Technical Report RM 710, Department of Statistics and Probability, Michigan State University

Estimating Non-stationary Spatial Covariance Matrix using Multi-resolution Knots

Siddhartha Nandy[1], Chae Young Lim[2] and Tapabrata Maiti[1]
[1] Department of Statistics & Probability
Michigan State University, East Lansing, MI 48824, U.S.A.
Email: nandysid, maiti@stt.msu.edu
[2] Department of Statistics
Seoul National University, Seoul, Korea, 08826
Email: twinwood@snu.ac.kr

Abstract

Providing a **b**est **l**inear **u**nbiased **p**redictor (BLUP) is always a challenge for a non-repetitive, irregularly spaced, spatial data. The estimation process as well as prediction involves inverting an $n \times n$ covariance matrix, which computationally requires $\boldsymbol{O}(n^3)$. Studies showed the potential observed process covariance matrix can be decomposed into two additive matrix components, measurement error and an underlying process which can be non-stationary. The non-stationary component is often assumed to be fixed but low rank. This assumption allows us to write the underlying process as a linear combination of fixed numbers of spatial random effects, known as **f**ixed **r**ank **k**riging (FRK). The benefit of smaller rank has been used to improve the computation time as $\boldsymbol{O}(nr^2)$, where r is the rank of the low rank covariance matrix. In this work we generalize FRK, by rewriting the underlying process as a linear combination of n random effects, although only a few among these are actually responsible to quantify the covariance structure. Further, FRK considers the covariance matrix of the random effect can be represented as product of $r \times r$ cholesky decomposition. The generalization leads us to a $n \times n$ cholesky decomposition and use a group-wise penalized likelihood where each row of the lower triangular matrix is penalized. More precisely, we present a two-step approach using group LASSO type shrinkage estimation technique for estimating the rank of the covariance matrix and finally the matrix itself. We investigate our findings over a set of simulation study and finally apply to a rainfall data obtained on Colorado, US.

Keywords. Covariance estimation; Cholesky decomposition; Block Coordinate Descent Algorithm; Low rank matrix estimation; Group LASSO.

1 Introduction

For most of statistical prediction problems, obtaining a BLUP is very crucial and generally modeling and estimating the mean does the trick. Although estimation of the underlying process covariance is instrumental for spatial BLUP also known as kriging. The concept of kriging was first introduced by D.G.Krige, a South African mining engineer (Cressie, 1990) and Matheron in 1962 coined the term to honor Krige. Kriging is a very popular tool used in earth climate modeling and environmental sciences. It uses quantification of spatial variability through covariance function and solving the standard kriging equation is often numerically cumbersome, and involves inversion of a $n \times n$ covariance matrix. With large n, which is quite reasonable for real data observed on global scale since computation cost increases with cubic power of the dimension n, spatial BLUP becomes challenging.

Hence, there have been several efforts to achieve a computationally feasible estimate. The foremost challenge of estimating covariance for a spatial set up arises due to absence of repetition. This may seem absurd if we realize this situation as a multivariate extension of computing variance from one observation. As odd as may it sound, the trick is to consider a specific sparsity structure for the covariance matrix under study. The covariance matrix is sparse when the covariance function is of finite range and due to sparsity the computation cost to invert a $n \times n$ matrix reduces considerably.

Before we delve in to the discussion of our contribution we would like to put forward a few other attempts to estimate large covariance matrices through literature review. In 1997 Barry and Pace used symmetric minimum degree algorithm when $n = 916$ for kriging. Rue and Tjelmeland (2002) approximated $\mathbf{\Sigma}^{-1}$ to be sparse precision matrix of a Gaussian Markov random field wrapped on a torus. For larger n, the first challenge in applying kriging is, increase in condition number of the covariance matrix, which plays a major role in building up the computation time and makes the kriging equation numerically unstable. On the other hand, to handle computational complexity, Kaufman $et.al.$ (2008), introduced the idea of covariance tapering which sparsify the covariance matrix element wise to approximate the likelihood. Some other worth mentioning efforts in tapering are Furrer $et.al.$ (2012), Stein (2013) e.t.c. Covariance tapering gains immense computational stability, keep interpolating property and also have asymptotic convergence of the taper estimator. But tapering is restricted only to isotropic covariance structure and the tapering radius needs to be determined.

Another alternative method, FRK was introduced by Cressie & Johannesson (2008). Unlike tapering, FRK is applicable to a more flexible class of non-stationary covariance matrix, and also reduces computational cost of kriging to $\boldsymbol{O}(n)$. For many non-stationary covariance model like ours, the observed process covariance matrix can be decomposed into two additive matrix components. The first is a measurement error modeled as white noise. While the second is an underlying process which can be non-stationary covariance structure and is often assumed to be fixed but low rank. The underlying process can be represented as a linear combination of r_n random effects. For FRK r_n plays the role of rank of the non-stationary component, is considered to be known r and fixed over n. In this work we would like to relax this assumption by allowing r_n changing over n.

Our goal in this paper is to achieve a data driven approach for finding the rank r_n. To do so let us assume even though there are unknown r_n random effects used to represent the underlying process, what if we start with some numbers of random effects and as we proceed, our algorithm will direct us toward a data driven value for r_n? Once we figure out that the dispersion matrix of this n dimensional random effect can be decomposed into cholesky factor, a closer look will teach us that dropping or selecting a particular random effect boils down to zero or non-zero row in the corresponding cholesky matrix. We consider a penalized likelihood approach where we penalize ℓ_2- norm within each row of the cholesky matrix and ℓ_1- norm between two different rows of the cholesky matrix.

The low rank non-stationary covariance matrix is decomposed, using a basis components (not necessarily orthogonal) and another component is dispersion matrix of random effects vector. The basis component depends primarily on the choice of the class of basis function and number of knot points. FRK recommends that the choice of basis function should be multi-resolutional, more precisely they used a local bi-square functions. This use of locally multi-resolutional knots has also been proved quite useful in the literature of kriging for large spatial data sets (Nychka (2015)) other than FRK. The choice of number of knot points and their positions is always crucial. The number of knot points is directly related to r_n, the number of random effects component. The foremost challenge in applying our method is choice of effective numbers of knot points necessary to construct the basis function under study.

Although our initial objective in this work is to provide a way to estimate the non-zero random effects and finally the covariance matrix, but like any other statistical prediction problem we shall be extending our findings in presence of covariates. Peng and Wu (2010), proved that condition number of the covariance matrix also increases with increase in input variables. To handle numerical instability, Peng and Wu (2010), suggested the idea of regularized kriging, which is a simple modification in the method of estimation. Unlike kriging, regularized kriging optimizes regularized or penalized likelihood. At this stage we have not considered dimension reduction challenges while extending our findings in presence of covariates but, for future studies, this can be a non-trivial and worthwhile extension.

A recent study on limitations of low rank kriging (Stein (2015)) shows an approximation in which observations are split into contiguous blocks and assumes independence across these blocks. It provides a much better approximation to the data likelihood than a low rank approximation requiring similar memory and calculations. It also shows that Kullback-Leibler divergence for low rank approximation is not reduced as much as it should have been in few settings. On the contrary the divergence is considerably reduced if there is a block structure. Keeping this in mind, and considering the fact that selections of knots work better under multiresolution setup, we consider the knots by superimposing resolutions.

Under some sensible assumptions this paper will motivate our readers to the idea of existence of a consistent covariance estimator of the spatial process using a low rank modeling, whose estimation has not been discussed before in any literature to the best of our knowledge. We will discuss the practical implications of our assumption later but, we still like to point out that without loss of generality we considered, the location knots for the bi-variate spline matrix are ordered in a specific way such that the true structure has the first r_n non-zero rows and rest $n - r_n$ zero rows. We also discuss how our findings fit in the situations of limitations of low rank kriging (Stein (2015)). To avoid further mathematical details here, this part of the comparison is in discussion section 6.

All kinds of approximation of the covariance function introduced so far, has a motive to reduce the computational cost. Most of these existing methods fail to capture both large scale (long-range) and small scale (short-range) dependence. However tapering captures small scale dependence and, low rank techniques captures large scale dependence. A new method is discussed using adding these two components (Sang and Huang 2012). We would like to point out our readers that however worthwhile this method of combining both low rank and tapering may look, this paper provides a more sound theoretical approach to support our algorithm and findings. Although estimation of low rank covariance matrix has it's limitations, the method has not always been criticized, rather well established in several situations by various authors. Most of the interesting work in this field, can be classified in two broad classes: statistics and machine learning. Among many others in the field of statistics we think, Fan and Li (2012), Banerjee $et.al.$ (2012), Tzeng and Huang (2015) e.t.c. are worth mentioning. On other the hand, the field of machine learning focuses on developing algorithms where, Frieze $et.al.$ (2004), Achlioptas and McSherry (2007), Journ$\acute{e}$e $et.al.$ (2010) are quite reasonable to browse through. Based on these literatures it is obviously worthwhile to contribute our time and to come up with a theoretical justification behind the possibility of low rank covariance matrix estimation.

Even when we keep the rank fixed, for a very large data set (order of tens of thousands to hundreds of thousands), kriging can be quite impossible and ad hoc local kriging neighborhoods are used (Cressie (1993)). Some recent developments include Nychka $et.al.$ (1996; 2002), Furrer $et.al.$ (2006) and many more. Among other alternative methods, some worth discussing are Radial basis interpolation functions (B$\ddot{u}$hlmann, (2004)), inverse distance weighting (Shepard, (1968)) or regression-based inverse distance weighting used by Joshep and Kang (2009) which is a fast interpolator and overcomplete basis surrogate method (Chen, Wang, and Wu (2010)). Surrogate basis representation is similar to lattice kriging (Nychka (2015)) where the basis functions are boundedly supported and over complete. But lattice kriging considers sparsity in the precision matrix through bounded basis function matrix and a parametric neighborhood matrix whereas we are considering sparsity in the covariance matrix through low rank factorization and cholesky decomposition of the low rank covariance matrix.

The rest of this paper is organized as follows. In Section 2, we explain the proposed approach for selecting and estimating nonzero rows (rank) and the corresponding low rank covariance matrix. In Section 3, we discuss the main theoretical results on the selection and estimation consistency. Following which in section 4 we present the block coordinate descent algorithm for block wise convex regularizing functions. Section 5 contains simulation results along with a real data example. Finally, we make some concluding remarks in section 6. Proofs of all theorems and lemmas are provided in the appendix.

2 Group Lasso for estimating low rank covariance matrix

The concept of using ℓ_1/ℓ_2 - penalty (Yuan $et.al.$ (2006), and Bühlmann $et.al.$ (2011)) component had been well established in the context of selecting varibles, if it is believed, that there exist an inherent group structure in the parameter space. But using this has not been quite clear in estimating rank of a low-rank matrix and estimating the matrix itself. Here in this section we want to present an ℓ_1/ℓ_2 - penalized approach in estimating the low rank non-stationary covariance matrix as an extension of FRK. The goal of FRK is to reduce computation cost of inversion of a matrix from cubic to linear in sample size. To explain the crucial difference between FRK and our method, low rank kriging, we need to introduce the following mathematical notations.

Consider $\boldsymbol{Y} = \{Y(\boldsymbol{s}); \boldsymbol{s} \in \mathscr{S}\}$ be a spatial process perturbed with measurement error $\boldsymbol{\epsilon} = \{\epsilon(\boldsymbol{s}); \boldsymbol{s} \in \mathscr{S}\}$ and let $\boldsymbol{X} = \{X(\boldsymbol{s}); \boldsymbol{s} \in \mathscr{S}\}$ be the process of potential observation where, $\boldsymbol{\epsilon}$ is a Gaussian process with mean 0 and $var\left(\epsilon(\boldsymbol{s})\right) = \sigma^2 v(\boldsymbol{s}) \in (0, \infty), \boldsymbol{s} \in \mathscr{S}$, for $\sigma^2 > 0$ and $v(\cdot)$ known. Now in general the underlying process $\boldsymbol{Y}$ has a mean structure, $Y(\boldsymbol{s}) = Z(\boldsymbol{s})'\boldsymbol{\beta} + \pi(\boldsymbol{s})$, for all $\boldsymbol{s} \in \mathscr{S}$ where, $\boldsymbol{\pi} = \{\pi(\boldsymbol{s}); \boldsymbol{s} \in \mathscr{S}\}$ follows a Gaussian distribution with mean 0, $0 < var\left(\pi(\boldsymbol{s})\right) < \infty$, for all $\boldsymbol{s} \in \mathscr{S}$, and a non-stationary spatial covariance function $cov\left(\pi(\boldsymbol{s}), \pi(\boldsymbol{s}')\right) = \sigma(\boldsymbol{s}, \boldsymbol{s}')$, for all $\boldsymbol{s}, \boldsymbol{s}' \in \mathscr{S}$. Also $\boldsymbol{Z} = \{Z(\boldsymbol{s}); \boldsymbol{s} \in \mathscr{S}\}$ represent known covariates and $\boldsymbol{\beta}$ be the vector of unknown coefficients. So, finally combining the underlying process and the measurement error we have,

$$X(\boldsymbol{s}) = Z(\boldsymbol{s})'\boldsymbol{\beta} + \pi(\boldsymbol{s}) + \epsilon(\boldsymbol{s}) \qquad\qquad \forall \boldsymbol{s} \in \mathscr{S}. \tag{1}$$

The process $X(\cdot)$ is observed only at a finite number of spatial locations $\boldsymbol{S}_n = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n\} \subset \mathscr{S}$. We allow $\boldsymbol{S}_n$ to be any irregular lattice in d-dimension with cardinality n. Now, we can go back explaining FRK. In general, the covariance function $\boldsymbol{\sigma}(\boldsymbol{s}, \boldsymbol{s}')$ has to be a positive definite function on $\mathbb{R}^n \times \mathbb{R}^n$. In practice $\boldsymbol{\Sigma}$ is unknown and being in a spatial sampling design we often consider $\boldsymbol{\sigma}(\boldsymbol{s}, \boldsymbol{s}')$ as stationary covariance function, but in this paper we want to keep it general and allow the possibility of it being non-stationary. We capture the spatial information through basis functions [*Cressie, N. et al. (2008)*],

$$\boldsymbol{R}(\boldsymbol{s}) = (R_1(\boldsymbol{s}), R_2(\boldsymbol{s}), \ldots, R_{r_n}(\boldsymbol{s}))', \qquad\qquad \forall \boldsymbol{s} \in \boldsymbol{S}_n$$

and for a positive definite matrix $\boldsymbol{\Omega}$, we have a model for our covariance function $\boldsymbol{\sigma}(\boldsymbol{s}, \boldsymbol{s}')$ as,

$$\boldsymbol{\sigma}(\boldsymbol{s}, \boldsymbol{s}') = \boldsymbol{R}(\boldsymbol{s})'\boldsymbol{\Omega}\boldsymbol{R}(\boldsymbol{s}'), \qquad\qquad \forall \boldsymbol{s}, \boldsymbol{s}' \in \boldsymbol{S}_n. \tag{2}$$

It is quite interesting to observe that above is a consequence of writing $\pi(\boldsymbol{s}) = \boldsymbol{R}(\boldsymbol{s})'\boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ is an r_n-dimensional vector with $var(\boldsymbol{\alpha}) = \boldsymbol{\Omega}$. The model for $\pi(\cdot)$ is often refered to as *spatial random-effects* model. Define matrix $\boldsymbol{R}$ with $\boldsymbol{R}(\boldsymbol{s}_i)'$ as the i^{th} row and correspondingly, $\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{\Omega}\boldsymbol{R}'$, where $\mathscr{R}(\boldsymbol{\Sigma}) \leq \mathscr{R}(\boldsymbol{\Omega}) = r_n$, where $\mathscr{R}(\cdot)$ is used to denote rank of a matrix. In practice since we do not have prior knowledge about the value of r_n, our contribution is to provide an estimate of the rank parameter r_n, while estimating the matrix itself as we start with L basis for each location sites,

$$\widetilde{\boldsymbol{R}}(\boldsymbol{s}) = \left(\widetilde{R}_1(\boldsymbol{s}), \widetilde{R}_2(\boldsymbol{s}), \ldots, \widetilde{R}_L(\boldsymbol{s})\right)', \qquad \forall \boldsymbol{s} \in \boldsymbol{S}_n.$$

In an ideal scenario we select the first r_n basis rows and, by dropping the rest we obtain $\boldsymbol{R}$. So we start with a model for our covariance function $\boldsymbol{\sigma}(\boldsymbol{s}, \boldsymbol{s}')$ as,

$$\boldsymbol{\sigma}(\boldsymbol{s}, \boldsymbol{s}') = \widetilde{\boldsymbol{R}}(\boldsymbol{s})'\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{R}}(\boldsymbol{s}'), \qquad \forall \boldsymbol{s}, \boldsymbol{s}' \in \boldsymbol{S}_n \tag{3}$$

Similar to equation (2), one can easily see that equation (3) is a consequence of writing $\pi(\boldsymbol{s}) = \widetilde{\boldsymbol{R}}(\boldsymbol{s})'\widetilde{\boldsymbol{\alpha}}$, where $\widetilde{\boldsymbol{\alpha}}$ is an $L-$dimensional vector with $var\ (\widetilde{\boldsymbol{\alpha}}) = \widetilde{\boldsymbol{\Omega}}$. Using this expression of random effect $\boldsymbol{\pi}(\boldsymbol{s})$, in (1) we get,

$$X(\boldsymbol{s}) = Z(\boldsymbol{s})'\boldsymbol{\beta} + \widetilde{\boldsymbol{R}}(\boldsymbol{s})'\widetilde{\boldsymbol{\alpha}} + \epsilon(\boldsymbol{s}) \qquad \forall \boldsymbol{s} \in \mathscr{S}. \tag{4}$$

