Decentralized Inference for Spatial Data Using Low Rank Model

1 Introduction

We consider the low-rank model of distributed spatial data as follows:

$$z(\boldsymbol{s}_{ji}) = \boldsymbol{x}_{ji}^T\boldsymbol{\beta} + \boldsymbol{b}(\boldsymbol{s}_{ji})^T\boldsymbol{\eta} + \varepsilon(\boldsymbol{s}_{ji}), i = 1, \dots n_j, j = 1, \dots, J$$

where j is the index of machine, i is the index of data point in each machine, $\boldsymbol{s}_{ji}$ is the location, $\boldsymbol{x}_{ji}$ is the covariate vector, $\boldsymbol{b}(\cdot)$ is a known vector-valued function, $\boldsymbol{\eta} \sim N(\boldsymbol{\mu}_0(\boldsymbol{\theta}), K_0(\boldsymbol{\theta}))$ is a Gaussian random vector with known $\boldsymbol{\mu}_0(\cdot), K_0(\cdot)$ and unknown $\boldsymbol{\theta}$, $\varepsilon(\boldsymbol{s}_{ji}) \sim N(0, \sigma^2)$ is a normal random variable whose variance σ^2 is unknown, $z(\boldsymbol{s}_{ji})$ is the observed measurement, $\boldsymbol{\beta}$ is the unknown cofficient vector. A typical example for this model is the predictive process model.

The goal is to make inference for the parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ and prediction at a new location with a covariate vector using data distributed across different machines.

The paper assumes that there is a central machine connected with the rest of machines. Namely, the central machine can communicaiton with all other machines. However, this connection network is vulnerable to the failure of the central machine. Additionally, communication congestion can occur when other machines send information to the central machine.

In this manuscript, we aim to investigate decentralized inference instead. In a decentralized network, the connection of machines forms a decentralized communication network. This approach is less susceptible to the failure of a single machine and can alleviate the problem of communication congestion.

However, there is a significant challenge in making inferences for the parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ by optimizing the log-likelihood. The reason is that the log-likelihood function cannot be represented as a summation of local functions, $\sum_j f_j(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$, with each local function f_j being constructed by the data in each machine j. This is problematic because existing methods in the decentralized optimization community require such a representation.

To tackle this challenge, we borrow the idea from variational inference. Instead of optimizing the log-likelihood directly, we optimize the evidence lower bound (ELBO). The ELBO can be represented as a summation of local functions, allowing existing methods for decentralized optimization to be directly applied.

Contribution The main contribution of this work is developing a decentralized method for this problem.

2 Method

For convenience of demonstration, we let $\boldsymbol{z}_j = (z(\boldsymbol{s}_{j1}), \dots, z(\boldsymbol{s}_{jn_j}))$, $X_j = (\boldsymbol{x}_{j1}, \dots, \boldsymbol{x}_{jn_j})^T$, $B_j = (\boldsymbol{b}(\boldsymbol{s}_{j1}), \dots, \boldsymbol{b}(\boldsymbol{s}_{jn_j}))^T$, $\boldsymbol{\varepsilon}_j = (\varepsilon(\boldsymbol{s}_{j1}), \dots, \varepsilon(\boldsymbol{s}_{jn_j}))$ and $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_J)$, $X = (X_1^T, \dots X_J^T)$, $B = (B_1^T, \dots B_J^T)$, $\boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_J)$. Then, the model can be rewritten as

$$\boldsymbol{z} = X\boldsymbol{\beta} + B\boldsymbol{\eta} + \boldsymbol{\varepsilon}.$$

Let $q(\cdot)$ be a density function, then the evidence lower bound (ELBO) as a funcitonal of q is defined as

$$\mathrm{ELBO}(q) := \int q(\boldsymbol{\eta}) \log \frac{p(\boldsymbol{y}|\boldsymbol{\eta}) p(\boldsymbol{\eta})}{q(\boldsymbol{\eta})} d\boldsymbol{\eta} = \mathbb{E}_q \log p(\boldsymbol{y}|\boldsymbol{\eta}) - \mathrm{KL}(q\|p)$$

According to the Jensen inequality, ELBO is a lower bound of the log-likelihood, that is, ELBO $\leqslant \log p(\boldsymbol{y})$, and the equality can be achieved when $q(\boldsymbol{\eta}) = p(\boldsymbol{\eta}|y)$. Since $p(\boldsymbol{\eta}|y)$ is a Guassian density function,

$$\max_{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}} \log p(\boldsymbol{y}) = \max_{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma} \{\mathbb{E}_{N(\boldsymbol{\eta}|\boldsymbol{\mu}, \Sigma)} \log p(\boldsymbol{y}|\boldsymbol{\eta}) - \mathrm{KL}(N(\boldsymbol{\eta}|\boldsymbol{\mu}, \Sigma)\|p)\},$$

where $N(\boldsymbol{\eta}|\boldsymbol{\mu}, \Sigma)$ represents the density function of a Guasisan vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Further, since $\log p(\boldsymbol{y}|\boldsymbol{\eta}) = \sum_j \log p(\boldsymbol{y}_j|\boldsymbol{\eta})$, we have

$$\max_{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}} \log p(\boldsymbol{y}) = \max_{\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{\mu}, \Sigma} \left\{ \sum_j \mathbb{E}_{N(\boldsymbol{\eta}|\boldsymbol{\mu}, \Sigma)} \log p(\boldsymbol{y}_j|\boldsymbol{\eta}) - \mathrm{KL}(N(\boldsymbol{\eta}|\boldsymbol{\mu}, \Sigma)\|p) \right\}.$$

Then, existing methods for decentralized optimization can be applied.

