexity, which guarantees computational stability and a global optimum. However, this L_1 penalty inevitably introduces estimation bias for large non-zero coefficients.

In contrast, non-convex penalties such as Group SCAD and Group MCP can formally relax the stringent irrepresentable condition. By smoothly tapering the penalty for large coefficients, SCAD and MCP shrink the estimation bias, leading to asymptotically unbiased estimates and guaranteeing the desirable "oracle property." Thus, while the fundamental property of selection consistency is robustly preserved across all three choices, they theoretically differ in their trade-offs between computational convexity and estimation bias.

To empirically evaluate the framework's sensitivity to penalty choices, we conducted additional numerical experiments under challenging high-dimensional settings. We compared the performance of fFASM utilizing Group Lasso, Group SCAD, and Group MCP.

Penalty	**IMSE**	**TPR**	**FPR**	**Model Size**
Group Lasso	0.137 (0.007)	0.750 (0.199)	0.154 (0.076)	10.10 (3.58)
Group SCAD	0.154 (0.010)	0.738 (0.222)	0.125 (0.095)	8.70 (4.84)
Group MCP	0.153 (0.009)	0.637 (0.206)	0.026 (0.033)	3.75 (2.12)

Table 6: Empirical performance comparison of different penalty choices within the fFASM framework. Standard errors are provided in parentheses.

As shown in Table 6, all three penalties maintain highly robust predictive performance, confirming the overall stability of the fFASM framework. The Integral Mean Square Error (IMSE) remains stable, ranging tightly from 0.137 to 0.154 across different penalty forms.

Based on the theoretical properties and the empirical behaviors observed in Table 6, we provide the following practical guidelines for practitioners:

- **Group Lasso** is recommended when the primary goal is maximizing sensitivity (True Positive Rate, TPR) and achieving the best predictive accuracy (minimizing IMSE). In our simulations, Lasso achieved the lowest IMSE (0.137) and highest TPR (0.750). However, practitioners should be aware of its tendency to slightly over-select covariates, leading to a higher False Positive Rate (FPR) and a larger model size.

- **Group SCAD and Group MCP** are highly recommended when strict sparsity control, low false discovery rate, and model interpretability are the top priorities. These non-convex penalties greatly reduce the FPR (dropping to 0.026 for MCP) and achieve a much more parsimonious, near-perfect model size, while still retaining a solid TPR and stable overall prediction error.

F.6 Robustness to Sparse and Irregular Sampling

In the main manuscript, our theoretical properties and primary simulations are established under the dense functional data regime. However, in many practical applications, functional trajectories are often observed sparsely and irregularly, contaminated with measurement errors. To address this and evaluate the robustness of the proposed fFASM framework, we conduct additional simulations under a sparse and irregular sampling regime.

Specifically, instead of generating dense observations over 51 regular grid points, we simulate sparse trajectories where the number of observations $n_i^{(g)}$ for the i-th subject and g-th covariate is randomly drawn from a discrete uniform distribution $\mathcal{U}[5, 15]$. The observation time points $t_{ij}^{(g)}$ are randomly and irregularly scattered across the domain $\mathcal{T} = [0, 1]$. Additionally, random measurement errors $\epsilon_i(t_{ij}) \sim N(0, 0.5^2)$ are injected into the observations. We utilize the Principal Analysis by Conditional Estimation (PACE) framework (Yao et al. 2005) to reconstruct the functional curves and extract the functional principal

component scores.

Sampling Regime	IMSE	TPR	FPR	Model Size
Dense (Main Setting)	0.148 (0.012)	0.750 (0.199)	0.154 (0.076)	10.10 (3.58)
Sparse	0.157 (0.015)	0.680 (0.210)	0.165 (0.082)	9.85 (4.12)

Table 7: Empirical performance of fFASM under dense versus sparse and irregular sampling regimes. Standard errors across 100 Monte Carlo repetitions are provided in parentheses.

As summarized in Table 7, transitioning from dense to sparse and irregular sampling inherently leads to a slight loss of information, reflected by a marginal increase in the Integral Mean Square Error (IMSE) and a minor drop in the True Positive Rate (TPR). However, the overall variable selection performance remains remarkably robust, and the model does not suffer from structural breakdown. This empirical robustness corroborates our theoretical design: the PACE framework effectively aggregates cross-sectional information to recover the covariance structure, ensuring that the estimated scores $\widehat{\boldsymbol{A}}$ remain reliable inputs for the subsequent factor augmentation and variable selection stages.

F.7 Solution Path and Selection Stability

To provide a deeper insight into the variable selection mechanism of the proposed fFASM and to address the stability of our selected active sets, we present the solution paths of the estimated functional coefficients.

In the factor-augmented penalized regression stage (Step 3 of Algorithm 1), the Group Lasso penalty is applied to the idiosyncratic components to perform variable selection. Figure 5 illustrates the solution paths of the L_2-norms of the functional coefficients, $\|\widehat{\beta}^{(g)}\|_{L_2}$, as the tuning parameter λ varies.