Also for simplicity let us first present our method for the case $\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{0}$. Let us now explain the relation between two versions of random effects or covariance model and, the method used to reduce this dimensionality cost. Ideally, $\boldsymbol{\Omega}$ is a sub-matrix of $\widetilde{\boldsymbol{\Omega}}$ with $\mathscr{R}\left(\widetilde{\boldsymbol{\Omega}}\right) = \mathscr{R}(\boldsymbol{\Omega})$ such that,

$$\begin{pmatrix} \boldsymbol{\Omega} & \mathbb{O}_{r_n \times (L-r_n)} \\ \mathbb{O}_{(L-r_n)\times r_n} & \mathbb{O}_{(L-r_n)\times(L-r_n)} \end{pmatrix} = \widetilde{\boldsymbol{\Omega}} = \widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}' = \begin{pmatrix} \boldsymbol{\Phi}\boldsymbol{\Phi}' & \mathbb{O}_{r_n \times (L-r_n)} \\ \mathbb{O}_{(L-r_n)\times r_n} & \mathbb{O}_{(L-r_n)\times(L-r_n)} \end{pmatrix}, \tag{5}$$

where, $\boldsymbol{\Phi}_{r_n \times r_n}$ is the cholesky decomposition of $\boldsymbol{\Omega}$, an unique lower triangular matrix. In practice, it may be necessary that we reorder the columns in our basis matrix $\widetilde{\boldsymbol{R}}$ to achieve the above structure. This reordering can be taken care by introducing a permutation matrix, explained in the appendix. So, for rest of the discussion we will consider $\boldsymbol{\Sigma} = \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Omega}}\widetilde{\boldsymbol{R}}'$. As mentioned earlier due to (5), $\boldsymbol{\Phi}_{r_n \times r_n}$ is principal sub-matrix of $\widetilde{\boldsymbol{\Phi}}_{n \times n}$ but since, we have limited knowledge about the value of r_n we propose to start with all L rows non-zero. Our proposed method allows us to select non-zero rows of $\widetilde{\boldsymbol{\Phi}}$, which eventually captures all information required to retreive $\boldsymbol{\Sigma}$. We drop a row from $\widetilde{\boldsymbol{\Phi}}$ if only if, all the elements in that row are smaller than some preset value. Hence a group wise penalization is sensible, as the shrinkage equation will have a similar nature of block-wise optimization. Denote $\widetilde{\boldsymbol{\varphi}}'_{(j)} = (\widetilde{\varphi}_{j1}, \widetilde{\varphi}_{j2}, \ldots, \widetilde{\varphi}_{jj}, 0, 0, \ldots, 0) = (\widetilde{\boldsymbol{\varphi}}'_j, 0, 0, \ldots, 0)$ to be the j^{th} row of $\widetilde{\boldsymbol{\Phi}}$, where number of zeros in j^{th} row is $L - j$.

Define, $\widetilde{\boldsymbol{\Phi}}^{vec}_{Fullset} = (\widetilde{\boldsymbol{\varphi}}'_1, \widetilde{\boldsymbol{\varphi}}'_2, \ldots, \widetilde{\boldsymbol{\varphi}}'_L)$ to be a row-wise vector representation of the lower triangular part of the matrix $\widetilde{\boldsymbol{\Phi}}$. For a weight vector $\boldsymbol{\psi} = (\psi_1, \psi_2, \ldots, \psi_L)'$, we define a weighted

ℓ_1/ℓ_2-norm, $\left\|\widetilde{\boldsymbol{\Phi}}^{vec}_{Fullset}\right\|_{2,1,\boldsymbol{\psi}} = \sum_{j=1}^{L}\psi_j\|\widetilde{\boldsymbol{\varphi}}_{(j)}\|_2$, where $\|\cdot\|_2$ is the ℓ_2-norm of a vector. So, we propose the following weighted ℓ_1/ℓ_2-penalized likelihood function,

$$\boldsymbol{Q}_n(\widetilde{\boldsymbol{\Phi}},\sigma^2,\tau_n,\boldsymbol{\psi}_n) = \boldsymbol{X}'\boldsymbol{\Xi}^{-1}\boldsymbol{X} + \log\det\boldsymbol{\Xi} + \tau_n\left\|\widetilde{\boldsymbol{\Phi}}^{vec}_{Fullset}\right\|_{2,1,\boldsymbol{\psi}_n}, \tag{6}$$

where τ_n is the regularization parameter, $\boldsymbol{\psi}_n = (\psi_{n1},\psi_{n2},\ldots,\psi_{nn})'$ is a suitable choice of a weight vector in the penalty term. We allow the possibility that the penalty parameter, τ_n, and the weight vector, $\boldsymbol{\psi}_n$, can depend on the sample size n. Now using the above covarince modeling for $\boldsymbol{\Sigma}$ *i.e.* $\boldsymbol{\Sigma} = \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'$ and using, $\boldsymbol{\Xi}^{-1} = \left(\sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'\right)^{-1}$, (6) can be rewritten as,

$$\begin{aligned}\boldsymbol{Q}_n(\widetilde{\boldsymbol{\Phi}},\sigma^2,\tau_n,\boldsymbol{\psi}_n) = \mathbf{Tr}\left(\boldsymbol{\Xi}_0\left(\sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'\right)^{-1}\right) &+ \log\det\left(\sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'\right)\\ &+ \tau_n\left\|\widetilde{\boldsymbol{\Phi}}^{vec}_{Fullset}\right\|_{2,1,\boldsymbol{\psi}_n},\end{aligned} \tag{7}$$

where $\boldsymbol{\Xi}_0 = \boldsymbol{X}\boldsymbol{X}'$ is the emperical variance covariance matrix. One can observe that the length of nonzero components in each row of $\widetilde{\boldsymbol{\Phi}}$ is varying since it is a lower triangular matrix and hence ideally we should put varying penalty quantity for each row of the matrix. A smart way to handle the problem is to rescale the j^{th}column of $\widetilde{\boldsymbol{R}}$ by $1/\psi_{nj}$. So we define $\widetilde{\boldsymbol{R}}^{\star}$ with j^{th} column equal to $1/\psi_{nj}$ times the j^{th} coulmn $\widetilde{\boldsymbol{R}}$, and accordingly we define $\widetilde{\boldsymbol{\Phi}}^{\star}$ with j^{th} row equal to ψ_{nj} times the j^{th} row $\widetilde{\boldsymbol{\Phi}}$ which leads to $\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}} = \widetilde{\boldsymbol{R}}^{\star}\widetilde{\boldsymbol{\Phi}}^{\star}$. This transformation helps us to acheive invariance in $\widetilde{\boldsymbol{\Sigma}}$ for adaptive group LASSO. Therefore the optimization problem in (7) boils down to an unweighted ℓ_1/ℓ_2-penalized likelihood function,

$$\begin{aligned}\boldsymbol{Q}_n(\widetilde{\boldsymbol{\Phi}},\sigma^2,\tau_n,\mathbf{1}) = \mathbf{Tr}\left(\boldsymbol{\Xi}_0\left(\sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'\right)^{-1}\right) &+ \log\det\left(\sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'\right)\\ &+ \tau_n\left\|\widetilde{\boldsymbol{\Phi}}^{vec}_{Fullset}\right\|_{2,1,\mathbf{1}},\end{aligned} \tag{8}$$

and we want to restrict our search over the space of lower triangular matrices, with absolutely bounded elements and $\sigma \leq K < \infty$. Let us denote our search space by $\mathscr{P}_0^N$, where $N = 0.5n(n+1)+1$, the total number of parameters is an increasing function of n. Observe that with this rescaling, magnitude of our spatial basis matrix $\widetilde{\boldsymbol{R}}$ will change over n which let us think that the largest or smallest eigen value of $\widetilde{\boldsymbol{R}}$ may not be fixed for varying sample size. As a choice for ψ_{nj} one might be interested with $\psi_{nj} = 1/j$, *i.e.* by scaling down the j^{th} row $\widetilde{\boldsymbol{\varphi}}_{(j)}$ by its number of nonzero components and then take the ℓ_2-norm. Define, $\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}(\tau_n),\hat{\sigma}^2\right) = \arg\min_{\mathscr{P}_0^N}\boldsymbol{Q}_n(\widetilde{\boldsymbol{\Phi}},\sigma^2,\tau_n,\mathbf{1})$. Based on $\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}$, $\hat{\sigma}^2$ and $\widetilde{\boldsymbol{R}}$ we compute $\widehat{\boldsymbol{\Xi}}_{gL} = \hat{\sigma}^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL}\widetilde{\boldsymbol{R}}'$.

3 Main Result

In this section, we want to present asymptotic properties of the group Lasso type estimators obtained in section 2. We observe a single realization of a Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $\Xi = \sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\Omega}\widetilde{\boldsymbol{R}}^{'} = \sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}^{'}\widetilde{\boldsymbol{R}}^{'}$, denoted as $\boldsymbol{X} = \boldsymbol{\epsilon} + \boldsymbol{\pi}$ such that $\boldsymbol{\epsilon}$ is a single realization of a Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $\sigma^2\mathbb{I}$ and $\boldsymbol{\pi}$ is a single realization of a Gaussian random vector with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. Let us denote the corresponding sample space as $\boldsymbol{\Omega}_0$ collection of all possible sample points $\boldsymbol{\omega}$ *i.e.*,

$$\boldsymbol{\Omega}_0 = \left\{\boldsymbol{\omega} = (\boldsymbol{\epsilon} + \boldsymbol{\pi}) \in \mathbb{R}^n \text{ with, } \boldsymbol{\Pi}_{(n-r_n)\times n}\boldsymbol{\pi} = \mathbf{0}_{n-r_n}\right\}.$$

Before we can delve in to any further details we need to define two pre-requisite quantities parameteric search space $\mathscr{P}_0^N$, and domain of the optimizing function $\boldsymbol{Q}_n$. Note that, the parametric search space $\mathscr{P}_0^N$, can be defined as,

$$\mathscr{P}_0^N := \left\{\widetilde{\varphi}_{ji} \in \mathscr{P}_0,\ \forall\, i = 1, \ldots, j\ \&,\ \forall\, j = 1, \ldots, n\ ;\ \sigma \in \mathscr{P}_0^+\right\},$$

where for some bounded subset $\boldsymbol{P}_1 \subset \mathbb{R}$ and $\boldsymbol{P}_2 \subset \mathbb{R}^+$ we have, $\mathscr{P}_0 \subset \boldsymbol{P}_1$ and $\mathscr{P}_0^+ \subset \boldsymbol{P}_2$. Let A_0 and $A_\star$ are the sets of zero and non-zero rows of $\widetilde{\boldsymbol{\Phi}}$ respectively. Without loss of generality, $A_\star = \{1, 2, \ldots, r_n\}$ and $A_0 = \{r_n + 1, \ldots, n\}$. The case of generality is discussed in remark 3. Also, we assume existence of a set $\tilde{A}_0$, such that, $\sum_{j\in\tilde{A}_0} \|\widetilde{\boldsymbol{\varphi}}_{(j)}\|_2 \le \eta_1$ for some $\eta_1 \ge 0$. Existence of $\tilde{A}_0$ is often refered to as generalized sparsity condition (GSC) in literature of group Lasso estimation. Define $\tilde{A}_\star = \{1, \ldots, n\} \setminus \tilde{A}_0$, $\hat{A}_{\boldsymbol{\varphi}} = \left\{j : \left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2 > 0, 1 \le j \le n\right\}$ and, $B_0 = A_\star \cup \widehat{A}_{\boldsymbol{\varphi}}$. We know $|A_\star| = r_n$ and $\left|\hat{A}_{\boldsymbol{\varphi}}\right| = \hat{r}_n$, where $\hat{r}_n = \mathscr{R}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}\right)$.

For a $m \times m$ non-negative definite symmetric matrix $\boldsymbol{D}$, we denote $q(\le m)$ eigen values by $0 \le \lambda_{\min}(\boldsymbol{D}) = \lambda_1(\boldsymbol{D}) \le \lambda_2(\boldsymbol{D}) \le \ldots \le \lambda_q(\boldsymbol{D}) = \lambda_{\max}(\boldsymbol{D})$. We denote $\|\cdot\|_{\mathscr{T}_1}$ to define trace norm of the matrix where, trace norm is the sum of it's singular values. Therefore for a non-negative definite matrix, $\|\boldsymbol{D}\|_{\mathscr{T}_1} = \sum_{\lambda(\cdot)>0} |\lambda(\boldsymbol{D})| = \mathbf{Tr}(\boldsymbol{D})$. Since $\|\boldsymbol{D}\|_F = \sqrt{\mathbf{Tr}(\boldsymbol{D}'\boldsymbol{D})} = \left(\sum_{\lambda(\cdot)>0} \lambda^2(\boldsymbol{D})\right)^{1/2} \le \sum_{\lambda(\cdot)>0} |\lambda(\boldsymbol{D})| \le q\lambda_{\max}(\boldsymbol{D})$. For any subset $B \subset \{1, 2, \ldots, n\}$ we define a lower triangluar matrix, $\widetilde{\boldsymbol{\Phi}}_B$ such that, j^{th} row of $\widetilde{\boldsymbol{\Phi}}_B$ will take $\widetilde{\boldsymbol{\varphi}}_{(j)}$ if $j \in B$ or, $\mathbf{0}$ if $j \in B^c$. The corresponding row-wise vector representation for $\widetilde{\boldsymbol{\Phi}}_B$ is $\widetilde{\boldsymbol{\Phi}}_B^{vec} = (\widetilde{\boldsymbol{\varphi}}_j', j \in B)'$. So for B_0, we obtain $\widetilde{\boldsymbol{\Phi}}_{B_0}$ and we can define, $\Xi_{B_0} = \sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}_{B_0}^{'}\widetilde{\boldsymbol{R}}^{'}$ and correspondingly we define, $\widehat{\boldsymbol{\Xi}}_{gL,B_0} = \hat{\sigma}^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}^{'}\widetilde{\boldsymbol{R}}^{'}$. In what follows, the theorem which supports consistency for both estimation and selection of the non-zero rows of $\widetilde{\boldsymbol{\Phi}}_{B_0}$.

Theorem 1 (Estimation consistency of covariance parameters)**.** *Let $\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}, \hat{\sigma}^2$ is the minimizer of* (8). *Also if $\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}$ denotes the group LASSO estimate of the j^{th} row of the minimizer $\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}$, then* *for some preassigned $\varrho, \varsigma > 0$ and, for some $\mathscr{M} \geq 1$, let us define,*

$$\mathscr{P}^c_{\infty,n} = \left\{ \omega \in \mathbf{\Omega}_0 \; ; \; \frac{\left\| \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)} \right\|_2^2}{\left\| \widetilde{\boldsymbol{\varphi}}_{(j)} \right\|_2^2} < \varrho^2, \forall\, j \in B_0 \; ; \; \left| \hat{\sigma}^2 - \sigma^2 \right| < \varsigma \; ; \; |B_0| < \mathscr{M} r_n \right\},$$

where $r_n = Dn^{\gamma} + \boldsymbol{O}(1)$, with $D \geq 1$ and $\gamma < 2/(15 + 11\alpha)$ for some $\alpha > 0$. Suppose that conditions in assumption A hold and, $\tau_n < Cn^{(4+\alpha)\gamma/2}$ for a sufficiently large constant C. Then we have, $\mathbb{P}\left(\mathscr{P}^c_{\infty,n}\right) \overset{n \to \infty}{\longrightarrow} 1$.

The above theorem provides estimation consistency of rows of lower triangular cholesky decomposition matrix $\widetilde{\mathbf{\Phi}}$, the nugget parameter σ^2, and boundedness of cardinality of the set B_0. Despite therorem 1, we do need the following theorem to support consistency of $\widehat{\mathbf{\Xi}}_{gL}$.