Notations

For a matrix $\boldsymbol{A}$, let $\|\boldsymbol{A}\|_{\mathrm{op}}$ be its operator norm. For a matrix function $\boldsymbol{A}(\boldsymbol{\theta})\colon \mathbb{R}^q \to \mathbb{R}^{n^2}$, denote $A_{k_1}(\boldsymbol{\theta}) = \frac{\partial \boldsymbol{A}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k}$, $\boldsymbol{A}_{k_1 k_2}(\boldsymbol{\theta}) = \frac{\partial^2 \boldsymbol{A}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{k_1} \partial \boldsymbol{\theta}_{k_2}}$,$\boldsymbol{A}_{k_1 k_2 k_3}(\boldsymbol{\theta}) = \frac{\partial^3 \boldsymbol{A}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{k_1} \partial \boldsymbol{\theta}_{k_2} \partial \boldsymbol{\theta}_{k_3}}$ for $k_1, k_2, k_3 = 1, \ldots., q$. Let $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p)$. Let $\mathbb{B}$,Θ,$\mathbb{S}$ be the parameter space of $\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma$, separately.

Conditions

1. The parameter spaces $\mathbb{B}$,Θ,$\mathbb{S}$ are compact and $\mathbb{S} \in [a, \infty)$ for some positive constant a.

2. $\|\boldsymbol{K}^{-1}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1)$. $\boldsymbol{B}(\boldsymbol{\theta})$ and $\boldsymbol{K}(\boldsymbol{\theta})$ are three times differentiable in each position, and for each $k_1, k_2, k_3 = 1, \ldots, q$,

$$\|\boldsymbol{B}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1), \|\boldsymbol{B}_{k_1}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1), \|\boldsymbol{B}_{k_1 k_2}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1), \|\boldsymbol{B}_{k_1 k_2 k_3}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1) \text{ for all } \boldsymbol{\theta} \in \Theta;$$

$$\|\boldsymbol{K}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1), \|\boldsymbol{K}_k(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1), \|\boldsymbol{K}_{k_1 k_2}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1), \|\boldsymbol{K}_{k_1 k_2 k_3}(\boldsymbol{\theta})\|_{\mathrm{op}} = O(1) \text{ for all } \boldsymbol{\theta} \in \Theta.$$

3. For each $i = 1, \ldots, p$, $\boldsymbol{x}_i$ is a sub-Gaussian vector where $\boldsymbol{x}_i$ is the i-the column of $\boldsymbol{X}$. $\lambda_{\max}\left(\frac{\mathbb{E}(\boldsymbol{X}^{\top}\boldsymbol{X})}{N}\right) = O(1)$.

4. There exist a positive constant such that $\lambda_{\min}(\boldsymbol{\mathcal{J}}) > \lambda^* > 0$. Here, $\boldsymbol{\mathcal{J}}$ is the expected Hessian matrix of negative log-likelihood function at the true point.

Theorem 1. (Convexity) *For any $\varepsilon > 0$, there exist $\delta > 0$ and M such that if $m > M$, the Hessian of the objective function $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$ is positive over the region $\{(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta})\colon \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\| + \|\sigma - \sigma^*\| + \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leqslant \delta, \boldsymbol{\Sigma} \succ 0\}$ with probability greater than $1 - \varepsilon$.*

Theorem 2. *Consistency. As $N \to \infty$, $\boldsymbol{\beta} \xrightarrow{p} \boldsymbol{\beta}^*, \sigma \xrightarrow{p} \sigma^*$. If $m \to \infty$, $\boldsymbol{\theta} \to \boldsymbol{\theta}^*$*

Theorem 3. *Asymptotic normality.*

How about that the low rank model is misspecified.

3 Appendix

3.1 Proofs of main results

Proof of Theorem 1. Denote

$$\nabla^2 f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{pmatrix}$$

where $\boldsymbol{H}_{11}$ corresponds to the Hessian of the parameters $\boldsymbol{\mu}, \boldsymbol{\Sigma}$, and $\boldsymbol{H}_{22}$ corresponds to the the Hessian of the parameters $\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}$. Note that we have the decomposition that

$$\begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{21} & \boldsymbol{H}_{22} \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{H}_{21}\boldsymbol{H}_{11}^{-1} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{H}_{11} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{H}_{22} - \boldsymbol{H}_{21}\boldsymbol{H}_{11}^{-1} H_{12} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{H}_{21}\boldsymbol{H}_{11}^{-1} & \boldsymbol{I} \end{pmatrix}^T$$

whenever $\boldsymbol{H}_{11}$ is positive. Then, it is sufficient to show that H_{11} is positive and the schur complement $\boldsymbol{H}_{22} - \boldsymbol{H}_{21}\boldsymbol{H}_{11}^{-1}\boldsymbol{H}_{12}$ is positive when $\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}$ are near $\boldsymbol{\beta}^*, \sigma^*, \boldsymbol{\theta}^*$ with high probability.

We first show that $\boldsymbol{H}_{11}$ is positive.

Define $f_1(\boldsymbol{\mu}, \boldsymbol{\Sigma})\!: = f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$ and $h(t) := f_1(\boldsymbol{\mu} + t\boldsymbol{\mu}', \boldsymbol{\Sigma} + t\,\boldsymbol{\Sigma}')$ for given $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}'$, a lower triangular matrix, then

$$h''(0) = \begin{pmatrix} \boldsymbol{\mu}'^\top & \mathrm{vec}(\boldsymbol{\Sigma}')^\top \end{pmatrix} \boldsymbol{H}_{11} \begin{pmatrix} \boldsymbol{\mu}' \\ \mathrm{vec}(\boldsymbol{\Sigma}') \end{pmatrix}.$$

Thus, it suffices to show that $h''(0) > 0$ for $\begin{pmatrix} \boldsymbol{\mu}'^\top & \mathrm{vec}(\boldsymbol{\Sigma}')^\top \end{pmatrix}^T \neq \mathbf{0}$. By direct computation,

$$h''(0) \;=\; \mathrm{tr}\left\{\left[\sigma^{-2}\sum_j \boldsymbol{B}_j^\top(\boldsymbol{\theta})\boldsymbol{B}_j(\boldsymbol{\theta}) + \boldsymbol{K}^{-1}(\boldsymbol{\theta})\right]\boldsymbol{\mu}'\boldsymbol{\mu}'^T\right\} + \mathrm{tr}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}').$$

It is obviously greater than zero.