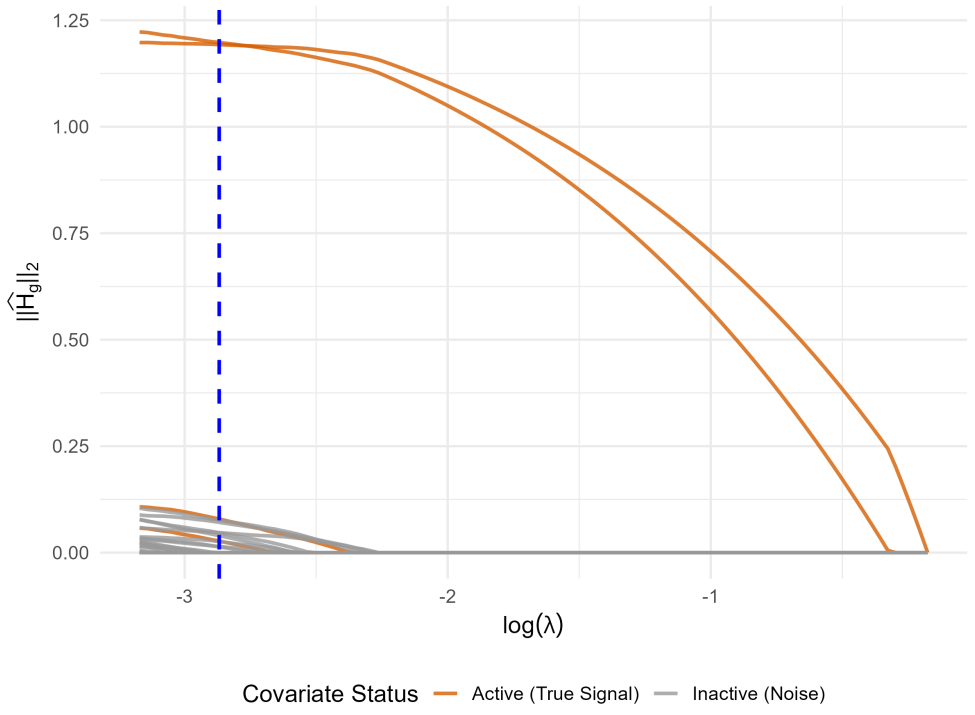

Figure 5: The solution path plot of the selected functional covariates.

Missing functional observations were imputed using the Principal Analysis by Conditional Estimation (PACE) framework (Yao et al. 2005). To ensure equitable penalization, both the scalar response and the extracted functional scores were centered and standardized across all spatial units (countries/provinces) separately prior to analysis, ensuring that the penalty parameter λ is properly scaled for the standardized variance

As shown in Figure 5, the truly active functional covariates enter the model at an early stage (i.e., at larger values of λ) and their coefficient norms remain stable as λ decreases. Conversely, the irrelevant noise covariates are heavily penalized and their norms are strictly shrunk to zero until λ becomes exceedingly small. The optimal λ, selected via cross-validation, accurately isolates the true active set, visually confirming the selection consistency established in Theorem 2.

References

Ahn, S. C. & Horenstein, A. R. (2013), 'Eigenvalue ratio test for the number of factors', *Econometrica* **81**(3), 1203–1227.

Akaike, H. (1974), 'A new look at the statistical model identification', *IEEE transactions on automatic control* **19**(6), 716–723.

Bai, J. & Ng, S. (2002), 'Determining the number of factors in approximate factor models', *Econometrica* **70**(1), 191–221.

Breheny, P. & Huang, J. (2015), 'Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors', *Statistics and computing* **25**(2), 1715–1727.

Fan, J., Ke, Y. & Wang, K. (2020), 'Factor-adjusted regularized model selection', *Journal of Econometrics* **216**(1), 71–85.

Fan, J., Liao, Y. & Mincheva, M. (2013*a*), 'Large covariance estimation by thresholding principal orthogonal complements', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **75**(4), 603–680.

Fan, J., Liao, Y. & Mincheva, M. (2013*b*), 'Large Covariance Estimation by Thresholding Principal Orthogonal Complements', *Journal of the Royal Statistical Society Series B: Statistical Methodology* **75**(4), 603–680.
URL: *https://doi.org/10.1111/rssb.12016*

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of statistical software* **33**(1), 1.

Golub, G. H. & Van Loan, C. F. (2013), *Matrix computations*, 4th edn, JHU press.

Kong, D., Xue, K., Yao, F. & Zhang, H. H. (2016), 'Partially functional linear regression in high dimensions', *Biometrika* **103**(1), 147–159.

Ramsay, J. & Silverman, B. (2005), *Functional Data Analysis*, Springer Series in Statistics, Springer.

Schwarz, G. (1978), 'Estimating the dimension of a model', *The annals of statistics* **6**(2), 461–464.

Yao, F., Müller, H.-G. & Wang, J.-L. (2005), 'Functional linear regression analysis for longitudinal data', *The Annals of Statistics* **33**(6), 2873–2903.