Theorem 2 (Estimation consistency of $\widehat{\mathbf{\Xi}}_{gL}$)**.** *Under the same assumptions A as in theorem 1 and additionally if we choose $n^2\varsigma^2 < \mathscr{M}^3 n^{10/(15+11\alpha)}$ we have,*

$$\frac{1}{n^2} \left\| \widehat{\mathbf{\Xi}}_{gL,B_0} - \mathbf{\Xi}_{B_0} \right\|_F^2 = \boldsymbol{O}_p \left(\frac{\mathscr{M}^3 \varrho^2}{n^{2\left(\frac{10+11\alpha}{15+11\alpha}\right)}} \right).$$

3.1 Extention to the case $\boldsymbol{Z\beta} \neq \mathbf{0}$

Our first and foremost goal was to estimate the rank and the process matrix $\mathbf{\Sigma}$, so we first presented the simplier case with $\boldsymbol{Z\beta} = \mathbf{0}$. But a problem of kriging is incomplete if we are unable to add predictor variables to the model. So, we present the revised version of ℓ_1/ℓ_2-penalized likelihood function,

$$\boldsymbol{Q}_n(\widetilde{\mathbf{\Phi}}, \sigma^2, \boldsymbol{\beta}, \tau_n, \boldsymbol{\psi}_n) = (\boldsymbol{X} - \boldsymbol{Z\beta})' \, \mathbf{\Xi}^{-1} (\boldsymbol{X} - \boldsymbol{Z\beta}) + \log\det \mathbf{\Xi} + \tau_n \left\| \widetilde{\mathbf{\Phi}}^{vec}_{Fullset} \right\|_{2,1,\boldsymbol{\psi}_n}, \quad (9)$$

Corresponding to (9) the objective function (8) can be rewritten as,

$$\begin{aligned} \boldsymbol{Q}_n(\widetilde{\mathbf{\Phi}}, \sigma^2, \boldsymbol{\beta}, \tau_n, \mathbf{1}) = \mathbf{Tr} \left(\mathbf{\Xi}_{\boldsymbol{\beta}} \left(\sigma^2 \mathbb{I} + \widetilde{\boldsymbol{R}} \widetilde{\mathbf{\Phi}} \widetilde{\mathbf{\Phi}}' \widetilde{\boldsymbol{R}}' \right)^{-1} \right) \; &+ \; \log\det \left(\sigma^2 \mathbb{I} + \widetilde{\boldsymbol{R}} \widetilde{\mathbf{\Phi}} \widetilde{\mathbf{\Phi}}' \widetilde{\boldsymbol{R}}' \right) \\ &+ \; \tau_n \left\| \widetilde{\mathbf{\Phi}}^{vec}_{Fullset} \right\|_{2,1,\mathbf{1}}, \qquad (10) \end{aligned}$$

where, $\mathbf{\Xi}_{\boldsymbol{\beta}} = (\boldsymbol{X} - \boldsymbol{Z\beta})\,(\boldsymbol{X} - \boldsymbol{Z\beta})'$. Define, $\left(\widehat{\widetilde{\mathbf{\Phi}}}_{gL}(\tau_n), \hat{\sigma}^2, \widehat{\boldsymbol{\beta}} \right) = \arg\min_{\mathscr{P}_0^N} \boldsymbol{Q}_n(\widetilde{\mathbf{\Phi}}, \sigma^2, \boldsymbol{\beta}, \tau_n)$. The following theorem provides a similiar selection and estimation consistency of rows of lower triangular cholesky decomposition matrix $\widetilde{\mathbf{\Phi}}$ as theorem 1, when $\boldsymbol{Z\beta} \neq 0$. While acheiving such a consistency gives us the benifit of supporting our conjecture of being able to select the correct non-zero rows of $\widetilde{\mathbf{\Phi}}$, $i.e.$, succesfully estimate true rank of the matrix $\mathbf{\Sigma}$ even when $\boldsymbol{Z\beta} \neq 0$.

Theorem 3. *Let $\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}, \hat{\sigma}^2$ is the minimizer of* (10)*. Also if $\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}$ denotes the group LASSO estimate of the j^{th} row of the minimizer $\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}$, then under the same conditions of theorem 1 we have,* $\mathbb{P}\left(\mathscr{P}^c_{\infty,n}\right) \overset{n\to\infty}{\longrightarrow} 1$, *and,*

$$\frac{n^{\frac{11+9\alpha}{15+11\alpha}}}{2}\lambda_{\min}\left(\widehat{\boldsymbol{\Delta}}_{gL,B_0}\right)\left\|\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right\|_2^2 = \boldsymbol{O}_p\left(\frac{n^{(4+2\alpha)\gamma}n^{\frac{22+18\alpha}{15+11\alpha}}}{n^2}+\frac{n^{\alpha\gamma}n^{\frac{11+9\alpha}{15+11\alpha}}}{n}\right).$$

The above theorem provides estimation consistency of rows of lower triangular cholesky decomposition matrix $\widetilde{\boldsymbol{\Phi}}$ and nugget parameter σ^2, similar to theorem 1 even while $\boldsymbol{Z}\boldsymbol{\beta} \neq 0$. Despite therorem 2, we do need a theorem to support consistency of $\widehat{\Xi}_{gL}$ under the case when, $\boldsymbol{Z}\boldsymbol{\beta} \neq 0$. But presenting a separate theorem is unnecessary given one can easily follow the steps of theorem 2 to obtain the same rate of consistency.

4 Algorithm

In this section we will present an cost-effective algorithm for the optimization problem posed in (8). We have a block-wise function, blocks being the rows of a lower triangular matrix $\widetilde{\boldsymbol{\Phi}}$, along with a group LASSO type penalty, groups corresponding to each block. There has been few significant efforts behind building efficient algorithm to minimize a penalized likelihood. Although group wise penalization is not a completely different ball game, it still requires some special attention, which exploits the group structure and considers penalizing ℓ_2-norm of each group.

We will be using a **B**lock **C**oordiante **D**escent (BCD) method for a block multi-convex function under regularizing constraints,

$$\min_{\boldsymbol{x}\in\mathcal{X}}\left\{F(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n) = f(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n)+\sum_{j=1}^{n} r_j(\boldsymbol{x}_j)\right\}, \tag{11}$$

where $\boldsymbol{x}$ is decomposed into n blocks and $r_j(\boldsymbol{x}_j)$ is the regularizing constraint for the j^{th} block. On comparing (11) with (8) we can see that, in our case we have n blocks, $\mathcal{X}$ is the collection of lower triangular matrices of the form, $\widetilde{\boldsymbol{\Phi}}$, $F(\widetilde{\boldsymbol{\Phi}}) = \boldsymbol{Q}_n(\widetilde{\boldsymbol{\Phi}}, \tau_n)$ with,

$$f(\widetilde{\boldsymbol{\varphi}}_{(1)},\widetilde{\boldsymbol{\varphi}}_{(2)},\ldots,\widetilde{\boldsymbol{\varphi}}_{(n)}) = \mathbf{Tr}\left(\Xi_0\left(\sigma^2\mathbb{I}+\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'\right)^{-1}\right)+\log\det\left(\sigma^2\mathbb{I}+\widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{R}}'\right) \tag{12a}$$

$$r_j\left(\widetilde{\boldsymbol{\varphi}}_{(j)}\right) = \tau_n\frac{\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2}{j} \tag{12b}$$

To ease the computation we use **Matrix determinant lemma** and **Sherman-Morisson-Woddbury** matrix indentiy. We follow "**prox-linear**" algorithm (Xu & Yin (2013)) where the update for $\widetilde{\boldsymbol{\varphi}}_{(j)}$ in the k^{th} step is denoted by $\widetilde{\boldsymbol{\varphi}}^k_{(j)}$ and is given by,

$$\widetilde{\boldsymbol{\varphi}}^k_{(j)} = \arg\min_{\widetilde{\boldsymbol{\varphi}}_{(j)}}\left\{\left\langle \hat{\mathrm{g}}^k_j, \widetilde{\boldsymbol{\varphi}}_{(j)}-\widehat{\widetilde{\boldsymbol{\varphi}}}^{k-1}_{(j)}\right\rangle+\frac{L^{k-1}_j}{2}\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}-\widehat{\widetilde{\boldsymbol{\varphi}}}^{k-1}_{(j)}\right\|_2^2+r_j\left(\widetilde{\boldsymbol{\varphi}}_{(j)}\right)\right\},\ \forall j\ \ \&\ \ k \tag{13}$$

where the extrapolated point $\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1}$ is given as $\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1} = \widetilde{\boldsymbol{\varphi}}_{(j)}^{k-1} + \omega_i^{k-1}\left(\widetilde{\boldsymbol{\varphi}}_{(j)}^{k-1} - \widetilde{\boldsymbol{\varphi}}_{(j)}^{k-2}\right)$, with $\omega_i^{k-1} \geq 0$ is the extrapolation weight, $\hat{\mathrm{g}}_j^k = \bigtriangledown f_j^k\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1}\right)$ and,

$$f_j^k\left(\widetilde{\boldsymbol{\varphi}}_{(j)}\right) \stackrel{def}{=} f\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{(1)}^{k}, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(2)}^{k}, \ldots, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(j-1)}^{k}, \widetilde{\boldsymbol{\varphi}}_{(j)}, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(j+1)}^{k-1}, \ldots, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(s)}^{k-1}\right), \forall\ j \&\ k.$$

The second term on the right hand side, is added on the contrary to standard block coordinate descent algorithm, to make sure that the k^{th} update is not too far from the $(k-1)^{th}$ update in L_2 sense. Before we can do that we need to prove block multi-convexity (lemma 4) and Lipschitz continuity (lemma 5) of $f(\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2, \ldots, \boldsymbol{\varphi}_n)$ and $\bigtriangledown f_j^k\left(\widetilde{\boldsymbol{\varphi}}_{(j)}\right)$ respectively.

Generally, L_j^{k-1} is some constant greater that zero, and plays the role similar to the penalty parameter τ_n in $r_j(\widetilde{\boldsymbol{\varphi}}_{(j)})$, so if the k^{th} update is too far from $(k-1)^{th}$ update in L_2-sense, our objective would be to penalize it more and control it, so unlike standard group LASSO problem, here we have to take care of two penalty parameters rather than just one. So, we have a more challenging problem to solve, but if scaled properly one can chose constant L_j^{k-1} as a scalar multiplie of τ_n. Let us introduce a new quantity $\eta = L_j^{k-1}/\tau_n$, which is used to explain the rest our algorithm and this is refered to as a dual parameter for our optimization method.

To update (13) we use the fact that if, r_j can be represented as an indicator function of convex set $\mathcal{D}_j$, *i.e.* $r_j = \delta_{\mathcal{D}_j}(\widetilde{\boldsymbol{\varphi}}_{(j)}) = \mathbb{I}\left(\widetilde{\boldsymbol{\varphi}}_{(j)} \in \mathcal{D}_j\right)$, then $\widetilde{\boldsymbol{\varphi}}_{(j)}^{k} = \mathcal{P}_{\mathcal{D}_j}\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}\right)$, where $\mathcal{P}_{\mathcal{D}_j}$ is the projection to the set $\mathcal{D}_j$. Since for a group wise LASSO penalty, $r_j\left(\widetilde{\boldsymbol{\varphi}}_{(j)}\right) = \tau_n \left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2/j$, which is equivalent to say that we need our pre-penalized update $\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}$, first scaled down by its length j, and then project it on a surface of the sphere with radius η. And for our group wise LASSO penalty, we define our component wise scaled projection function is, $\mathcal{P}_{\mathcal{D}_j}(t) = \operatorname{sgn}(t)\max(\sqrt{|t|/j} - \sqrt{\eta}, 0)$. So the update rule (13) can be simplified and the following can be used component wise to obtain the j^{th} row,

$$\widetilde{\boldsymbol{\varphi}}_{(j)}^{k} = \operatorname{sgn}\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}\right)\left(\sqrt{\text{abs}\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}\right)/j} - \sqrt{\eta}\right)_+, \forall j\ \&\ k \qquad (14)$$

where all the above functions defined on the vector $\widehat{\widetilde{\boldsymbol{\varphi}}}_{(j)}^{k-1} - \hat{\mathrm{g}}_j^k/L_j^{k-1}$ are used component wise. Define the corresponding lower triangular matrix as $\widetilde{\boldsymbol{\Phi}}^k = \text{row-bind}(\widetilde{\boldsymbol{\varphi}}_{(1)}^{k\prime}, \widetilde{\boldsymbol{\varphi}}_{(2)}^{k\prime}, \cdots, \widetilde{\boldsymbol{\varphi}}_{(n)}^{k\prime})$ and now let us present the working algorithm for our optimization and following which we also provide a small modification in situations where a subsequent extrapolated update does not reduces the optimizing functional value.

(M 1) In case of $\boldsymbol{Q}(\widetilde{\boldsymbol{\Phi}}^k) \geq \boldsymbol{Q}(\widetilde{\boldsymbol{\Phi}}^{k-1})$ we modify the above algorithm by redoing the k^{th} iteration with $\widehat{\widetilde{\boldsymbol{\varphi}}}_i^{k-1} = \widetilde{\boldsymbol{\varphi}}_i^{k-1}$, *i.e.*, with out extrapolation.

Algorithm 1 Group LASSO algorithm for estimating a low rank covariance matrix

```
1: Initialization: $\widetilde{\boldsymbol{\Phi}}^{-1}$ and $\widetilde{\boldsymbol{\Phi}}^{0}$ lower triangular matrices as first two initial roots with no zero rows
2: Prefix: $\eta > 0$ and $\epsilon > 0$ prespecified
3: for $k = 1, 2, 3, \ldots$ do
4:     for $j = 1, 2, 3, \ldots, n$ do
         $\widehat{\widetilde{\boldsymbol{\varphi}}}^{k}_{(j)} \longleftarrow$ using (14)
5:     end for
6:     return Lower triangular matrix $\widehat{\widetilde{\boldsymbol{\Phi}}}^{k}$
         $\widetilde{\boldsymbol{\Phi}}^{-1} \longleftarrow \widetilde{\boldsymbol{\Phi}}^{0}$ and $\widetilde{\boldsymbol{\Phi}}^{0} \longleftarrow \widehat{\widetilde{\boldsymbol{\Phi}}}^{k}$
7:     for $j = 1, 2, \ldots, n$ do
         $temp_j \longleftarrow \left\| \widehat{\widetilde{\boldsymbol{\varphi}}}^{k}_{(j)} - \widehat{\widetilde{\boldsymbol{\varphi}}}^{k-1}_{(j)} \right\|_2$
8:     end for
9:     if $\max temp < \lambda$ then break and go to line 18
10:     else
11:         Go back to line 4 with $k = k + 1$
12:     end if
13: end for
14: return Lower triangular matrix $\widehat{\widetilde{\boldsymbol{\Phi}}}^{k}$
```

5 Numerical investigation

5.1 Simulation study

We follow spatial sampling design with an increasing domain asymptotic framework where sample sizes increases in proportion to the area of the sampling region. So we consider $m \times m$ square lattices where $m = 20, 25, 30, 35$ which makes sample sizes $n = 400, 625, 900$, respectively. For each choice we need to consider some true value of $\mathscr{R}(\boldsymbol{\Sigma})$, rank of $\boldsymbol{\Sigma}$, for different n we choose $\mathscr{R}(\boldsymbol{\Sigma}) = 30, 35, 40$. We generate our error term from a mean zero and nonstationary Gaussian process from a covariance function given by (3) and we consider different choices of $\widetilde{\boldsymbol{R}}(\boldsymbol{s})$ for example **R**adial **B**asis **F**unction (RBF), **W**endland **B**asis **F**ucntion (WBF), **F**ourier **B**asis **F**unction (FBF) etc. The data has been generated from model (1) for all possible combination of m, $\mathscr{R}(\boldsymbol{\Sigma})$ and $\widetilde{\boldsymbol{R}}(\boldsymbol{s})$, we generate n data points. From summarizing all the simulation results we believe that the method starts to work better for larger n.

If one considers a dyadic break of the two dimensional spatial domain, and pick centers of each of the regions as their knot points, then the first resolution will have 2^2 knots, second resolution will have 2^4 knots, $i.e.$ the k^{th} resolution will have 2^{2k} knot points. We have applied the concept of reversible jumps into our algorithm by considering a starting value of the number of effective knot points. For example lets say we start by considering all the knot points from the first two resolutions effective. After every iteration, we let our model to change by either dropping one of the knots which might have considered to be important earlier or selecting one of the knots which has not been considered to be important earlier.

5.2 Real data examples

The data set we used is part of a group of R data sets for monthly min-max temperatures and precipitation over the period $1895 - 1997$. It is a subset extracted from the more extensive

Lattice size	Local bi-square Basis Function			Wendland Basis Function		
(s)	$r = 30$	$r = 35$	$r = 40$	$r = 30$	$r = 35$	$r = 40$
20	**27.59 (0.81)**	**30.87 (0.66)**	**33.71 (1.60)**	**29.19 (0.14)**	**32.17 (0.56)**	**37.71 (1.05)**
25	**29.17 (1.49)**	**31.07 (2.66)**	**35.87 (1.02)**	**30.27 (0.91)**	**34.72 (1.01)**	**38.27 (0.52)**
30	**30.01 (1.05)**	**34.59 (1.89)**	**40.05 (2.89)**	**30.11 (1.52)**	**34.90 (0.91)**	**40.05 (1.09)**
35	**30.15 (0.91)**	**33.12 (0.88)**	**42.11 (1.05)**	**30.33 (0.25)**	**36.59 (1.90)**	**41.25 (0.29)**

Table 1: **Mean (Standard Devation)** of 200 Monte Carlo simulations for rank estimation of the nonstationary covariance matrix $\boldsymbol{\Sigma}$

US data record described in at (`www.image.ucar.edu/Data/US.monthly.met`). Observed monthly precipitation, min and max temperatures for the conterminous US $1895 - 1997$. We have taken a subset of the stations in Colorado. Temperature is in degrees C and precipitation is total monthly accumulation in millimeters. Note that minimum (maximum) monthly tempertuare is the mean of the daily minimum (maximum) temperatures. A rectagular lon-lat region $[-109.5, -101] \times [36.5, 41.5]$ larger than the boundary of Colorado comprises approximately 400 stations. Although there are additional stations reported in this domain, stations that only report preicipitation or only report temperatures have been excluded. In addition stations that have mismatches between locations and elevations from the two meta data files have also been excluded. The net result is 367 stations that have colocated temperatures and precipitation. We have used minimum temperature data as the observed process to apply our method and obtain the image plots below.

6 Discussion

Our work is quite significant from several perspective, although we would like to point out that it gives a dimension reduction perspective of estimation of low rank covariance matrix. As mentioned earlier Cressie & Johannesson, (2008) pointed out the benefit of assuming a fixed but lower rank than the actual dimension of the covariance matrix. They pointed out that inversion time of $n \times n$ covariance matrix, which is $\boldsymbol{O}(n^3)$ can now be reduced to $\boldsymbol{O}(nr^2)$, where r is assumed to be the known fixed rank. A previous knowledge about the value of r is quite hard to believe and our contribution is to figure out a relevant way to get around this. Although at this point we do not claim that we are able to provide an unbiased estimate of rank, but our result does provide consistent estimate of the covariance matrix along with linear model parameters. We also extended the work by Cressie & Johannesson, in the sense that our method allows one to assume that r can vary over n, the sample size, more precisely $r = r_n$ it can increase in a polynomial of n.