We then show that $\boldsymbol{H}_{11} - \boldsymbol{H}_{12}\boldsymbol{H}_{22}^{-1}\boldsymbol{H}_{21}$ is positive when $\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}$ are near $\boldsymbol{\beta}^*, \sigma^*, \boldsymbol{\theta}^*$ with high probability.

For fixed $\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}$, let $\boldsymbol{\mu}(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$ and $\boldsymbol{\Sigma}(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$ be the minimizer of $f(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$ and define $f_2(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}) = f(\boldsymbol{\mu}(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}), \boldsymbol{\Sigma}(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}), \boldsymbol{\beta}, \sigma, \boldsymbol{\theta})$, then the Hessian of f_2 is exactly $\boldsymbol{H}_{11} - \boldsymbol{H}_{12}\boldsymbol{H}_{22}^{-1}\boldsymbol{H}_{21}$. Then, it suffices to show $\lambda_{\min}[\nabla^2 f_2(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta})] > 0$ when $\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}$ are near $\boldsymbol{\beta}^*, \sigma^*, \boldsymbol{\theta}^*$ with high probability.

$f_2(\boldsymbol{\beta}, \sigma, \boldsymbol{\theta}) = \frac{N}{2}\log 2\pi + \frac{1}{2}\log|\sigma^2\boldsymbol{I} + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{B}^\top(\boldsymbol{\theta})| + (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^T[\sigma^2\boldsymbol{I} + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{B}^\top(\boldsymbol{\theta})]^{-1}(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})$

Let $\boldsymbol{C}(\tilde{\boldsymbol{\theta}}) = \sigma^2\boldsymbol{I} + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{B}^\top(\boldsymbol{\theta})$ where $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}^\top, \sigma)^\top$. Then, for $k = 1, \dots, q+1$,

$$\frac{\partial^2 f_2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} = \boldsymbol{X}^\top\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{X}, \frac{\partial^2 f_2}{\partial\boldsymbol{\beta}\partial\tilde{\boldsymbol{\theta}}_k} = -2\boldsymbol{X}^\top\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})$$

$$\frac{\partial^2 f_2}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l} = \mathrm{tr}\left(\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_l}\right) + \mathrm{tr}\left([\boldsymbol{C}(\tilde{\boldsymbol{\theta}}) - (\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})(\boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta})^\top]\frac{\partial^2\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l}\right)$$

Let $\boldsymbol{u} = \boldsymbol{z} - \boldsymbol{X}\boldsymbol{\beta}^*$, then

$$\frac{\partial^2 f_2}{\partial\boldsymbol{\beta}\partial\tilde{\boldsymbol{\theta}}_k} = -2\boldsymbol{X}^\top\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{u} + -2\boldsymbol{X}^\top\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta})$$

$$\begin{aligned}\frac{\partial^2 f_2}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l} &= \mathrm{tr}\left(\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_l}\right) + \mathrm{tr}\left([\boldsymbol{C}(\tilde{\boldsymbol{\theta}}^*) - \boldsymbol{u}\boldsymbol{u}^\top]\frac{\partial^2\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l}\right) + \\ &\mathrm{tr}\left([\boldsymbol{C}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{C}(\tilde{\boldsymbol{\theta}}^*) + 2\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\boldsymbol{u}^\top + \boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top\boldsymbol{X}^\top]\frac{\partial^2\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l}\right)\end{aligned}$$

Let $N_k = m$ for $k = 1, \dots, q$ and $N_{q+1} = N$, we have

$$\sup_{\boldsymbol{\theta}\in\Theta, \sigma\in\mathbb{S}}\left\|\frac{1}{N_k}\boldsymbol{X}^\top\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{u}\right\| = o_{\mathbb{P}}(1), \tag{1}$$

$$\sup_{\boldsymbol{\theta}\in\Theta, \sigma\in\mathbb{S}}\left\|\frac{1}{N_k}\boldsymbol{X}^\top\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{X}(\boldsymbol{\beta}^* - \boldsymbol{\beta})\right\| = O_{\mathbb{P}}(\|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|) \tag{2}$$

$$\sup_{\boldsymbol{\theta}\in\Theta, \sigma\in\mathbb{S}}\left|\frac{1}{\min\{N_k, N_l\}}\mathrm{tr}\left([\boldsymbol{C}(\tilde{\boldsymbol{\theta}}^*) - \boldsymbol{u}\boldsymbol{u}^\top]\frac{\partial^2\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l}\right)\right| = o_{\mathbb{P}}(1) \tag{3}$$

$$\sup_{\boldsymbol{\theta}\in\Theta, \sigma\in\mathbb{S}}\left|\frac{1}{\min\{N_k, N_l\}}\mathrm{tr}\left([\boldsymbol{C}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{C}(\tilde{\boldsymbol{\theta}}^*)]\frac{\partial^2\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l}\right)\right| = O(\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\|) \tag{4}$$

$$\sup_{\boldsymbol{\theta}\in\Theta, \sigma\in\mathbb{S}}\left|\frac{1}{\min\{N_k, N_l\}}\mathrm{tr}\left(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)\boldsymbol{u}^\top\frac{\partial^2\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l}\right)\right| = O_{\mathbb{P}}(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|) \tag{5}$$

$$\sup_{\boldsymbol{\theta}\in\Theta, \sigma\in\mathbb{S}}\left|\frac{1}{\min\{N_k, N_l\}}\mathrm{tr}\left(\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}^*)(\boldsymbol{\beta} - \boldsymbol{\beta}^*)^\top\boldsymbol{X}^\top\frac{\partial^2\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k\partial\tilde{\boldsymbol{\theta}}_l}\right)\right| = O_{\mathbb{P}}(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2) \tag{6}$$

$$\sup_{\boldsymbol{\theta}\in\Theta, \sigma\in\mathbb{S}}\left\|\frac{1}{N}\boldsymbol{X}^\top\boldsymbol{C}(\tilde{\boldsymbol{\theta}})\boldsymbol{X} - \frac{1}{N}\mathbb{E}\boldsymbol{X}^\top\boldsymbol{C}(\tilde{\boldsymbol{\theta}})\boldsymbol{X}\right\|_{\mathrm{op}} = o_{\mathbb{P}}(1) \tag{7}$$

$$\left\|\frac{1}{N}\mathbb{E}\boldsymbol{X}^\top\boldsymbol{C}(\tilde{\boldsymbol{\theta}})\boldsymbol{X} - \frac{1}{N}\mathbb{E}\boldsymbol{X}^\top\boldsymbol{C}(\tilde{\boldsymbol{\theta}}^*)\boldsymbol{X}\right\|_{\mathrm{op}} = O(\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\|) \tag{8}$$

$$\frac{1}{\min\{N_k, N_l\}}\mathrm{tr}\left(\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_l} - \boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}}^*)\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}}^*)}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}}^*)\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}}^*)}{\partial\tilde{\boldsymbol{\theta}}_l}\right) = O(\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\|) \tag{9}$$

We will postpone the proof of the above equations until the end to enhance readability.

Therefore, the Hessian matrix can be written as

$$\begin{pmatrix} \frac{\partial^2 f_2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \frac{\partial^2 f_2}{\partial\boldsymbol{\beta}\partial\sigma} & \frac{\partial^2 f_2}{\partial\boldsymbol{\beta}\partial\boldsymbol{\theta}^T} \\ \frac{\partial^2 f_2}{\partial\sigma\partial\boldsymbol{\beta}} & \frac{\partial^2 f_2}{\partial\sigma\partial\sigma} & \frac{\partial^2 f_2}{\partial\sigma\partial\boldsymbol{\theta}} \\ \frac{\partial^2 f_2}{\partial\boldsymbol{\theta}^T\partial\boldsymbol{\beta}} & \frac{\partial^2 f_2}{\partial\boldsymbol{\theta}^T\partial\sigma} & \frac{\partial^2 f_2}{\partial\boldsymbol{\theta}^T\partial\sigma} \end{pmatrix} = \begin{pmatrix} N\frac{1}{N}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & N\frac{1}{N}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\sigma\partial\sigma} & m\frac{1}{m}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\sigma\partial\boldsymbol{\theta}} \\ \boldsymbol{O} & m\frac{1}{m}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\boldsymbol{\theta}\partial\sigma} & m\frac{1}{m}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \end{pmatrix} + \begin{pmatrix} N\,\boldsymbol{R}_{11} & N\,\boldsymbol{R}_{12} & m\,\boldsymbol{R}_{13} \\ N\,\boldsymbol{R}_{21} & N\,\boldsymbol{R}_{22} & m\,\boldsymbol{R}_{23} \\ m\,\boldsymbol{R}_{31} & m\,\boldsymbol{R}_{32} & m\,\boldsymbol{R}_{33} \end{pmatrix}$$

where each element of $\boldsymbol{R}_{ij}$ is $o_{\mathbb{P}}(1) + O_{\mathbb{P}}(\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\| + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|)$ for $i, j = 1, 2, 3$. The Hessian matrix can further be written as

$$\boldsymbol{H}_1 + \boldsymbol{H}_2$$

where

$$\boldsymbol{H}_1 = m\begin{pmatrix} \frac{1}{N}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & N\frac{1}{N}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\sigma\partial\sigma} & \frac{1}{m}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\sigma\partial\boldsymbol{\theta}} \\ \boldsymbol{O} & \frac{1}{m}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\boldsymbol{\theta}\partial\sigma} & \frac{1}{m}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} \end{pmatrix} + m\begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \boldsymbol{R}_{13} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} & \boldsymbol{R}_{23} \\ \boldsymbol{R}_{31} & \boldsymbol{R}_{32} & \boldsymbol{R}_{33} \end{pmatrix},$$

and

$$\boldsymbol{H}_2 = (N - m)\begin{pmatrix} \frac{1}{N}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \frac{1}{N}\mathbb{E}\frac{\partial^2 f_2(\boldsymbol{\beta}^*,\tilde{\boldsymbol{\theta}}^*)}{\partial\sigma\partial\sigma} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \end{pmatrix} + (N - m)\begin{pmatrix} \boldsymbol{R}_{11} & \boldsymbol{R}_{12} & \boldsymbol{O} \\ \boldsymbol{R}_{21} & \boldsymbol{R}_{22} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \end{pmatrix}.$$

Therefore, the minimal eigenvalue is lower bounded by

$$m[\lambda^* + o_{\mathbb{P}}(1) + O_{\mathbb{P}}(\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\| + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|)] + (N - m)\|\boldsymbol{x}\|^2[\lambda^* + o_{\mathbb{P}}(1) + O_{\mathbb{P}}(\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\| + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\|)]$$

for some vector $\boldsymbol{x}$ with $\|\boldsymbol{x}\| \leqslant 1$. Then, for each $\varepsilon > 0$, there exist constants δ and M, if $\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\| + \|\boldsymbol{\beta}^* - \boldsymbol{\beta}\| \leqslant \delta$ and $m > M$, then the minimal eigenvalue is lower bounded $m\lambda^*/2$ with probability greater than $1 - \varepsilon$.

Now, we turn to the proof of the equations XX. Here, we will prove only XX, as the proof for the others is similar.