Now let us compare our finding with another recent study (Stein, 2015), which provides some examples and discusses scenarios where appoximating a true full rank covariance matrix $\boldsymbol{\Psi}_0$ with a matrix $\boldsymbol{\Psi}_1 = \rho^2\mathbb{I} + \boldsymbol{\Upsilon}$, where $\boldsymbol{\Upsilon}$ is a low rank matrix, does not reduces the Kulback-Liebler divergence considerably. As necessary, interesting and relevant this may sound, we

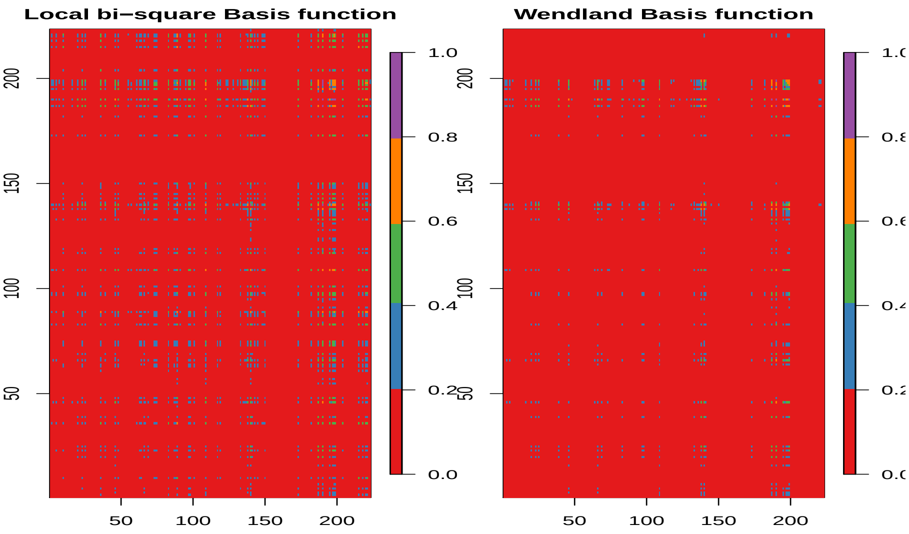

Figure 1: Quantile Image plot of $\widehat{\Xi}_{gL}$, the estimated covariance matrix of the observed process

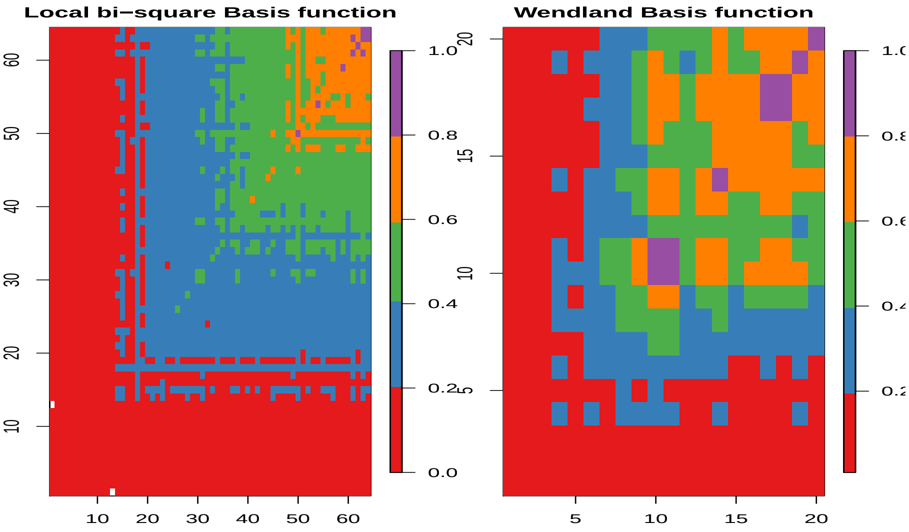

Figure 2: Quantile Image plot of $\widehat{\Phi}_{gL}$ estimated covariance matrix of the random effects vector

would like to point out dissimilarties. Firstly, unlike any full rank covariance matrix $\boldsymbol{\Psi}_0$, we assume true covariance matrix has the structure $\boldsymbol{\Psi}_0 = \rho^2\mathbb{I} + \boldsymbol{\Upsilon}$ and our approach estimates $\boldsymbol{\Psi}_0$ through estimates of ρ and $\boldsymbol{\Upsilon}$. Using the concept of capturing spatial dependence through a set of basis functions (Cressie & Johannesson, 2008) our model is further specified by considering the low rank compnent as, $\boldsymbol{\Upsilon} = \widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{K}}\widetilde{\boldsymbol{S}}^{'}$, where $\widetilde{\boldsymbol{K}}$ is a $n \times n$ matrix of rank r_n. As mentioned earlier r_n is a polynomial in n, we would like to refer our readers to assumption **(A 1)** which says $r_n = Dn^{\gamma} + \boldsymbol{O}(1)$, with $D > 0$ and $\gamma < 2/(15 + 11\alpha)$ with $\alpha > 0$. Although one might feel the necessity of estimating the nuisance parameter α. But let us point out the fact that our results works for any value of $\alpha > 0$. Even if we choose $\alpha = \alpha_n \longrightarrow 0$, $\gamma < 1/7.5$. This implies our finding covers Case 3 and a subset of Case 2 in Stein (2015). In the paper by Stein (2015) it is been pointed out KL divergence do not reduces sufficiently enough under a very special situation of stationary periodic process on line, which can be extended to be a process on surface, although can be quite challenging even for stationary periodic process. On the contrary our finding provides theoritical justification of consistent estimation of $\boldsymbol{\Psi}_0 = \rho^2\mathbb{I} + \widetilde{\boldsymbol{S}}\widetilde{\boldsymbol{K}}\widetilde{\boldsymbol{S}}^{'}$ in a more general set up.

Appendices

A Assumptions

(A 1) Let, the true rank $r_n = Dn^\gamma + \boldsymbol{O}(1)$, with $D > 0$, be increasing with n at some power $\gamma < 2/(15 + 11\alpha)$ with $\alpha > 0$.

(A 2) $\widetilde{\boldsymbol{R}}$ belongs to $\left\{\overline{\boldsymbol{R}}\,; \mathscr{C}_1 r_n \leq \lambda_{\min}\left(\overline{\boldsymbol{R}}'\overline{\boldsymbol{R}}\right) \leq \lambda_{\max}\left(\overline{\boldsymbol{R}}'\overline{\boldsymbol{R}}\right) \leq \mathscr{C}_2 r_n\right\}$, a class of well-conditioned matrices.

(A 3) Define $\widetilde{\boldsymbol{\Phi}}_{\mathscr{W}}$ to be collection of $n \times n$ lower triangular matrix with ℓ_2-norm of first r_n rows nonzero $i.e.$, for any preassigned $\phi_1, \phi_2 > 0$,

$$\widetilde{\boldsymbol{\Phi}} \in \widetilde{\boldsymbol{\Phi}}_{\mathscr{W}} = \left\{\overline{\boldsymbol{\Phi}}; \phi_1^2 < \left\|\overline{\boldsymbol{\varphi}}_{(j)}\right\|_2^2 < \phi_2^2,\ \forall\, j = 1, 2, \ldots, r_n\right\}.$$

Remark 1. Assumption **(A 1)** informs us that, the number of non-zero rows of $\widetilde{\boldsymbol{\Phi}}$ (or, rank r_n), changes in γ^{th} order polynomial of n. More precisely $\gamma < 2/(15 + 11\alpha)$. A more detailed explanation of the need of introducing such a parameter will be discussed later. Although it should be greater than zero and not depending on n. We leave the choice of α to the user. Observe that since we do not have a lower bound for γ, it can take any value strictly less that $2/(15 + 11\alpha)$. Although $\gamma < 0$, refers to cases with a rank 1 process covariance matrix $\boldsymbol{\Sigma}$. The choice $\gamma = 0$ has a specific significance, in which case $r_n = r$ refers to the case of **FRK**.

Remark 2. The assumption **(A 2)** implies that, the condition number of the matrix $\widetilde{\boldsymbol{R}}$, $\kappa\left(\widetilde{\boldsymbol{R}}\right) \leq \mathscr{C}_1/\mathscr{C}_2$. Under assumption **(A 1)**, **(A 2)** can be rewritten as,

$$\widetilde{\boldsymbol{R}} \in \left\{\overline{\boldsymbol{R}}\,; \mathscr{C}_1 n^\gamma \leq \lambda_{\min}\left(\overline{\boldsymbol{R}}'\overline{\boldsymbol{R}}\right) \leq \lambda_{\max}\left(\overline{\boldsymbol{R}}'\overline{\boldsymbol{R}}\right) \leq \mathscr{C}_2 n^\gamma\right\}.$$

Remark 3. Although WLOG, $A_\star = \{1, 2, \ldots, r_n\}$, in case the data indicate differently $i.e.$ $A_\star$ is any subset of $\{1, 2, \ldots, n\}$ with cardinality r_n, we can rotate $\widetilde{\boldsymbol{R}}(\boldsymbol{s})$ using a projection (permutation) matrix, $\widetilde{\boldsymbol{P}}_{A_\star}$ with $\widetilde{\boldsymbol{P}}_{A_\star}\widetilde{\boldsymbol{P}}'_{A_\star} = \mathbb{I}$. We define the rotation as, $\widetilde{\boldsymbol{R}}(\boldsymbol{s})\widetilde{\boldsymbol{P}}_{A_\star} = \left(\widetilde{\boldsymbol{R}}_{A_\star}(\boldsymbol{s}), \widetilde{\boldsymbol{R}}_{A_\star^c}(\boldsymbol{s})\right)'$. Also corresponding to $\widetilde{\boldsymbol{P}}_{A_\star}$, we can reorder our data. We denote corresponding data vector and sample variance covariance matrix as $\widetilde{\boldsymbol{P}}_{A_\star}\boldsymbol{X}$ and $\Xi_0(\widetilde{\boldsymbol{P}}_{A_\star}) = \widetilde{\boldsymbol{P}}_{A_\star}\boldsymbol{X}(\widetilde{\boldsymbol{P}}_{A_\star}\boldsymbol{X})' = \widetilde{\boldsymbol{P}}_{A_\star}\boldsymbol{X}\boldsymbol{X}'\widetilde{\boldsymbol{P}}'_{A_\star}$ respectively. Since our theoritical findings are based on $\lambda_{\min}(\Xi_0(\widetilde{\boldsymbol{P}}_{A_\star}))$ which is invariant over $\widetilde{\boldsymbol{P}}_{A_\star}$. Correspondingly, we define, $\Xi\left(\widetilde{\boldsymbol{P}}_{A_\star}\right) = \sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{P}}_{A_\star}\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'\widetilde{\boldsymbol{P}}'_{A_\star}\widetilde{\boldsymbol{R}}'$ and, $\widehat{\Xi}_{gL}\left(\widetilde{\boldsymbol{P}}_{A_\star}\right) = \hat{\sigma}^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{P}}_{A_\star}\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL}\widetilde{\boldsymbol{P}}'_{A_\star}\widetilde{\boldsymbol{R}}'$ and observe that both Ξ and $\widehat{\Xi}_{gL}$ remains unaltered. Henceforth with out loss of generality we will drop $\widetilde{\boldsymbol{P}}_A$ from rest of our calculations.

B Proofs of theorems and lemmas

Lemma 1 (Restricted reverse triangle inequality in ℓ_1- norm)**.** *Let a and b are two real numbers such that for some $\mathscr{A}_1, \mathscr{A}_2, \mathscr{B}_1, \mathscr{B}_2 > 0$, we have either,*

$$\boldsymbol{I} : \mathscr{A}_1 \leq a \leq \mathscr{A}_2 \textit{ and } \mathscr{B}_1 \leq |b| \leq \mathscr{B}_2 \textit{ with, } 0 < \mathscr{B}_1 \leq |b| \leq \mathscr{B}_2 < \mathscr{A}_1 \leq a \leq \mathscr{A}_2,$$

or,

$$\boldsymbol{II} : \mathscr{A}_1 \leq |a| \leq \mathscr{A}_2 \textit{ and } \mathscr{B}_1 \leq |b| \leq \mathscr{B}_2 \textit{ with, } 0 < \mathscr{B}_1 \leq |b| \leq \mathscr{B}_2 < \mathscr{A}_1 \leq |a| \leq \mathscr{A}_2,$$

then,

$$|a+b| > k_{\mathscr{A}_2}\left(|a| + |b|\right)$$

where, $k_{\mathscr{A}_2} = C/\mathscr{A}_2$.

Proof. For the case $\boldsymbol{I}$, we have the condition $0 < |b| < \mathscr{B}_2 < \mathscr{A}_1 < a$,

$$a > |b| \Leftrightarrow a + b > |b| + b.$$

Now either of the two possibilities, $b < 0$ in which case $b + |b| = 0$ or $b > 0$, then $b + |b| = 2b$. For both cases we can write, $a + b > 0$. The strict inequality allows us to say there exists a $C > 0$ such that $a + b > 2C$ and we have the following.

$$\begin{aligned} |a+b| &= \frac{|a+b|}{|a|+|b|}\left(|a|+|b|\right) \\ &\geq \frac{|a+b|}{\mathscr{A}_2 + \mathscr{B}_2}\left(|a|+|b|\right) \\ &> \frac{C}{\mathscr{A}_2}\left(|a|+|b|\right) \qquad \left(\text{ Since, } \mathscr{B}_2 \leq \mathscr{A}_2\right). \end{aligned}$$

Similary for the case $\boldsymbol{II}$, we have conditions $0 < b < \mathscr{B}_2 < \mathscr{A}_1 < a$, or $0 > -b > -\mathscr{B}_2 > -\mathscr{A}_1 > -a$, and for both cases, one can show $a + b > 0$. Hence we can prove the same as for case $\boldsymbol{I}$. □

Lemma 2. *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ are $m \times m$ non-negative definite symmetric matrices with $\lambda_{\min}(\boldsymbol{B}) = 0$ then,*

$$\mathbf{Tr}(\boldsymbol{A}) \leq \mathbf{Tr}(\boldsymbol{A}+\boldsymbol{B}) \leq \mathbf{Tr}(\boldsymbol{A}) + m\lambda_{\max}(\boldsymbol{B}). \tag{15}$$

Proof. Note that using the famous Weyl's theorem (Theorem 3.21, Schott (2005)) on $\boldsymbol{A}$ and $\boldsymbol{B}$ we have,

$$\lambda_h(\boldsymbol{A}) + \lambda_{\min}(\boldsymbol{B}) \leq \lambda_h(\boldsymbol{A}+\boldsymbol{B}) \leq \lambda_h(\boldsymbol{A}) + \lambda_{\max}(\boldsymbol{B}).$$

The above statement follows that,

$$\sum_{h=1}^{k}\lambda_h(\boldsymbol{A}) + k\lambda_{\min}(\boldsymbol{B}) \le \sum_{h=1}^{k}\lambda_h(\boldsymbol{A}+\boldsymbol{B}) \le \sum_{h=1}^{k}\lambda_h(\boldsymbol{A}) + k\lambda_{\max}(\boldsymbol{B}),\ \forall k = 1,\dots,m,$$

and if we choose $k = m$ we get,

$$\mathbf{Tr}(\boldsymbol{A}) + m\lambda_{\min}(\boldsymbol{B}) \le \mathbf{Tr}(\boldsymbol{A}+\boldsymbol{B}) \le \mathbf{Tr}(\boldsymbol{A}) + m\lambda_{\max}(\boldsymbol{B}).$$

Further if we use $\lambda_{\min}(\boldsymbol{B}) = 0$ then, we get (15). □

The left hand side of (15) gives monotonicity of property of the eigen values of of symmetric matrices where the inequality is strict if $\boldsymbol{B}$ is positive definite. One can also directly use Courant-Fischer min-max theorem to prove the above. Another important bound for the trace of a matrix product is given through the following lemma used in several places for the proof of our theorems.

Lemma 3. *If $\boldsymbol{A}$ and $\boldsymbol{B}$ are $m \times m$ nonnegative definite matrices, then,*

$$\lambda_{\min}\left(\boldsymbol{B}\right)\mathbf{Tr}\left(\boldsymbol{A}\right) \overset{LHS}{\le} \mathbf{Tr}\left(\boldsymbol{AB}\right) \overset{RHS}{\le} \lambda_{\max}\left(\boldsymbol{B}\right)\mathbf{Tr}\left(\boldsymbol{A}\right), \tag{16}$$

where LHS and RHS are acronyms for left hand side and right hand side respectively.