It suffices to show

$$\sup_{\tilde{\boldsymbol{\theta}}} |h_{i,k}(\tilde{\boldsymbol{\theta}})| = O_{\mathbb{P}}(1) \text{ with } h_{i,k}(\tilde{\boldsymbol{\theta}}) = m^{-\frac{1}{2}}\boldsymbol{x}_i^\top \boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial\boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial\tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{u}$$

We use Dudley's entropy integral as stated in the following for the proof.

Definition: A stochastic process $\{X_\theta : \theta \in T\}$ is called **sub-Exponential** with respect to a metric d on T if, for all $\theta, \theta' \in T$ and for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda(X_\theta - X_{\theta'}))] \leq \exp(\lambda^2 d(\theta, \theta')^2) \text{ for } \lambda \leqslant \frac{1}{d(\theta, \theta')}$$

Theorem: Let X_θ be a zero-mean stochastic process that is sub-Gaussian with respect to a pseudo-metric d on the indexing set T. Then

$$\mathbb{E}\left[\sup_{\theta,\theta' \in T} |X_\theta - X_{\theta'}|\right] \leq 8\int_0^D \log(1 + N(\epsilon, T, d))\, d\epsilon$$

where $N(\epsilon, T, d)$ is the covering number of T with respect to the pseudo-metric d and D is the diameter of T.

Note that

$$\frac{\partial \boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}_k} = \boldsymbol{B}_k(\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{B}^\top(\boldsymbol{\theta}) + \boldsymbol{B}_k(\boldsymbol{\theta})\boldsymbol{K}(\boldsymbol{\theta})\boldsymbol{B}_k^\top(\boldsymbol{\theta}) + \boldsymbol{B}(\boldsymbol{\theta})\boldsymbol{K}_k(\boldsymbol{\theta})\boldsymbol{B}^\top(\boldsymbol{\theta}).$$

Then, according to lemma XX in XX and Assumptions XX, it has at most m nonzero bounded singular values. Let $\boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}})\colon= \boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial \boldsymbol{C}(\tilde{\boldsymbol{\theta}})}{\partial \tilde{\boldsymbol{\theta}}_k}\boldsymbol{C}^{-1}(\tilde{\boldsymbol{\theta}})\boldsymbol{C}^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}^*)$, then by singular value decompostion, we have

$$\boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}') = \boldsymbol{U}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\Sigma(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')V^T(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}'),$$

where $\boldsymbol{U}(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}') = \big(\boldsymbol{u}_1(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}'), \ldots, \boldsymbol{u}_N(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)$ and $V(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}') = \big(\boldsymbol{v}_1(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}'), \ldots, \boldsymbol{v}_N(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)$ are two and $\Sigma(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}') = \operatorname{diag}\big(\sigma_1(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}'), \ldots, \sigma_m(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}'), 0, \ldots, 0\big)$. Thus,

$$\begin{aligned} & \mathbb{E}[\exp(\lambda\,(h_{i,k}(\tilde{\boldsymbol{\theta}}) - h_{i,k}(\tilde{\boldsymbol{\theta}}')))] \\ = \; & \mathbb{E}\Big[\exp\Big(\lambda\, m^{-\frac{1}{2}}\boldsymbol{x}_i^\top\big[\boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}')\big]\boldsymbol{e}\Big)\Big] \\ = \; & \mathbb{E}\left[\exp\left(\lambda\, m^{-\frac{1}{2}}\sum_{j=1}^m \boldsymbol{x}_i^\top\boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\sigma_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\boldsymbol{v}_j^T(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\boldsymbol{e}\right)\right]. \end{aligned} \tag{10}$$

Since $\boldsymbol{v}_1^T(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\boldsymbol{e}, \boldsymbol{v}_2^T(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\boldsymbol{e}, \ldots, \boldsymbol{v}_m^T(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\boldsymbol{e}$ are standard Gaussian random vector conditional on $\boldsymbol{X}$, above term further equals to

$$\begin{aligned} & \mathbb{E}\prod_{j=1}^m \mathbb{E}\Big\{\exp\Big[\lambda\, m^{-\frac{1}{2}}\boldsymbol{x}_i^\top\boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\sigma_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\boldsymbol{v}_j^T(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\boldsymbol{e}\Big]\Big|\boldsymbol{X}\Big\} \\ \leqslant \; & \mathbb{E}\exp\left[\lambda^2 m^{-1}\sum_{j=1}^m \big(\boldsymbol{x}_i^\top\boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)^2\sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\right]. \end{aligned} \tag{11}$$

Since, for any $p > 0$,

$$\left\| m^{-1}\sum_{j=1}^m \big(\boldsymbol{x}_i^\top\boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)^2\sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\right\|_{L^p} \leqslant m^{-1}\sum_{j=1}^m \big\|\big(\boldsymbol{x}_i^\top\boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)^2\big\|_{L^p}\sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}'),$$

and $\boldsymbol{x}_i$ is sub-Gaussian, $m^{-1}\sum_{j=1}^m \big(\boldsymbol{x}_i^\top\boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)^2\sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')$ is sub-exponential with paramater proportional to $m^{-1}\sum_{j=1}^m \sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')$. Thus, for some constant $c > 0$ and $\lambda \leqslant c\big(m^{-1}\sum_{j=1}^m \sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)^{-1/2}$,

$$\mathbb{E}\exp\left[\lambda^2 m^{-1}\sum_{j=1}^m \big(\boldsymbol{x}_i^\top\boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big)^2\sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\right] \leqslant \exp\left\{\lambda^2 c^2 m^{-1}\sum_{j=1}^m \sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\right\}. \tag{12}$$

Combining equations XX, we have,

$$\mathbb{E}\left[\exp\left(\lambda\frac{h_{i,k}(\tilde{\boldsymbol{\theta}}) - h_{i,k}(\tilde{\boldsymbol{\theta}}')}{c}\right)\right] \leqslant \exp\big(\lambda^2 d_k^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\big) \text{ for } \lambda \leqslant \frac{1}{d_k(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')}, \tag{13}$$

where

$$d_k(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')\colon= \sqrt{m^{-1}\sum_{j=1}^m \sigma_j^2(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}')} = m^{-\frac{1}{2}}\big\|\boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}')\big\|_F.$$

For fixed $\tilde{\boldsymbol{\theta}}'$, let

$$h(\tilde{\boldsymbol{\theta}}) = \frac{1}{m}\mathrm{tr}\{[\boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}) - \boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}')]^2\},$$

then, by Taylor expansion that $h(\tilde{\boldsymbol{\theta}}) = h(\tilde{\boldsymbol{\theta}}') + (\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}')^\top \nabla h(\tilde{\boldsymbol{\theta}}') + (\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}')^\top \nabla^2 h(\tau\tilde{\boldsymbol{\theta}} + (1-\tau)\tilde{\boldsymbol{\theta}}')(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}')$ for some $0 < \tau < 1$ and $\nabla h(\tilde{\boldsymbol{\theta}}') = 0$,

$$h(\tilde{\boldsymbol{\theta}}) = (\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}')^\top \nabla^2 h(\tau\tilde{\boldsymbol{\theta}} + (1-\tau)\tilde{\boldsymbol{\theta}}')(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}') = O(\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}'\|^2). \tag{14}$$

From XX, we finally have, for some constant $c > 0$,

$$\mathbb{E}\left[\exp\left(\lambda\frac{h_{i,k}(\tilde{\boldsymbol{\theta}}) - h_{i,k}(\tilde{\boldsymbol{\theta}}')}{c}\right)\right] \leqslant \exp\left(\lambda^2\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}'\|^2\right) \text{ for } \lambda \leqslant \frac{1}{\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}'\|}. \tag{15}$$

The covering number $N(\epsilon, \Theta \times \mathbb{S}, \|\cdot\|)$ with respect to the Eculidea distance $\|\cdot\|$ is proportional to $\left(1 + \frac{2D}{\varepsilon}\right)^{q+1}$. Therefore,

$$\mathbb{E}\left[\sup_{\tilde{\boldsymbol{\theta}} \in \Theta \times \mathbb{S}} |h_{i,k}(\tilde{\boldsymbol{\theta}}) - h_{i,k}(\tilde{\boldsymbol{\theta}}')|\right] \leq 8\int_0^D \log(1 + N(\epsilon, \Theta \times \mathbb{S}, \|\cdot\|))\, d\epsilon = O(1)$$

Note that, by singular value decompostion again, $\boldsymbol{M}_{1k}(\tilde{\boldsymbol{\theta}}) = \boldsymbol{U}(\tilde{\boldsymbol{\theta}})\Sigma(\tilde{\boldsymbol{\theta}})V^T(\tilde{\boldsymbol{\theta}})$, where $\boldsymbol{U}(\tilde{\boldsymbol{\theta}}) = (\boldsymbol{u}_1(\tilde{\boldsymbol{\theta}}),..., \boldsymbol{u}_N(\tilde{\boldsymbol{\theta}}))$ and $V(\tilde{\boldsymbol{\theta}}) = (\boldsymbol{v}_1(\tilde{\boldsymbol{\theta}}), \ldots, \boldsymbol{v}_N(\tilde{\boldsymbol{\theta}}))$ are two othogonal matrices, and $\Sigma(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{\theta}}') = \mathrm{diag}\,(\sigma_1(\tilde{\boldsymbol{\theta}},), \ldots, \sigma_m(\tilde{\boldsymbol{\theta}}), 0, \ldots, 0)$. Thus,

$$h_{i,k}(\tilde{\boldsymbol{\theta}}^*) = m^{-\frac{1}{2}}\sum_{j=1}^m \boldsymbol{x}_i^\top \boldsymbol{u}_j(\tilde{\boldsymbol{\theta}}^*)\sigma_j \boldsymbol{v}_j^T \boldsymbol{e}$$

has mean zero and finite variance, which means that

$$\mathbb{E}|h_{i,k}(\tilde{\boldsymbol{\theta}}^*)| = O(1).$$

Finally,

$$\mathbb{E}\left[\sup_{\tilde{\boldsymbol{\theta}} \in \Theta \times \mathbb{S}} |h_{i,k}(\tilde{\boldsymbol{\theta}})|\right] \leqslant \mathbb{E}\left[\sup_{\tilde{\boldsymbol{\theta}} \in \Theta \times \mathbb{S}} |h_{i,k}(\tilde{\boldsymbol{\theta}}) - h_{i,k}(\tilde{\boldsymbol{\theta}}')|\right] + \mathbb{E}|h_{i,k}(\tilde{\boldsymbol{\theta}}^*)| = O(1)$$

□

Proof of Theorem 2. Let $L_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) = \log|C(\tilde{\boldsymbol{\theta}})| + (\boldsymbol{z} - X\boldsymbol{\beta})^T C^{-1}(\tilde{\boldsymbol{\theta}})(\boldsymbol{z} - X\boldsymbol{\beta})$ where $C(\tilde{\boldsymbol{\theta}}) = \sigma^2 I + B(\boldsymbol{\theta})K_0(\boldsymbol{\theta})B^T(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}^T, \sigma)^T$. It is sufficient to prove that: for any given constant $\varepsilon > 0$, if $\|\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}^*\|^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 > \varepsilon^2$, then there exists some constant $\eta > 0$ such that

$$N_m^{-1}[L_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) - L_N(\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\beta}^*)] \geqslant \eta + o_{\mathbb{P}}(1)$$

Let $\mathbb{L}_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) = \mathbb{E}L_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta})$, and note that

$$\begin{aligned} L_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) - L_N(\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\beta}^*) &= \mathbb{L}_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) - \mathbb{L}_N(\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\beta}^*) \\ &\quad + L_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) - L_N(\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\beta}^*) - [\mathbb{L}_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) - \mathbb{L}_N(\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\beta}^*)] \end{aligned}$$

the two terms on the right hand side in above equation.