Proof. For proof of this lemma we use the following bounds from Marshall and Olkin (1979),

$$\sum_{i=1}^{m}\lambda_i(\boldsymbol{A})\lambda_{m-i+1}(\boldsymbol{B}) \le \sum_{i=1}^{m}\lambda_i(\boldsymbol{AB}) \le \sum_{i=1}^{m}\lambda_i(\boldsymbol{A})\lambda_i(\boldsymbol{B}).$$

Using the above we get,

$$\lambda_{\min}(\boldsymbol{B})\sum_{i=1}^{m}\lambda_i(\boldsymbol{A}) \le \sum_{i=1}^{m}\lambda_i(\boldsymbol{AB}) \le \lambda_{\max}(\boldsymbol{B})\sum_{i=1}^{m}\lambda_i(\boldsymbol{A}),$$

and finally equation (16). □

Lemma 4 (Block Multi-convexity)**.** *Under assumption* **(A 2)** *show that for,* $f(\widetilde{\boldsymbol{\varphi}}_{(1)}, \widetilde{\boldsymbol{\varphi}}_{(2)}, \dots, \widetilde{\boldsymbol{\varphi}}_{(n)})$ *as defined according as* (12a)*, where* $\widetilde{\boldsymbol{\varphi}}'_{(j)} = \big(\widetilde{\boldsymbol{\varphi}}_j, 0, 0, \dots, 0\big) = (\varphi_{j1}, \varphi_{j2}, \dots, \varphi_{jj}, 0, 0, \dots, 0)$ *is a block multi-convex function.*

Proof. If for each j, f is a convex function of $\widetilde{\boldsymbol{\varphi}}_{(j)}$, while the other blocks are fixed we call the function to be block multi-convex function. Define,

$$\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}_0\right) = \left(\left(\frac{\partial^2}{\partial\widetilde{\varphi}_{jk}\partial\widetilde{\varphi}_{jk'}} f\left(\widetilde{\boldsymbol{\Phi}}\right)\Big|_{\widetilde{\boldsymbol{\Phi}}=\widetilde{\boldsymbol{\Phi}}_0}\right)\right)_{k=1,k'=1}^{j,j}, \qquad \forall j = 1,\dots,n, \tag{17}$$

as the $j \times j$ Hessian matrix, with respect to $\widetilde{\widetilde{\boldsymbol{\varphi}}}_{(j)}$, where $\widetilde{\boldsymbol{\Phi}}_0$ is a solution to,

$$\frac{\partial}{\partial \widetilde{\boldsymbol{\varphi}}_j} f(\widetilde{\boldsymbol{\Phi}}) = \mathbf{0}, \qquad \forall j = 1, \ldots, n. \tag{18}$$

It is enough to show that $\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}_0\right)$ is a positive definite matrix. Note that for all j and k,

$$\begin{aligned}
\frac{\partial}{\partial \widetilde{\varphi}_{jk}} f(\widetilde{\boldsymbol{\Phi}}) &= \mathbf{Tr}\left[-\left(\Xi^{-1}\Xi_0\Xi^{-1}\right)\widetilde{\boldsymbol{R}}\frac{\partial\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'}{\partial\widetilde{\varphi}_{jk}}\widetilde{\boldsymbol{R}}'\right] + \mathbf{Tr}\left[\Xi^{-1}\widetilde{\boldsymbol{R}}\frac{\partial\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'}{\partial\widetilde{\varphi}_{jk}}\widetilde{\boldsymbol{R}}'\right] \\
&= \mathbf{Tr}\left[-\left(\Xi^{-1}\Xi_0\Xi^{-1}\right)\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right] + \mathbf{Tr}\left[\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right] \\
&= \mathbf{Tr}\left[\left(\mathbb{I} - \Xi^{-1}\Xi_0\right)\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right] \\
&= \mathbf{Tr}\left[\Xi^{-1}\left(\Xi - \Xi_0\right)\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right], \text{ where } \boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k} = \widetilde{\boldsymbol{R}}\frac{\partial\widetilde{\boldsymbol{\Phi}}\widetilde{\boldsymbol{\Phi}}'}{\partial\widetilde{\varphi}_{jk}}\widetilde{\boldsymbol{R}}' \\
&= \mathbf{Tr}\left[\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\Xi^{-1}\left(\Xi - \Xi_0\right)\Xi^{-1}\right] \qquad (19) \\
&= \left[vec\left(\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right)\right]'\left(\Xi^{-1}\otimes\Xi^{-1}\right)\left[vec\left(\Xi - \Xi_0\right)\right],
\end{aligned}$$

where, $\otimes$ is used to denote Kronecker product. The last identity is due to, $\mathbf{Tr}(\boldsymbol{V}_1\boldsymbol{V}_2\boldsymbol{V}_3\boldsymbol{V}_4) = [vec(\boldsymbol{V}_1')]'(\boldsymbol{V}_4'\otimes\boldsymbol{V}_2)[vec(\boldsymbol{V}_3)]$ (Theorem 8.12, Schott (2005)) with $\boldsymbol{V}_1 = \boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}$, $\boldsymbol{V}_2 = \boldsymbol{V}_4 = \Xi^{-1}$, and $\boldsymbol{V}_3 = \Xi - \Xi_0$. Note that $\Xi^{-1}\otimes\Xi^{-1}$ is positive definite matrix, $vec\left(\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right) \geq \mathbf{0}$. Since $\widetilde{\boldsymbol{\Phi}}_0$ is solution to equation (18), $\Xi\left(\widetilde{\boldsymbol{\Phi}}_0\right) - \Xi_0 = \mathbf{0}$ matrix, where $\Xi\left(\widetilde{\boldsymbol{\Phi}}_0\right) = \sigma^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widetilde{\boldsymbol{\Phi}}_0\widetilde{\boldsymbol{\Phi}}_0'\widetilde{\boldsymbol{R}}'$. Let the $(k,k')^{th}$ element of the Hessian matrix is denoted as $\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}\right)[k,k']$ and observe that,

$$\begin{aligned}
\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}\right)[k,k'] &= \overbrace{-\mathbf{Tr}\left[\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k'}\Xi^{-1}\left(\Xi - \Xi_0\right)\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right]}^{\text{First Component}} + \overbrace{\mathbf{Tr}\left[\Xi^{-1}\left(\Xi - \Xi_0\right)\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),kk'}\right]}^{\text{Second Component}} \\
&+ \overbrace{-\mathbf{Tr}\left[\Xi^{-1}\left(\Xi - \Xi_0\right)\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k'}\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right]}^{\text{Third Component}} + \overbrace{\mathbf{Tr}\left[\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k'}\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right]}^{\text{Fourth Component}}. \; (20)
\end{aligned}$$

The above equation (20) is obtained by differentiating (19) with respect to $\widetilde{\varphi}_{jk'}$. The first three component at $\widetilde{\boldsymbol{\Phi}}_0$ in equation (20) are zero. Hence, $\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}_0\right)[k,k']$ is just the fourth component in equation (20) computed under $\widetilde{\boldsymbol{\Phi}}_0$ and as a special case $\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}_0\right)[k,k]$, the k^{th} diagonal element equals to $\mathbf{Tr}\left[\left(\Xi^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right)^2\middle|\widetilde{\boldsymbol{\Phi}} = \widetilde{\boldsymbol{\Phi}}_0\right]$ which makes all the diagonal elements positive.

For a vector $\boldsymbol{a} = (a_1, \ldots, a_j)' \neq \mathbf{0}$ of length j,

$$\begin{aligned}
\boldsymbol{a}'\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}_0\right)\boldsymbol{a} &= \sum_{k=1}^{j}\sum_{k'=1}^{j} a_k\, \mathbf{Tr}\left[\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k'}\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\middle|\widetilde{\boldsymbol{\Phi}} = \widetilde{\boldsymbol{\Phi}}_0\right] a_{k'} \\
&= \sum_{k=1}^{j}\sum_{k'=1}^{j} \mathbf{Tr}\left[a_k\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k'}\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}a_{k'}\middle|\widetilde{\boldsymbol{\Phi}} = \widetilde{\boldsymbol{\Phi}}_0\right] \\
&= \mathbf{Tr}\left[\sum_{k=1}^{j}\sum_{k'=1}^{j} a_{k'}\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k'}\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}a_k\middle|\widetilde{\boldsymbol{\Phi}} = \widetilde{\boldsymbol{\Phi}}_0\right] \\
&= \mathbf{Tr}\left[\sum_{k=1}^{j} a_k\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\sum_{k'=1}^{j} a_{k'}\boldsymbol{\Xi}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k'}\middle|\widetilde{\boldsymbol{\Phi}} = \widetilde{\boldsymbol{\Phi}}_0\right] \\
&= \mathbf{Tr}\left[\left(\boldsymbol{\Xi}^{-1}\sum_{k=1}^{j} a_k\boldsymbol{\Delta}_{\boldsymbol{\Sigma}}^{(j),k}\right)^2\middle|\widetilde{\boldsymbol{\Phi}} = \widetilde{\boldsymbol{\Phi}}_0\right] > 0. \qquad (21)
\end{aligned}$$

Hence $\mathbb{H}_j\left(\widetilde{\boldsymbol{\Phi}}_0\right)$ is a positive definite matrix. □

Lemma 5 (Lipschitz continuity)**.** *For $f_j^k(\widetilde{\boldsymbol{\varphi}}_{(j)})$ as defined in section 4, show $\bigtriangledown f_j^k(\widetilde{\boldsymbol{\varphi}}_{(j)})$ is Lipschitz continuous in $\widetilde{\boldsymbol{\varphi}}_{(j)}$.*

Proof. Once we follow through the steps of lemma 4, we can see, $\bigtriangledown f_j^k(\widetilde{\boldsymbol{\varphi}}_{(j)})$ is a vector of length j with the k^{th} component as $\mathbf{Tr}\left[\boldsymbol{\Xi}_{(j)}^{-1}\left(\boldsymbol{\Xi}_{(j)} - \boldsymbol{\Xi}_0\right)\boldsymbol{\Xi}_{(j)}^{-1}\boldsymbol{\Delta}_{\boldsymbol{\Sigma}_{(j)}}^{(j),k}\right]$, where $\boldsymbol{\Xi}_{(j)}$ and $\boldsymbol{\Sigma}_{(j)}$ are computed using $\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{(1)}^{k-1}, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(2)}^{k-1}, \ldots, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(j-1)}^{k-1}, \widetilde{\boldsymbol{\varphi}}_{(j)}, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(j+1)}^{k-1}, \ldots, \widehat{\widetilde{\boldsymbol{\varphi}}}_{(n)}^{k-1}\right)$ as the rows of $\widetilde{\boldsymbol{\Phi}}$. So $\bigtriangledown f_j^k(\widetilde{\boldsymbol{\varphi}}_{(j)})$ is a differentiable function w.r.t $\widetilde{\boldsymbol{\varphi}}_{(j)}$ and hence it is Lipschitz continuous. □

Lemma 6. *For the two $n \times n$ positive definite matrices $\boldsymbol{\Sigma}_{B_0}$ and $\widehat{\boldsymbol{\Sigma}}_{B_0}$,*

$$\sum_{\lambda(\cdot)>0}\left|\lambda\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}\right) - \lambda\left(\boldsymbol{\Sigma}_{B_0}\right)\right| \leq |B_0| \sum_{\lambda(\cdot)\neq 0}\left|\lambda\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right)\right|, \qquad (22)$$

where the summation in left side of (22) *is over all possible eigen values of $\boldsymbol{\Sigma}_{B_0}$ and $\widehat{\boldsymbol{\Sigma}}_{gL,B_0}$ and the right side of* (22) *is over all possible eigen values of $\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}$.*

Proof. By part (b) of Theorem 1.20 in Simon (1979) we know, for any pair of finite dimensional self-adjoint matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$ if, we denote $\lambda_n(\cdot)$ denotes eigen values of a matrix then,

$$\lambda_m\left(\boldsymbol{A}\right) - \lambda_m\left(\boldsymbol{B}\right) = \sum_{n=1}^{N} e_{mn}\lambda_n\left(\boldsymbol{A} - \boldsymbol{B}\right)$$

where, $\boldsymbol{E} = \left(e_{mn}\right)_{m=1,n=1}^{N,N}$ is a **d**oubly **s**ub-**s**tochastic (dss) matrix. We define, a matrix $\boldsymbol{E} = \left(e_{mn}\right)_{m=1,n=1}^{N,N}$ to be dss iff,

$$\sum_{n=1}^{N}|e_{mn}| \leq 1,\ \forall\ m = 1, 2, \ldots, N, \text{ and } \sum_{m=1}^{N}|e_{mn}| \leq 1,\ \forall\ n = 1, 2, \ldots, N.$$

Therefore,

$$\begin{aligned}
\left|\lambda_m\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}\right)-\lambda_m\left(\boldsymbol{\Sigma}_{B_0}\right)\right| &= \left|\sum_{\lambda_n(\cdot)\neq 0} e_{mn}\lambda_n\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right| \\
&\leq \left(\sum_{\lambda_n(\cdot)\neq 0} e_{mn}^2\right)^{1/2}\left(\sum_{\lambda_n(\cdot)\neq 0}\lambda_n^2\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right)^{1/2} \\
&\leq \sum_{\lambda_n(\cdot)\neq 0}|e_{mn}|\sum_{\lambda_n(\cdot)\neq 0}\left|\lambda_n\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right| \\
&\leq \sum_{\lambda_n(\cdot)\neq 0}\left|\lambda_n\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right|, \qquad (23)
\end{aligned}$$

where, the first inequality in equation (23) is due to Cauchy-Bunyakovsky-Schwarz (**CBS**) inequality. The second inequality is obtained by using ℓ_2- norm is smaller than ℓ_1- norm. Finally the last inequality is due to $\boldsymbol{E}=((e_{m,n}))_{m=1,n=1}^{N,N}$ is a dss matrix. So, if we add both sides of (23) over all non-zero eigen values of both $\widehat{\boldsymbol{\Sigma}}_{gL,B_0}$ and $\boldsymbol{\Sigma}_{B_0}$ so we have,

$$\begin{aligned}
\sum_{\lambda_m(\cdot)>0}\left|\lambda_m\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}\right)-\lambda_m\left(\boldsymbol{\Sigma}_{B_0}\right)\right| &\leq \sum_{\lambda_m(\cdot)>0}\left\{\sum_{\lambda_n(\cdot)\neq 0}\left|\lambda_n\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right|\right\} \\
&= \sum_{\lambda_n(\cdot)\neq 0}\left|\lambda_n\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right|\left(\sum_{\lambda_m(\cdot)>0}1\right) \\
&= |B_0|\sum_{\lambda_n(\cdot)\neq 0}\left|\lambda_n\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right|,
\end{aligned}$$

and thus we have lemma 6. □

For an index set $\tilde{A}_1$ that satisfies $\tilde{A}_{\boldsymbol{\varphi}}=\{j:\|\hat{\boldsymbol{\varphi}}_{gL,j}\|_2>0\}\subseteq\tilde{A}_1\subseteq\tilde{A}_{\boldsymbol{\varphi}}\cup\tilde{A}_\star$, we consider the following sets:

	“Large” $\|\boldsymbol{\varphi}_j\|_2$ (i.e. $\tilde{A}_\star$)	“Small” $\|\boldsymbol{\varphi}_j\|_2$ (i.e. $\tilde{A}_0$)
$\tilde{A}_1$	$\tilde{A}_3$	$\tilde{A}_4$
$\tilde{A}_2=\tilde{A}_1^c$	$\tilde{A}_5$	$\tilde{A}_6$

We can deduce some relations from the above table $\tilde{A}_3=\tilde{A}_1\cap\tilde{A}_\star$, $\tilde{A}_4=\tilde{A}_1\cap\tilde{A}_0$, $\tilde{A}_5=\tilde{A}_1^c\cap\tilde{A}_\star$, $\tilde{A}_6=\tilde{A}_2\cap\tilde{A}_0$, and hence we have $\tilde{A}_3\cup\tilde{A}_4=\tilde{A}_1$, $\tilde{A}_5\cup\tilde{A}_6=\tilde{A}_2$, and $\tilde{A}_3\cap\tilde{A}_4=\tilde{A}_5\cap\tilde{A}_6=\phi$. Also, let $|\tilde{A}_1|=r_1$.

For some preassigned $\varrho, \varsigma > 0, \mathscr{M} > 1$, and $\varrho_n = Cn^{\alpha\gamma}, \varsigma_n, \mathscr{M}_n = Cn^{\alpha\gamma}$ for some generic constant C, and $\alpha > 0$, define an increasing sequence of sets, $\mathscr{P}_n = P_{1n} \cup P_{2n} \cup P_{3n}$ where,

$$\begin{aligned}
P_{1n} &= \left\{ \boldsymbol{\omega} \in \boldsymbol{\Omega}_0 \; ; \; \varrho^2 \leq \frac{\left\| \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)} \right\|_2^2}{\left\| \widetilde{\boldsymbol{\varphi}}_{(j)} \right\|_2^2} \leq \varrho_n^2, \text{ for some } j \in B_0 \right\}, \\
P_{2n} &= \left\{ \boldsymbol{\omega} \in \boldsymbol{\Omega}_0 \; ; \; \varsigma \leq \left| \hat{\sigma}^2 - \sigma^2 \right| \leq \varsigma_n \right\}, \text{ and,} \\
P_{3n} &= \left\{ \boldsymbol{\omega} \in \boldsymbol{\Omega}_0 \; ; \; \mathscr{M} r_n \leq |B_0| \leq \mathscr{M}_n r_n \right\}.
\end{aligned}$$

Since, $\varrho_n, \varsigma_n, \mathscr{M}_n \uparrow \infty$,

$$\begin{aligned}
P_{1n} &\nearrow \left\{ \boldsymbol{\omega} \in \boldsymbol{\Omega}_0 \; ; \; \varrho^2 \leq \frac{\left\| \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)} \right\|_2^2}{\left\| \widetilde{\boldsymbol{\varphi}}_{(j)} \right\|_2^2}, \text{ for some } j \in B_0 \right\} \overset{def}{=} P_{1\infty}, \\
P_{2n} &\nearrow \left\{ \boldsymbol{\omega} \in \boldsymbol{\Omega}_0 \; ; \; \varsigma \leq \left| \hat{\sigma}^2 - \sigma^2 \right| \right\} \overset{def}{=} P_{2\infty}, \text{ and,} \\
P_{3n} &\nearrow \left\{ \boldsymbol{\omega} \in \boldsymbol{\Omega}_0 \; ; \; \mathscr{M} r_n \leq |B_0| \right\} \overset{def}{=} P_{3\infty}.
\end{aligned}$$

Let us now define, $\mathscr{P}_{\infty,n} = P_{1\infty} \cup P_{2\infty} \cup P_{2\infty}$. Hence as $n \longrightarrow \infty$, $\mathscr{P}_n \cap \mathscr{P}_{\infty,n}^c \searrow \phi$, an empty set. Define,

$$\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}} = \frac{\lambda_{\min}\left(\frac{\boldsymbol{\Xi}_{\boldsymbol{\beta}}}{n}\right)}{\lambda_{\max}\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right) \lambda_{\max}\left(\boldsymbol{\Xi}_{B_0}\right)},$$

where $\boldsymbol{\Xi}_{\boldsymbol{\beta}} = (\boldsymbol{X} - \boldsymbol{Z\beta})(\boldsymbol{X} - \boldsymbol{Z\beta})'$ and $\boldsymbol{X}$ has Gaussian process with mean $\boldsymbol{Z\beta}$ and variance covariance matrix $\boldsymbol{\Xi}$. To prove Theorem 1 and 3 we need boundedness of $\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}}$, which is given in the following lemma.

Lemma 7 (Boundedness for $\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}}$). *Under assumptions* **(A 1)** - **(A 3)** *on* $\mathscr{P}_n$, $\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}}$ *satisfies*

$$\frac{1}{\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}}} \leq \boldsymbol{O}_p\left(n^{(3+4\alpha)\gamma}\right)$$

Proof. Note that under assumption **(A 1)**, on $\mathscr{P}_n$ $|B_0| \leq \mathscr{M}_n r_n = Cn^{(1+\alpha)\gamma}$, since $\mathscr{M}_n =$

$Cn^{\alpha\gamma}$. Therefore,

$$\begin{aligned}
\lambda_{\max}\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right) &= \hat{\sigma}^2 + \lambda_{\max}\left(\widetilde{\boldsymbol{R}}\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}\widetilde{\boldsymbol{R}}'\right) \\
&\leq \hat{\sigma}^2 + \textbf{Tr}\left(\widetilde{\boldsymbol{R}}\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}\widetilde{\boldsymbol{R}}'\right) \\
&= \hat{\sigma}^2 + \textbf{Tr}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right) \\
&\leq \hat{\sigma}^2 + \lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\textbf{Tr}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}\right) \text{ (By RHS of lemma 3)} \\
&\leq \hat{\sigma}^2 + |B_0|\lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\lambda_{\max}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}\right) \\
&\leq \hat{\sigma}^2 + Cn^{(1+\alpha)\gamma}\lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\max_{j\in B_0}\widehat{\widetilde{\boldsymbol{\varphi}}}^2_{gL,jj} \\
&\leq \hat{\sigma}^2 + Cn^{(2+\alpha)\gamma}\max_{j\in B_0}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2^2 \text{(By assumption (}\textbf{A 1}\text{) and (}\textbf{A 2}\text{))} \qquad (24) \\
&\leq \hat{\sigma}^2 + Cn^{(2+\alpha)\gamma}\max_{j\in B_0}\left\{\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2 + \left\|\widetilde{\boldsymbol{\varphi}}^2_{(j)}\right\|_2^2\right\} \\
&\leq \hat{\sigma}^2 + Cn^{(2+\alpha)\gamma}\max_{j\in B_0}\left\{\varrho_n^2\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2\right\} \\
&\leq Cn^{(2+3\alpha)\gamma}, \text{(By assumption (}\textbf{A 3}\text{) and } \varrho_n = Cn^{\alpha\gamma}\text{)} \qquad (25)
\end{aligned}$$

where C is a generic constant. If we follow the same trick as in equation (24) we have,

$$\begin{aligned}
\lambda_{\max}\left(\boldsymbol{\Xi}_{B_0}\right) &\leq \sigma^2 + Cn^{(1+\alpha)\gamma}\max_{j\in B_0}\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2 \\
&\leq Cn^{(1+\alpha)\gamma}. \text{(By assumption (}\textbf{A 3}\text{))} \qquad (26)
\end{aligned}$$

Hence using equations (25) and (26) we have,

$$\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}} \geq \frac{C}{n^{(3+4\alpha)\gamma}}\boldsymbol{O}_p\left(\lambda_{\min}\left(\left(\boldsymbol{X} - \boldsymbol{Z\beta}\right)\left(\boldsymbol{X} - \boldsymbol{Z\beta}\right)'\right)/n\right),$$

where C is a generic constant. We know $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)'$ follows a Gaussian distribution with mean $\boldsymbol{Z\beta}$ and covariance matrix $\Xi_{A_\star}$ and, $c_1 = \lambda_{\min}(\Xi_{A_\star})$ which gives,

$$c_1\lambda_{\min}\left(\frac{\left(\boldsymbol{X} - \boldsymbol{Z\beta}\right)\left(\boldsymbol{X} - \boldsymbol{Z\beta}\right)'}{n}\right) \geq \lambda_{\min}\left(\frac{\left(\boldsymbol{X} - \boldsymbol{Z\beta}\right)\Xi_{A_\star}^{-1}\left(\boldsymbol{X} - \boldsymbol{Z\beta}\right)'}{n}\right) = \lambda_{\min}\left(\frac{\boldsymbol{X}_0\boldsymbol{X}_0'}{n}\right)$$

where, $\boldsymbol{X}_0$ follows a Gaussian distribution with mean $\boldsymbol{0}$ and identity as its covariance matrix. By Theorem 2 and Remark 1 in Bai & Yin (1993), we know that $\lambda_{\min}\left(\frac{\boldsymbol{X}_0\boldsymbol{X}_0'}{n}\right) \xrightarrow{a.s.} 1$. And since almost sure convergence implies convergence in probability, we have $\lambda_{\min}\left(\frac{\boldsymbol{X}_0\boldsymbol{X}_0'}{n}\right) = \boldsymbol{O}_p\left(c_2\right)$,

$$\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}} \geq \frac{C}{n^{(3+4\alpha)\gamma}}\lambda_{\min}\left(\frac{\boldsymbol{X}_0\boldsymbol{X}_0'}{n}\right) = \boldsymbol{O}_p\left(\frac{1}{n^{(3+4\alpha)\gamma}}\right).$$

This also works for the case $\boldsymbol{Z\beta} = 0$, where $\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}=0}$ is defined as,

$$\boldsymbol{\Lambda}_n^{\boldsymbol{\beta}=0} = \frac{\lambda_{\min}\left(\frac{\boldsymbol{\Xi}_0}{n}\right)}{\lambda_{\max}\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right)\lambda_{\max}\left(\boldsymbol{\Xi}_{B_0}\right)} = \boldsymbol{\Lambda}_n,$$

where $\boldsymbol{\Xi}_0 = \boldsymbol{XX}'$. Hence we have lemma 7. □

Lemma 8. *Under assumption* **(A 1) - (A 3)** *prove that on* $\mathscr{P}_n$,

$$\begin{aligned} G_1 n^{\gamma} &\leq n^{\gamma}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 \\ &\leq \lambda_{\min}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 \\ &\leq \mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right) \leq G_2 n^{(5+\alpha)\gamma/2}, \end{aligned}$$

where, G_1, G_2 *are generic constants. Therefore,*

$$G_1 n^{\gamma} \leq \mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right) \leq G_2 n^{(5+\alpha)\gamma/2}. \tag{27}$$

Proof. Let us first define, $\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Omega}}_{B_0} = \widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}$ and observe that,

$$\begin{aligned} \widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Omega}}_{B_0} &= \widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0} - \widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0} + \widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0} \\ &= \widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}\right)' + \left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}\right)\widetilde{\boldsymbol{\Phi}}'_{B_0} \\ &= \left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0} + \widetilde{\boldsymbol{\Phi}}_{B_0}\right)\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}\right)' + \left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}\right)\widetilde{\boldsymbol{\Phi}}'_{B_0} \\ &= \left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)\left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)' + \widetilde{\boldsymbol{\Phi}}_{B_0}\left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)' + \left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)\widetilde{\boldsymbol{\Phi}}'_{B_0}, \end{aligned} \tag{28}$$

where, $\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0} = \widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}$. Hence, $\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Omega}}_{B_0}$ can be represented as sum to two matrices, $\boldsymbol{U}_1 = \left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)\left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)'$ and $\boldsymbol{U}_2 = \widetilde{\boldsymbol{\Phi}}_{B_0}\left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)' + \left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)\widetilde{\boldsymbol{\Phi}}'_{B_0}$ where $\boldsymbol{U}_1$, $\boldsymbol{U}_2$ are $n \times n$ symmetric matrix with $\mathscr{R}(\boldsymbol{U}_2) = r_n$, since $\widetilde{\boldsymbol{\Phi}}_{B_0}$ is a lower triangular matrix with exactly r_n non-zero rows. Note that, under assumption **(A 3)**, P_{1n} can be rewritten as,

$$P_{1n} = \left\{\boldsymbol{\omega} \in \boldsymbol{\Omega}_0 \ ; \ C_1\varrho^2 \leq \left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2 \leq C_2\varrho_n^2, \text{ for some } j \in B_0\right\},$$

where, C_1, C_2 are generic constants. That means there exist one $j \in B_0$ such that $\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2 \geq$

$C_1\varrho^2$. Hence,

$$\begin{aligned}
C_1\varrho \le \left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2 &= \left\{\sum_{i=1}^{j}\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,ji}-\widetilde{\boldsymbol{\varphi}}_{ji}\right)^2\right\}^{1/2} \\
&\le \sum_{i=1}^{j}\left|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,ji}-\widetilde{\boldsymbol{\varphi}}_{ji}\right| \\
&= \sum_{i=1}^{j-1}\left|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,ji}-\widetilde{\boldsymbol{\varphi}}_{ji}\right| + \left|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,jj}-\widetilde{\boldsymbol{\varphi}}_{jj}\right| \\
C_1\varrho - \sum_{i=1}^{j-1}\left|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,ji}-\widetilde{\boldsymbol{\varphi}}_{ji}\right| &\le \left|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,jj}-\widetilde{\boldsymbol{\varphi}}_{jj}\right|. \qquad (29)
\end{aligned}$$

So if we choose ϱ, ϕ_1 such that, $2\sum_{i=1}^{j-1}\left|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,ji}-\widetilde{\boldsymbol{\varphi}}_{ji}\right| \le C_1\varrho$, then under P_{1n}, $\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}$ is diagonally dominant and hence on $\mathscr{P}_n$, $\boldsymbol{U}_1$, $\boldsymbol{U}_2$, and $\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Omega}}_{B_0}$ are non-negative definite, so if we use (28) we get,

$$\begin{aligned}
\mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right) &= \mathbf{Tr}\left[\widetilde{\boldsymbol{R}}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right)\widetilde{\boldsymbol{R}}'\right] \\
&= \mathbf{Tr}\left[\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right)\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right] \\
&\ge \lambda_{\min}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\mathbf{Tr}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right) \text{(LHS of lemma 3)} \\
&\ge \lambda_{\min}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\mathbf{Tr}\left[\left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)\left(\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Phi}}_{B_0}\right)'\right] \\
&= \lambda_{\min}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0}-\widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2, \qquad (30)
\end{aligned}$$

Finally using assumption **(A 3)** on the set $\mathscr{P}_n$ we have,

$$\mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right) \ge n^\gamma\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0}-\widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 \ge \phi_1^2\varrho^2 n^\gamma = G_1 n^\gamma,$$

where, G_1 is a generic constant used in (27). On the other hand by right side of equation (16),

$$
\begin{aligned}
\left|\mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right| &= \left|\mathbf{Tr}\left[\widetilde{\boldsymbol{R}}\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right)\widetilde{\boldsymbol{R}}'\right]\right| \\
&= \left|\mathbf{Tr}\left[\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right)\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right]\right| \\
&\leq \lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\left|\sum_{\lambda(\cdot)\neq 0}\lambda\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right)\right| \text{(RHS of lemma 3)} \\
&\leq \lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\sum_{\lambda(\cdot)\neq 0}\left|\lambda\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right)\right| \\
&\leq \lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)|B_0|^{1/2}\left\{\sum_{\lambda(\cdot)\neq 0}\left|\lambda^2\left(\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right)\right|\right\}^{1/2} \\
&= \lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)|B_0|^{1/2}\left\|\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right\|_F \\
&\leq n^{\gamma}n^{(1+\alpha)\gamma/2}\left\|\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right\|_F . \qquad (31)
\end{aligned}
$$

Note that,

$$
\begin{aligned}
&\left\|\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right\|_F \\
&\quad= \sqrt{\sum_{i\in B_0}\sum_{j\in B_0}\left\{\widehat{\widetilde{\boldsymbol{\varphi}}}'_{gL,(i)}\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}-\widetilde{\boldsymbol{\varphi}}'_{(i)}\widetilde{\boldsymbol{\varphi}}_{(j)}\right\}^2} \\
&\quad= \sqrt{\sum_{i\in B_0}\sum_{j\in B_0}\left\{\widehat{\widetilde{\boldsymbol{\varphi}}}'_{gL,(i)}\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}-\widehat{\widetilde{\boldsymbol{\varphi}}}'_{gL,(i)}\widetilde{\boldsymbol{\varphi}}_{(j)}+\widehat{\widetilde{\boldsymbol{\varphi}}}'_{gL,(i)}\widetilde{\boldsymbol{\varphi}}_{(j)}-\widetilde{\boldsymbol{\varphi}}'_{(i)}\widetilde{\boldsymbol{\varphi}}_{(j)}\right\}^2} \\
&\quad= \sqrt{\sum_{i\in B_0}\sum_{j\in B_0}\left\{\widehat{\widetilde{\boldsymbol{\varphi}}}'_{gL,(i)}\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}-\widetilde{\boldsymbol{\varphi}}_{(j)}\right)+\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(i)}-\widetilde{\boldsymbol{\varphi}}_{(i)}\right)'\widetilde{\boldsymbol{\varphi}}_{(j)}\right\}^2} \\
&\quad\leq \sqrt{2\sum_{i\in B_0}\sum_{j\in B_0}\left\{\widehat{\widetilde{\boldsymbol{\varphi}}}'_{gL,(i)}\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}-\widetilde{\boldsymbol{\varphi}}_{(j)}\right)\right\}^2+\left\{\left(\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(i)}-\widetilde{\boldsymbol{\varphi}}_{(i)}\right)'\widetilde{\boldsymbol{\varphi}}_{(j)}\right\}^2}.
\end{aligned}
$$

Hence by CBS inequality we have,

$$\begin{aligned}
&\left\|\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0} - \widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right\|_F \\
&\quad \leq \sqrt{2\sum_{i\in B_0}\sum_{j\in B_0}\left(\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(i)}\right\|_2^2\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2 + \left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(i)} - \widetilde{\boldsymbol{\varphi}}_{(i)}\right\|_2^2\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2\right)} \\
&\quad = \sqrt{\left\{2\sum_{i\in B_0}\left(\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(i)}\right\|_2^2 + \left\|\widetilde{\boldsymbol{\varphi}}_{(i)}\right\|_2^2\right)\right\}\left\{\sum_{j\in B_0}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2\right\}} \\
&\quad = \sqrt{\left\{2\sum_{i\in B_0}\left(\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(i)} - \widetilde{\boldsymbol{\varphi}}_{(i)} + \widetilde{\boldsymbol{\varphi}}_{(i)}\right\|_2^2 + \left\|\widetilde{\boldsymbol{\varphi}}_{(i)}\right\|_2^2\right)\right\}\left\{\sum_{j\in B_0}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2\right\}} \\
&\quad \leq \sqrt{\left\{2\sum_{i\in B_0}\left(\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(i)} - \widetilde{\boldsymbol{\varphi}}_{(i)}\right\|_2^2 + \left\|\widetilde{\boldsymbol{\varphi}}_{(i)}\right\|_2^2 + \left\|\widetilde{\boldsymbol{\varphi}}_{(i)}\right\|_2^2\right)\right\}\left\{\sum_{j\in B_0}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2\right\}} \\
&\quad \leq G_2\sqrt{n^\gamma\varrho^2\phi_2^2 + \varrho_n^2\phi_2^2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 \\
&\quad \leq G_2 n^{\gamma/2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2, \qquad (32)
\end{aligned}$$

where, the last inequality holds only on $\mathscr{P}_0^N$, G_2 is a generic constant, and $\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 = \sum_{j\in B_0}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2$. Therefore combining equation (31) and (32),

$$\begin{aligned}
\left|\mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right)\right| &\leq G_2 n^\gamma n^{(1+\alpha)\gamma/2} n^{\gamma/2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 \\
&\leq G_2 n^{(4+\alpha)\gamma/2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 \leq G_2 n^{(5+\alpha)\gamma/2}
\end{aligned}$$

on $\mathscr{P}_n$, where the last inequality is due to assumptions **(A 1)**, **(A 2)** and G_2 is a generic constant. Hence we have equation (27). □