Firstly, **if**

$$\lambda_{\min}\Big[C^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}^*)C^{-1}(\tilde{\boldsymbol{\theta}})C^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}^*)\Big]\gtrsim c, \lambda_{\max}\Big[C^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}^*)C^{-1}(\tilde{\boldsymbol{\theta}})C^{\frac{1}{2}}(\tilde{\boldsymbol{\theta}}^*)\Big]\lesssim c$$

$$\lambda_{\min}[N^{-1}X^TC^{-1}(\tilde{\boldsymbol{\theta}})X]\gtrsim c$$

$$\mathrm{tr}^2(B(\boldsymbol{\theta})K_0(\boldsymbol{\theta})B^T(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)K_0(\boldsymbol{\theta}^*)B^T(\boldsymbol{\theta}^*))<N_m\|B(\boldsymbol{\theta})K_0(\boldsymbol{\theta})B^T(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)K_0(\boldsymbol{\theta}^*)B^T(\boldsymbol{\theta}^*)\|_F^2$$

$$\inf_{\boldsymbol{\theta}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|>\delta}\frac{1}{m}\|B(\boldsymbol{\theta})K_0(\boldsymbol{\theta})B^T(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)K_0(\boldsymbol{\theta}^*)B^T(\boldsymbol{\theta}^*)\|_F^2\gtrsim c$$

for some positive constants c, then,

$$\begin{aligned}
& \mathbb{L}_N(\tilde{\boldsymbol{\theta}},\boldsymbol{\beta})-\mathbb{L}_N(\tilde{\boldsymbol{\theta}}^*,\boldsymbol{\beta}^*)\\
= \; & \log|C(\tilde{\boldsymbol{\theta}})|-\log|C(\widetilde{\boldsymbol{\theta}^*})|+\mathrm{tr}(C^{-1}(\tilde{\boldsymbol{\theta}})C(\widetilde{\boldsymbol{\theta}^*}))-n+(\boldsymbol{\beta}-\boldsymbol{\beta}^*)^TX^TC^{-1}(\tilde{\boldsymbol{\theta}})X(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\\
\gtrsim \; & \|C(\tilde{\boldsymbol{\theta}})-C(\widetilde{\boldsymbol{\theta}^*})\|_F^2+N\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2\\
= \; & |\sigma^2-(\sigma^*)^2|^2N+[\sigma^2-(\sigma^*)^2]\mathrm{tr}(B(\boldsymbol{\theta})K_0(\boldsymbol{\theta})B^T(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)K_0(\boldsymbol{\theta}^*)B^T(\boldsymbol{\theta}^*))\\
& +\|B(\boldsymbol{\theta})K_0(\boldsymbol{\theta})B^T(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)K_0(\boldsymbol{\theta}^*)B^T(\boldsymbol{\theta}^*)\|_F^2+N\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2\\
\gtrsim \; & (\sigma-\sigma^*)^2N+1(\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|>\delta)N_m+N\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2-|(\sigma-\sigma^*)|O(N_m)
\end{aligned}$$

Let $\varepsilon_1>0$ be the constant such that such that $(\sigma-\sigma^*)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2<\varepsilon_1^2$ implies that $\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|>\delta$ and $A|\sigma-\sigma^*|<\frac{1}{2}$. Then, when $(\sigma-\sigma^*)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2<\varepsilon_1^2$, we have

$$\mathbb{L}_N(\tilde{\boldsymbol{\theta}},\boldsymbol{\beta})-\mathbb{L}_N(\tilde{\boldsymbol{\theta}}^*,\boldsymbol{\beta}^*)\gtrsim[(\sigma-\sigma^*)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2]N+m$$

On the contrary, when $(\sigma-\sigma^*)^2+\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|^2\geqslant\varepsilon_1^2$, we have

$$\mathbb{L}_N(\tilde{\boldsymbol{\theta}},\boldsymbol{\beta})-\mathbb{L}_N(\tilde{\boldsymbol{\theta}}^*,\boldsymbol{\beta}^*)\gtrsim N-O(m)$$

where A is another positive constant.

Second, **if**

$$\sup_{\boldsymbol{\theta}}\frac{1}{m}\mathrm{tr}\left(C^{-1}(\tilde{\boldsymbol{\theta}})\frac{\partial[B(\boldsymbol{\theta})K_0(\boldsymbol{\theta})B^T(\boldsymbol{\theta})]}{\partial\boldsymbol{\theta}_k}C^{-1}(\tilde{\boldsymbol{\theta}})[\boldsymbol{u}\boldsymbol{u}^T-C(\widetilde{\boldsymbol{\theta}^*})]\right)=o_{\mathbb{P}}(1)$$

$$\sup_{\tilde{\boldsymbol{\theta}}}\left|N^{-\frac{1}{2}}\mathrm{tr}(C^{-1}(\tilde{\boldsymbol{\theta}})C^{-1}(\tilde{\boldsymbol{\theta}})[\boldsymbol{u}\boldsymbol{u}^T-C(\widetilde{\boldsymbol{\theta}^*})])\right|=O_{\mathbb{P}}(1)$$

$$\sup_{\tilde{\boldsymbol{\theta}},\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|=1}\left|N^{-\frac{1}{2}}\boldsymbol{u}^TC^{-1}(\tilde{\boldsymbol{\theta}})X\boldsymbol{\alpha}\right|=O_{\mathbb{P}}(1)$$