PROOF OF THEOREM 1. Note that since, $\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}, \hat{\sigma}^2$ is minimizer to (8) then,

$$\mathbf{Tr}(\boldsymbol{\Xi}_0\widehat{\boldsymbol{\Xi}}_{gL}^{-1}) + \log\det\widehat{\boldsymbol{\Xi}}_{gL} + \tau_n\sum_{j=1}^n\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2 \leq \mathbf{Tr}(\boldsymbol{\Xi}_0\boldsymbol{\Xi}^{-1}) + \log\det\boldsymbol{\Xi} + \tau_n\sum_{j=1}^n\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2,$$

and by definition of B_0 one can see precisely we need,

$$\mathbf{Tr}(\boldsymbol{\Xi}_0\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}) + \log\det\widehat{\boldsymbol{\Xi}}_{gL,B_0} + \tau_n\sum_{j\in B_0}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2 \leq \mathbf{Tr}(\boldsymbol{\Xi}_0\boldsymbol{\Xi}_{B_0}^{-1}) + \log\det\boldsymbol{\Xi}_{B_0} + \tau_n\sum_{j\in B_0}\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2.$$

By interchanging sides of the above inequality we have,

$$\textbf{Tr}\left(\boldsymbol{\Xi}_0\left[\widehat{\boldsymbol{\Xi}}^{-1}_{gL,B_0}-\boldsymbol{\Xi}^{-1}_{B_0}\right]\right) \quad\leq\quad \left(\log\frac{\det\boldsymbol{\Xi}_{B_0}}{\det\widehat{\boldsymbol{\Xi}}_{gL,B_0}}\right)-\tau_n\sum_{j\in B_0}\left(\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2-\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2\right).\text{(33)}$$

Now using the fact, $\log(x)$ is continuously differentiable on $[a,\infty)$ for any $a>0$, and hence Lipschitz continuous. Observing that $\lambda_{\min}\left(\boldsymbol{\Xi}_{B_0}\right)=\sigma^2$ and, $\lambda_{\min}\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right)\geq\hat{\sigma}^2>0$ therefore each eigen values of $\boldsymbol{\Xi}_{B_0}$, and, $\widehat{\boldsymbol{\Xi}}_{gL,B_0}$ is positive. Hence,

$$\begin{aligned}
&\left|\log\det\boldsymbol{\Xi}_{B_0}-\log\det\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right| \\
&\qquad=\left|\log\left(\prod_{\lambda(\cdot)\neq 0}\lambda\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right)\right)-\log\left(\prod_{\lambda(\cdot)\neq 0}\lambda\left(\boldsymbol{\Xi}_{B_0}\right)\right)\right| \\
&\qquad\leq\sum_{\lambda(\cdot)>0}\left|\log\left(\lambda\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right)\right)-\log\left(\lambda\left(\boldsymbol{\Xi}_{B_0}\right)\right)\right| \\
&\qquad\leq M\sum_{\lambda(\cdot)>0}\left|\lambda\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right)-\lambda\left(\boldsymbol{\Xi}_{B_0}\right)\right| \qquad\text{(By Lipschitz continuity)} \\
&\qquad\leq M\left\{\sum_{\lambda(\cdot)>0}\left|\lambda\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}\right)-\lambda\left(\boldsymbol{\Sigma}_{B_0}\right)\right|+n\left|\hat{\sigma}^2-\sigma^2\right|\right\} \\
&\qquad\leq M\left\{|B_0|\sum_{\lambda(\cdot)\neq 0}\left|\lambda\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right|+n\left|\hat{\sigma}^2-\sigma^2\right|\right\} \qquad\text{(By lemma 6)} \\
&\qquad\leq M|B_0|\left\{\sum_{\lambda(\cdot)\neq 0}\lambda^2\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right)\right\}^{1/2}+Mn\left|\hat{\sigma}^2-\sigma^2\right| \\
&\qquad=M|B_0|\left\|\widehat{\boldsymbol{\Sigma}}_{gL,B_0}-\boldsymbol{\Sigma}_{B_0}\right\|_F+Mn\left|\hat{\sigma}^2-\sigma^2\right| \\
&\qquad\leq M\left|B_0\right|\lambda_{\max}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\left\|\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL,B_0}\widehat{\widetilde{\boldsymbol{\Phi}}}'_{gL,B_0}-\widetilde{\boldsymbol{\Phi}}_{B_0}\widetilde{\boldsymbol{\Phi}}'_{B_0}\right\|_F+Mn\left|\hat{\sigma}^2-\sigma^2\right|, \qquad (34)
\end{aligned}$$

where, $M=\left(\hat{\sigma}^2\wedge\sigma^2\right)^{-1}$. On combining (32) with (34) on $\mathscr{P}_n$ we have,

$$\begin{aligned}
\left|\log\det\boldsymbol{\Xi}_{B_0}-\log\det\widehat{\boldsymbol{\Xi}}_{gL,B_0}\right| \quad&\leq\quad M\left|B_0\right|n^{\gamma}n^{\gamma/2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0}-\widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2+Mn\left|\hat{\sigma}^2-\sigma^2\right| \\
&\leq\quad M\mathscr{M}_n r_n n^{3\gamma/2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0}-\widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2+Mn\left|\hat{\sigma}^2-\sigma^2\right| \\
&\leq\quad Mn^{(5+2\alpha)\gamma/2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0}-\widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2+Mn\left|\hat{\sigma}^2-\sigma^2\right|. \quad (35)
\end{aligned}$$

And now using backward triangle inequality we have,

$$\begin{aligned}\left|\tau_n \sum_{j\in B_0}\left(\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2 - \left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2\right)\right| &\le \tau_n \sum_{j\in B_0}\left\|\widetilde{\boldsymbol{\varphi}}_{(j)} - \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2\\ &= \tau_n\sqrt{|B_0|}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 .\end{aligned}\tag{36}$$

Combining (33),(35),(36), and since on $\mathscr{P}_n$, $|B_0| \le \mathscr{M}_n r_n = Cn^{(1+\alpha)\gamma}$ we have,

$$\mathbf{Tr}\left(\Xi_0\left[\widehat{\Xi}^{-1}_{gL,B_0} - \Xi^{-1}_{B_0}\right]\right) \le \left(Mn^{\frac{(5+2\alpha)\gamma}{2}} + \tau_n n^{\frac{(1+\alpha)\gamma}{2}}\right)\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 + Mn\left|\hat{\sigma}^2 - \sigma^2\right|, \tag{37}$$

on $\mathscr{P}_n$. Now for two positive definite matrices, $\boldsymbol{A}$ and $\boldsymbol{B}$ let us observe the identity,

$$\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\left(\boldsymbol{A} - \boldsymbol{B}\right)\boldsymbol{B}^{-1}.$$

By choosing $\boldsymbol{A} = \widehat{\Xi}_{gL,B_0}$ and $\boldsymbol{B} = \Xi_{B_0}$ we have,

$$\begin{aligned}&\left|\mathbf{Tr}\left(\Xi_0\left[\Xi^{-1}_{B_0} - \widehat{\Xi}^{-1}_{gL,B_0}\right]\right)\right|\\ &= \left|n\,\mathbf{Tr}\left(\frac{\Xi_0}{n}\left[\widehat{\Xi}^{-1}_{gL,B_0}\left(\widehat{\Xi}_{gL,B_0} - \Xi_{B_0}\right)\Xi^{-1}_{B_0}\right]\right)\right|\\ &= \left|n\left[vec\left(\frac{\Xi_0}{n}\right)\right]'\left(\Xi^{-1}_{B_0}\otimes\widehat{\Xi}^{-1}_{gLB_0}\right)vec\left(\widehat{\Xi}_{gL,B_0} - \Xi_{B_0}\right)\right|,\end{aligned}\tag{38}$$

where, $\otimes$ is used to denote Kronecker product. The last identity (38) is due to, $\mathbf{Tr}(\boldsymbol{V}_1\boldsymbol{V}_2\boldsymbol{V}_3\boldsymbol{V}_4) = [vec(\boldsymbol{V}_1')]'\left(\boldsymbol{V}_4'\otimes\boldsymbol{V}_2\right)[vec(\boldsymbol{V}_3)]$ (Theorem 8.12, Schott (2005)) with $\boldsymbol{V}_1 = \Xi_0/n$, $\boldsymbol{V}_2 = \widehat{\Xi}^{-1}_{gL,B_0}$, $\boldsymbol{V}_3 = \widehat{\Xi}_{gL,B_0} - \Xi_{B_0}$, and $\boldsymbol{V}_4 = \Xi^{-1}_{B_0}$. Since $\Xi_0/n - \lambda_{\min}\left(\Xi_0/n\right)\mathbb{I}$, $\widehat{\Xi}^{-1}_{gL,B_0}$, and $\Xi^{-1}_{B_0}$ are positive definite matrices,

$$\begin{aligned}&\left|\mathbf{Tr}\left(\Xi_0\left[\Xi^{-1}_{B_0} - \widehat{\Xi}^{-1}_{gL,B_0}\right]\right)\right|\\ &\ge \frac{\lambda_{\min}\left(\frac{\Xi_0}{n}\right)}{\lambda_{\max}\left(\widehat{\Xi}_{gL,B_0}\right)\lambda_{\max}\left(\Xi_{B_0}\right)}\left|n\left[vec\left(\mathbb{I}\right)\right]'\left(\mathbb{I}\otimes\mathbb{I}\right)vec\left(\widehat{\Xi}_{gL,B_0} - \Xi_{B_0}\right)\right|\\ &= \boldsymbol{\Lambda}_n\left|n\,\mathbf{Tr}\left(\widehat{\Xi}_{gL,B_0} - \Xi_{B_0}\right)\right|,\end{aligned}\tag{39}$$

where $\boldsymbol{\Lambda}_n$ is defined according as lemma 7. The last identity is obtained by using Theorem 8.12, Schott (2005) with $\boldsymbol{V}_1 = \boldsymbol{V}_2 = \boldsymbol{V}_4 = \mathbb{I}$ and $\boldsymbol{V}_3 = \widehat{\Xi}_{gL,B_0} - \Xi_{B_0}$. Note that on $\mathscr{P}_n$, $\mathscr{B}_1 = \varsigma \le |\hat{\sigma}^2 - \sigma^2| \le \varsigma_n = \mathscr{B}_2$, also by lemma 8 on $\mathscr{P}_n$,

$$\mathscr{A}_1 = G_1 n^{\gamma} \le \mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right) \le G_2 n^{(5+\alpha)\gamma/2} = \mathscr{A}_2.$$

If we choose ς_n such that $\mathscr{B}_2 = \varsigma_n < G_1 n^{\gamma} = \mathscr{A}_1$, by using restricted reverse triangle inequality in ℓ_1- norm (lemma 1) with $(a,b) = \left(n\,\mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right), n^2\left(\hat{\sigma}^2 - \sigma^2\right)\right)$ and a sutiable choice of $k_n^{-1} = \boldsymbol{O}(\mathscr{A}_2)$, we get a lower bound for the right hand side of (39),

$$\begin{aligned}
\boldsymbol{\Lambda}_n \left| n\,\mathbf{Tr}\left(\widehat{\boldsymbol{\Xi}}_{gL,B_0} - \boldsymbol{\Xi}_{B_0}\right)\right| &= \boldsymbol{\Lambda}_n \left| n\,\mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right) + n^2\left(\hat{\sigma}^2-\sigma^2\right)\right| \\
&\geq k_n\boldsymbol{\Lambda}_n\left\{n\left|\mathbf{Tr}\left(\widehat{\boldsymbol{\Sigma}}_{gL,B_0} - \boldsymbol{\Sigma}_{B_0}\right)\right| + n^2\left|\hat{\sigma}^2-\sigma^2\right|\right\} \\
&\geq k_n\boldsymbol{\Lambda}_n\left\{n^{\gamma+1}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + n^2\left|\hat{\sigma}^2-\sigma^2\right|\right\},
\end{aligned} \tag{40}$$

where the last inequality is based on lemma 8. Combining equations (40), (38), (37) and assumption **(A 1)** we have,

$$\begin{aligned}
& k_n\boldsymbol{\Lambda}_n n^{\gamma+1}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + k_n\boldsymbol{\Lambda}_n n^2\left|\hat{\sigma}^2-\sigma^2\right| \\
&\leq \left(Mn^{\frac{(5+2\alpha)\gamma}{2}} + \tau_n n^{\frac{(1+\alpha)\gamma}{2}}\right)\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 + Mn\left|\hat{\sigma}^2-\sigma^2\right| \\
&= \frac{\left(Mn^{\frac{(5+2\alpha)\gamma}{2}} + \tau_n n^{\frac{(1+\alpha)\gamma}{2}}\right)}{\sqrt{k_n\boldsymbol{\Lambda}_n n^{\gamma+1}}}\sqrt{k_n\boldsymbol{\Lambda}_n n^{\gamma+1}}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 + Mn\left|\hat{\sigma}^2-\sigma^2\right| \\
&\leq \frac{\left(Mn^{\frac{(5+2\alpha)\gamma}{2}} + \tau_n n^{\frac{(1+\alpha)\gamma}{2}}\right)^2}{2k_n\boldsymbol{\Lambda}_n n^{\gamma+1}} + \frac{k_n\boldsymbol{\Lambda}_n n^{\gamma+1}}{2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + Mn\left|\hat{\sigma}^2-\sigma^2\right|,
\end{aligned} \tag{41}$$

where the last inequality is based on $2ab \leq a^2 + b^2$. If we choose $\tau_n < Cn^{(4+\alpha)\gamma/2}$,

$$\begin{aligned}
\frac{k_n\boldsymbol{\Lambda}_n n^{\gamma+1}}{2}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + n^2k_n\boldsymbol{\Lambda}_n\left(1 - M\frac{k_n^{-1}\boldsymbol{\Lambda}_n^{-1}}{n}\right)\left|\hat{\sigma}^2-\sigma^2\right| &\leq \frac{2n^{(5+2\alpha)\gamma}}{k_n\boldsymbol{\Lambda}_n n^{\gamma+1}} \\
\text{or, } n^{\gamma}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + 2n\left(1 - M\frac{k_n^{-1}\boldsymbol{\Lambda}_n^{-1}}{n}\right)\left|\hat{\sigma}^2-\sigma^2\right| &\leq \frac{4n^{(5+2\alpha)\gamma}}{k_n^2\boldsymbol{\Lambda}_n^2 n^{\gamma+2}}.
\end{aligned}$$

Finally using $1/\boldsymbol{\Lambda}_n \leq \boldsymbol{O}_p\left(n^{(3+4\alpha)\gamma}\right)$ (by lemma 7) and $k_n^{-1} = \boldsymbol{O}\left(n^{(5+\alpha)\gamma/2}\right)$,

$$\begin{aligned}
\frac{k_n^2\boldsymbol{\Lambda}_n^2 n^{\gamma+2}}{n^{(5+2\alpha)\gamma}} &\geq \boldsymbol{O}_p\left(\frac{n^2}{n^{(15+11\alpha)\gamma}}\right) \\
&= \boldsymbol{O}_p\left(n^{2-(15+11\alpha)\gamma}\right) \uparrow \infty,
\end{aligned} \tag{42}$$

since $\gamma < 2/(15+11\alpha)$. Also,

$$\begin{aligned}
\frac{k_n^{-1}\boldsymbol{\Lambda}_n^{-1}}{n} &\leq \boldsymbol{O}_p\left(\frac{n^{(3+4\alpha)\gamma}n^{(5+\alpha)\gamma/2}}{n}\right) \\
&= \boldsymbol{O}_p\left(\frac{n^{(11+9\alpha)\gamma/2}}{n}\right) < \boldsymbol{O}_p\left(\frac{n^{\frac{11+9\alpha}{15+11\alpha}}}{n}\right) \downarrow 0,
\end{aligned}$$

since $11 + 9\alpha < 15 + 11\alpha$. Therefore,

$$n^{\gamma}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + 2n\left|\hat{\sigma}^2-\sigma^2\right| \leq \boldsymbol{O}_p\left(\frac{1}{n^{2-(15+11\alpha)\gamma}}\right). \tag{43}$$

Under assumption **(A 1)**, $2-(15+11\alpha)\gamma > 0$. So if possible, let us assume that there exists $\delta > 0$, such that for large enough N_δ, $\forall\ n \geq N_\delta$, $\mathbb{P}(\mathscr{P}_n) > \delta$ and hence we have, $\forall\ n \geq N_\delta$, $\delta < \mathbb{P}(\mathscr{P}_n) < \mathbb{P}(\mathscr{P}_{\infty,n})$. So on the set $\mathscr{P}_{\infty,n}$,

$$\left\{n^{\frac{\gamma}{2}} \left\| \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0} \right\|_2 \right\}^2 + 2\left\{n^{\frac{1}{2}} \left|\hat{\sigma}^2 - \sigma^2\right|^{\frac{1}{2}}\right\}^2 \quad = \quad \boldsymbol{O}_p\left(\frac{1}{n^{2-(15+11\alpha)\gamma}}\right).$$

$i.e.$, both $\left\{n^{\frac{\gamma}{2}} \left\| \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0} \right\|_2 \right\}^2$ and, $\left\{n^{\frac{1}{2}} \left|\hat{\sigma}^2 - \sigma^2\right|^{\frac{1}{2}}\right\}^2$ goes to 0, which is a contradiction. Hence our assumption, that there exists $n \geq N_\delta$ such that $\mathbb{P}(\mathscr{P}_{\infty,n}) > \delta$ is violated and we can conclude that, $\mathbb{P}(\mathscr{P}^c_{\infty,n}) \overset{n\to\infty}{\longrightarrow} 1$. □