by the intermediate value theorem,

$$\begin{aligned}
& L_N(\tilde{\boldsymbol{\theta}},\boldsymbol{\beta})-L_N(\tilde{\boldsymbol{\theta}}^*,\boldsymbol{\beta}^*)-[\mathbb{L}_N(\tilde{\boldsymbol{\theta}},\boldsymbol{\beta})-\mathbb{L}_N(\tilde{\boldsymbol{\theta}}^*,\boldsymbol{\beta}^*)]\\
= \; & \sum_k\left[\frac{\partial L_n(\tilde{\boldsymbol{\theta}}^\tau,\boldsymbol{\beta}^\tau)}{\partial\tilde{\boldsymbol{\theta}}_k}-\frac{\partial\mathbb{L}_n(\tilde{\boldsymbol{\theta}}^\tau,\boldsymbol{\beta}^\tau)}{\partial\tilde{\boldsymbol{\theta}}_k}\right](\tilde{\boldsymbol{\theta}}_k-\tilde{\boldsymbol{\theta}}_k^*)+\left[\frac{\partial L_n(\tilde{\boldsymbol{\theta}}^\tau,\boldsymbol{\beta}^\tau)}{\partial\boldsymbol{\beta}}-\frac{\partial\mathbb{L}_n(\tilde{\boldsymbol{\theta}}^\tau,\boldsymbol{\beta}^\tau)}{\partial\boldsymbol{\beta}}\right]^T(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\\
= \; & \sum_k\mathrm{tr}\left(C^{-1}(\tilde{\boldsymbol{\theta}}^\tau)\frac{\partial C(\tilde{\boldsymbol{\theta}}^\tau)}{\partial\tilde{\boldsymbol{\theta}}_k}C^{-1}(\tilde{\boldsymbol{\theta}}^\tau)[\boldsymbol{u}\boldsymbol{u}^T-C(\widetilde{\boldsymbol{\theta}^*})+2\tau X(\boldsymbol{\beta}^*-\boldsymbol{\beta})\boldsymbol{u}^T]\right)(\tilde{\boldsymbol{\theta}}_k-\tilde{\boldsymbol{\theta}}_k^*)\\
& +\boldsymbol{u}^TC^{-1}(\tilde{\boldsymbol{\theta}}^\tau)X(\boldsymbol{\beta}-\boldsymbol{\beta}^*)\\
= \; & \sqrt{N}(\sigma-\sigma^*)O_{\mathbb{P}}(1)+m\sum_k(\boldsymbol{\theta}_k-\boldsymbol{\theta}_k^*)o_{\mathbb{P}}(1)+\sqrt{N}\|\boldsymbol{\beta}-\boldsymbol{\beta}^*\|O_{\mathbb{P}}(1)
\end{aligned}$$

Combine above results, if $(\sigma - \sigma^*)^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 < \varepsilon_1^2$,

$$\begin{aligned}
& L_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) - L_N(\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\beta}^*) \\
\gtrsim \ & [(\sigma - \sigma^*)^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2] N + [(\sigma - \sigma^*) + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|] O_{\mathbb{P}}(\sqrt{N}) + m(1 + o_{\mathbb{P}}(1)) \\
\gtrsim \ & m(1 + o_{\mathbb{P}}(1)) + O_{\mathbb{P}}(1)
\end{aligned}$$

and if $(\sigma - \sigma^*)^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2 \geqslant \varepsilon_1^2$,

$$\begin{aligned}
& L_N(\tilde{\boldsymbol{\theta}}, \boldsymbol{\beta}) - L_N(\tilde{\boldsymbol{\theta}}^*, \boldsymbol{\beta}^*) \\
\gtrsim \ & [(\sigma - \sigma^*)^2 + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2] N + [(\sigma - \sigma^*) + \|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|] O_{\mathbb{P}}(\sqrt{N}) + N_m(1 + o_{\mathbb{P}}(1)) \\
\gtrsim \ & N - O_{\mathbb{P}}(\sqrt{N}) - O(m)
\end{aligned}$$

□

3.2 Technical lemmas and their proofs

For a matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, define $\sigma_i(\boldsymbol{A})$ as its i-th singular value where $\sigma_1(\boldsymbol{A}) \geqslant \sigma_2(\boldsymbol{A}) \geqslant \ldots \geqslant \sigma_n(\boldsymbol{A})$.

Lemma 4. *For two matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n \times n}$,*

$$\sigma_i(\boldsymbol{A}\boldsymbol{B}) \leqslant \sigma_i(\boldsymbol{A}) \sigma_1(\boldsymbol{B}), i = 1, \ldots, n.$$

Lemma 5. *Suppose that (1) $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ is low rank matrix with $\operatorname{rank}(\boldsymbol{M}) \leqslant m$ and $\sigma_i(\boldsymbol{A}) = O(1)$; (2) $\boldsymbol{x} \in \mathbb{R}^n$ is a random vector with $\sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\| = 1} \mathbb{E}(\boldsymbol{\alpha}^\top \boldsymbol{x}_i)^2 = O(1)$; (3)conditional on $\boldsymbol{x}$, $\boldsymbol{e}$ is a standard Gaussian random vector, then*

$$\frac{1}{m} \boldsymbol{x}^\top \boldsymbol{M} \boldsymbol{e} = o_{\mathbb{P}}(1)$$

Ignore $\boldsymbol{\beta}$, let $\boldsymbol{\xi} = H_{f_2}^{1/2} \tilde{\boldsymbol{\theta}}$, define $g(\boldsymbol{\xi}) = f_2\big(H_{f_2}^{-1/2} \boldsymbol{\xi}\big)$, then $\nabla^2 g(\boldsymbol{\xi}) = \nabla^2 f(\tilde{\boldsymbol{\theta}}) H_{f_2}^{-1}$

$$\mathbb{E} \nabla^2 g(\boldsymbol{\xi}) = [\mathbb{E} \nabla^2 f(\tilde{\boldsymbol{\theta}})] \, H_{f_2}^{-1}$$

$$\operatorname{Var}([\nabla^2 g(\boldsymbol{\xi})]_{kl}) =$$