PROOF OF THEOREM 2. Note that,

$$\begin{aligned}
&\left\| \widehat{\boldsymbol{\Xi}}_{gL,B_0} - \boldsymbol{\Xi}_{B_0} \right\|_F \\
&\quad = \left\{ \sum_{\lambda \neq 0} \lambda^2 \left(\widehat{\boldsymbol{\Xi}}_{gL,B_0} - \boldsymbol{\Xi}_{B_0} \right) \right\}^{1/2} \\
&\quad \leq \sum_{\lambda \neq 0} \left| \lambda \left(\widehat{\boldsymbol{\Xi}}_{gL,B_0} - \boldsymbol{\Xi}_{B_0} \right) \right| \\
&\quad \leq \left\{ \sum_{\lambda \neq 0} \left| \lambda \left(\boldsymbol{\Sigma}_{B_0} - \widehat{\boldsymbol{\Sigma}}_{gL,B_0} \right) \right| + n \left| \hat{\sigma}^2 - \sigma^2 \right| \right\} \\
&\quad \leq |B_0| \, \lambda_{\max}\left(\widetilde{\boldsymbol{R}}' \widetilde{\boldsymbol{R}} \right) \left\| \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0} \right\|_2 + n \left| \hat{\sigma}^2 - \sigma^2 \right|. \qquad \text{(By (34) and (32))}
\end{aligned}$$

Therefore, again by using $(a+b)^2 \leq 2a^2 + 2b^2$,

$$\begin{aligned}
\frac{1}{n^2} \left\| \widehat{\boldsymbol{\Xi}}_{gL,B_0} - \boldsymbol{\Xi}_{B_0} \right\|_F^2 &\leq \frac{1}{n^2} \left(|B_0|^2 \, \lambda^2_{\max}\left(\widetilde{\boldsymbol{R}}' \widetilde{\boldsymbol{R}} \right) \left\| \widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0} \right\|_2^2 + n^2 \left| \hat{\sigma}^2 - \sigma^2 \right|^2 \right) \\
&\leq \boldsymbol{O}_p \left(\frac{|B_0|^3 \, \lambda^2_{\max}\left(\widetilde{\boldsymbol{R}}' \widetilde{\boldsymbol{R}} \right) \varrho^2}{n^2} + \varsigma^2 \right) \text{ (By theorem 1)} \\
&\leq \boldsymbol{O}_p \left(\frac{\mathscr{M}^3 n^{5\gamma} \varrho^2}{n^2} + \varsigma^2 \right).
\end{aligned}$$

The above probability bound is achieved using the fact $\mathbb{P}(\mathscr{P}^c_\infty) \longrightarrow 1$ from theorem 1. Since, $\varsigma > 0$ can be any preassigned positive number and since since $\gamma < 2/(15+11\alpha)$, if we choose $n^2\varsigma^2 < \mathscr{M}^3 n^{10/(15+11\alpha)}$,

$$\frac{1}{n^2} \left\| \widehat{\boldsymbol{\Xi}}_{gL,B_0} - \boldsymbol{\Xi}_{B_0} \right\|_F^2 \quad = \quad \boldsymbol{O}_p \left(\frac{\mathscr{M}^3 \varrho^2}{n^{2\left(1 - \frac{5}{15+11\alpha}\right)}} \right) = \boldsymbol{O}_p \left(\frac{\mathscr{M}^3 \varrho^2}{n^{2\left(\frac{10+11\alpha}{15+11\alpha}\right)}} \right), \tag{44}$$

and hence we have theorem 2. □

PROOF OF THEOREM 3. Note that since, $\widehat{\widetilde{\boldsymbol{\Phi}}}_{gL}, \hat{\sigma}^2, \& \widehat{\boldsymbol{\beta}}$ is minimizer to (10) then,

$$\mathbf{Tr}\left(\boldsymbol{\Xi}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Xi}}_{gL}^{-1}\right) + \log\det\widehat{\boldsymbol{\Xi}}_{gL} + \tau_n\sum_{j=1}^{n}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2 \leq \mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\boldsymbol{\Xi}^{-1}\right) + \log\det\boldsymbol{\Xi} + \tau_n\sum_{j=1}^{n}\|\widetilde{\boldsymbol{\varphi}}_{(j)}\|_2,$$

and by definition of B_0 one can see precisely we need,

$$\mathbf{Tr}\left(\boldsymbol{\Xi}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\right)+\log\det\widehat{\boldsymbol{\Xi}}_{gL,B_0}+\tau_n\sum_{j\in B_0}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)}\right\|_2 \leq \mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\boldsymbol{\Xi}_{B_0}^{-1}\right)+\log\det\boldsymbol{\Xi}_{B_0}+\tau_n\sum_{j\in B_0}\left\|\widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2.$$

Obeserve that,

$$\begin{aligned}
&\mathbf{Tr}\left(\boldsymbol{\Xi}_{\widehat{\boldsymbol{\beta}}}\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\right)\\
&= \left(\boldsymbol{X}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{X}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)\\
&= \left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\beta}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{\beta}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}}\right)\\
&= \left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}\right) - \left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)\\
&- \left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}\right) + \left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)\\
&= \mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\right) - \left\{\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1/2}\left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}\right)\right\}'\left\{\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1/2}\boldsymbol{Z}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}\\
&- \left\{\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1/2}\boldsymbol{Z}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}'\left\{\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1/2}\left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}\right)\right\} + \left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)\\
&= \mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\right) - \boldsymbol{\xi}_{B_0}'\boldsymbol{\eta}_{B_0} - \boldsymbol{\eta}_{B_0}'\boldsymbol{\xi}_{B_0} + \left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right), \qquad (45)
\end{aligned}$$

by defining $\boldsymbol{\xi}_{B_0} = \left\{\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1/2}\left(\boldsymbol{X}-\boldsymbol{Z}\boldsymbol{\beta}\right)\right\}$ and, $\boldsymbol{\eta}_{B_0} = \left\{\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1/2}\boldsymbol{Z}\left(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}\right)\right\}$. Hence, changing sides and combining equations (45) and (36),

$$\begin{aligned}
&\mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\left[\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}-\boldsymbol{\Xi}_{B_0}^{-1}\right]\right) + \left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}_{gL,B_0}^{-1}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right)\\
&\leq \left(\log\frac{\det\boldsymbol{\Xi}_{B_0}}{\det\widehat{\boldsymbol{\Xi}}_{gL,B_0}}\right) + \tau_n\sqrt{|B_0|}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0}-\widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 + \boldsymbol{\xi}_{B_0}'\boldsymbol{\eta}_{B_0} + \boldsymbol{\eta}_{B_0}'\boldsymbol{\xi}_{B_0}. \qquad (46)
\end{aligned}$$

Now,

$$\begin{aligned}
\boldsymbol{\eta}_{B_0}'\boldsymbol{\xi}_{B_0} = \boldsymbol{\xi}_{B_0}'\boldsymbol{\eta}_{B_0} \quad &\leq \quad \left(\boldsymbol{\xi}_{B_0}'\boldsymbol{\xi}_{B_0}\right)^{\frac{1}{2}}\left(\boldsymbol{\eta}_{B_0}'\boldsymbol{\eta}_{B_0}\right)^{\frac{1}{2}} \qquad \text{[By CBS Inequality]}\\
&= \quad 2\left(\boldsymbol{\xi}_{B_0}'\boldsymbol{\xi}_{B_0}\right)^{\frac{1}{2}}\frac{\left(\boldsymbol{\eta}_{B_0}'\boldsymbol{\eta}_{B_0}\right)^{\frac{1}{2}}}{2}\\
&\leq \quad \left(\boldsymbol{\xi}_{B_0}'\boldsymbol{\xi}_{B_0}\right) + \frac{\left(\boldsymbol{\eta}_{B_0}'\boldsymbol{\eta}_{B_0}\right)}{4}. \qquad (47)
\end{aligned}$$

The last inequality is based on $2ab \leq a^2 + b^2$. Now using the definition of $\boldsymbol{\xi}_{B_0}$ and $\boldsymbol{\eta}_{B_0}$, we have, $\boldsymbol{\xi}'_{B_0}\boldsymbol{\xi}_{B_0} = \mathbf{Tr}\left(\Xi_{\boldsymbol{\beta}}\widehat{\Xi}^{-1}_{gL,B_0}\right)$ and $\boldsymbol{\eta}'_{B_0}\boldsymbol{\eta}_{B_0} = \left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right)' \widehat{\Xi}^{-1}_{gL,B_0}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right)$ and also using (35) and (46) on $\mathscr{P}_n$ we have,

$$\begin{aligned}
\mathbf{Tr}\left(\Xi_{\boldsymbol{\beta}}\left[\widehat{\Xi}^{-1}_{gL,B_0} - \Xi^{-1}_{B_0}\right]\right) &+ \frac{1}{2}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right)' \widehat{\Xi}^{-1}_{gL,B_0}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right) \\
&\leq \left(Mn^{\frac{(5+2\alpha)\gamma}{2}} + \tau_n n^{\frac{(1+\alpha)\gamma}{2}}\right)\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2 \\
&+ Mn\left|\hat{\sigma}^2 - \sigma^2\right| + 2\,\mathbf{Tr}\left(\Xi_{\boldsymbol{\beta}}\widehat{\Xi}^{-1}_{gL,B_0}\right). \qquad (48)
\end{aligned}$$

By recalling **Sherman-Morisson-Woddbury** matrix indentiy on $\widehat{\Xi}_{gL,B_0} = \hat{\sigma}^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widehat{\widetilde{\boldsymbol{\Omega}}}_{gL,B_0}\widetilde{\boldsymbol{R}}'$ and, defining $\boldsymbol{F} = \hat{\sigma}^2\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0} + \widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}$, we learn that,

$$\begin{aligned}
\widehat{\Xi}^{-1}_{gL,B_0} &= \left(\hat{\sigma}^2\mathbb{I} + \widetilde{\boldsymbol{R}}\widehat{\widetilde{\boldsymbol{\Omega}}}_{gL,B_0}\widetilde{\boldsymbol{R}}'\right)^{-1} = \frac{1}{\hat{\sigma}^2}\left(\mathbb{I} + \widetilde{\boldsymbol{R}}\frac{\widehat{\widetilde{\boldsymbol{\Omega}}}_{gL,B_0}}{\hat{\sigma}^2}\widetilde{\boldsymbol{R}}'\right)^{-1} \\
&= \frac{1}{\hat{\sigma}^2}\left[\mathbb{I} - \widetilde{\boldsymbol{R}}\left(\hat{\sigma}^2\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0} + \widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)^{-1}\widetilde{\boldsymbol{R}}'\right] \overset{def}{=} \frac{1}{\hat{\sigma}^2}\left(\mathbb{I} - \widetilde{\boldsymbol{R}}\boldsymbol{F}^{-1}\widetilde{\boldsymbol{R}}'\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbf{Tr}\left(\widehat{\Xi}^{-1}_{gL,B_0}\right) &= \frac{1}{\hat{\sigma}^2}\left\{n - \mathbf{Tr}\left(\widetilde{\boldsymbol{R}}\boldsymbol{F}^{-1}\widetilde{\boldsymbol{R}}'\right)\right\} \\
&= \frac{1}{\hat{\sigma}^2}\left\{n - \mathbf{Tr}\left(\boldsymbol{F}^{-1}\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right)\right\} \\
&= \frac{1}{\hat{\sigma}^2}\left\{n - \mathbf{Tr}\left[\boldsymbol{F}^{-1}\left(\boldsymbol{F} - \hat{\sigma}^2\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0}\right)\right]\right\} \\
&= \frac{1}{\hat{\sigma}^2}\left\{n - n + \hat{\sigma}^2\,\mathbf{Tr}\left[\boldsymbol{F}^{-1}\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0}\right]\right\} \\
&= \mathbf{Tr}\left[\boldsymbol{F}^{-1}\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0}\right] \leq \lambda_{\max}\left(\boldsymbol{F}^{-1}\right)\mathbf{Tr}\left[\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0}\right] \text{(RHS of lemma 3)} \\
&\leq n^{-\gamma}\,\mathbf{Tr}\left[\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0}\right],
\end{aligned}$$

where the last inequality is based on the fact $\lambda_{\max}\left(\boldsymbol{F}^{-1}\right) = \lambda^{-1}_{\min}\left(\boldsymbol{F}\right) = \lambda^{-1}_{\min}\left(\widetilde{\boldsymbol{R}}'\widetilde{\boldsymbol{R}}\right) \leq Cn^{-\gamma}$, by assumption **(A 2)**. Hence using RHS of lemma 3,

$$
\begin{aligned}
\boldsymbol{\xi}'_{B_0}\boldsymbol{\xi}_{B_0} = \mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\Xi}}^{-1}_{gL,B_0}\right) &\leq Cn^{-\gamma}\lambda_{\max}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\right)\mathbf{Tr}\left(\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0}\right) \\
&= Cn^{-\gamma}\lambda_{\max}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\right)\sum_{\lambda(\cdot)>0}\lambda\left(\widehat{\widetilde{\boldsymbol{\Omega}}}^{-1}_{gL,B_0}\right) \\
&= Cn^{-\gamma}\lambda_{\max}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\right)\sum_{\lambda(\cdot)>0}\left\{\lambda\left(\widehat{\widetilde{\boldsymbol{\Omega}}}_{gL,B_0}\right)\right\}^{-1} \\
&\leq Cn^{-\gamma}\sum_{j\in B_0}\left\{\widehat{\widetilde{\varphi}}^{2}_{gL,jj}\right\}^{-1} \\
&\leq Cn^{-\gamma}|B_0|\max_{j\in B_0}\left\{\widehat{\widetilde{\varphi}}^{2}_{gL,jj}\right\}^{-1} \\
&= Cn^{-\gamma}|B_0|\left\{\min_{j\in B_0}\widehat{\widetilde{\varphi}}^{2}_{gL,jj}\right\}^{-1}, \qquad (49)
\end{aligned}
$$

where C is a generic constant. Note that on $\mathscr{P}_n$,

$$
\left|\widehat{\widetilde{\varphi}}_{gL,jj}\right| = \left|\widehat{\widetilde{\varphi}}_{gL,jj} - \widetilde{\varphi}_{jj} + \widetilde{\varphi}_{jj}\right| \geq C\left(\left|\widehat{\widetilde{\varphi}}_{gL,jj} - \widetilde{\varphi}_{jj}\right| + |\widetilde{\varphi}_{jj}|\right) \geq C\left|\widehat{\widetilde{\varphi}}_{gL,jj} - \widetilde{\varphi}_{jj}\right|,
$$

where the first inequality is obtained using lemma 1 with $a = \widehat{\widetilde{\varphi}}_{gL,jj} - \widetilde{\varphi}_{jj}$ and $b = \widetilde{\varphi}_{jj}$. The choice of constants, $\mathscr{B}_1$ and $\mathscr{B}_2$ can be obtained if we recall that the elements of the lower triangular matrix $\widetilde{\boldsymbol{\Phi}}$ belongs to a bounded set $\mathscr{P}_0 \subset \boldsymbol{P}_1$,

$$
\mathscr{B}_1 = \phi_1 < |\widetilde{\varphi}_{jj}| < \phi_2 = \mathscr{B}_2.
$$

One the other hand, the choice of constants, $\mathscr{A}_1$ and $\mathscr{A}_2$ can be obtained on $\mathscr{P}_n$ using the fact that $\widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Omega}}_{B_0}$ is diagonally dominant,

$$
\mathscr{A}_1 = C_1\varrho < \left|\widehat{\widetilde{\varphi}}_{gL,jj} - \widetilde{\varphi}_{jj}\right| < C_2 n^{\alpha\gamma+\frac{1}{2}} = \mathscr{A}_2.
$$

Although we need to choose these constants such that $\mathscr{B}_2 < \mathscr{A}_1$, which can be attained by using, . Therefore,

$$
\begin{aligned}
\left|\widehat{\widetilde{\varphi}}_{gL,jj}\right|^2 &\geq C\left|\widehat{\widetilde{\varphi}}_{gL,jj} - \widetilde{\varphi}_{jj}\right|^2 \\
&\geq C\left\{\sum_{i=1}^{j}\left|\widehat{\widetilde{\varphi}}_{gL,ji} - \widetilde{\varphi}_{ji}\right|\right\}^2 \Big(\text{ Since, } \widehat{\nabla}_{gL}\widetilde{\boldsymbol{\Omega}}_{B_0} \text{ is diagonally dominant }\Big) \\
&\geq C\sum_{i=1}^{j}\left|\widehat{\widetilde{\varphi}}_{gL,ji} - \widetilde{\varphi}_{ji}\right|^2 = C\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,(j)} - \widetilde{\boldsymbol{\varphi}}_{(j)}\right\|_2^2 \geq C\varrho^2. \qquad (50)
\end{aligned}
$$

Hence combining (49) and (50) on $\mathscr{P}_n$ we have,

$$\boldsymbol{\xi}'_{B_0}\boldsymbol{\xi}_{B_0} = \mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\widehat{\boldsymbol{\Xi}}^{-1}_{gL,B_0}\right) \ \le \ Cn^{-\gamma}\mathscr{M}_n r_n \le n^{\alpha\gamma}. \tag{51}$$

Also observe,

$$\begin{aligned}
\frac{1}{2}\boldsymbol{\eta}'_{B_0}\boldsymbol{\eta}_{B_0} &= \frac{1}{2}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Xi}}^{-1}_{gL,B_0}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right) \\
&= \frac{n}{2}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right)'\left[\frac{\widehat{\boldsymbol{\Xi}}^{-1}_{gL,B_0}}{n}\right]\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right) \\
&= \frac{n}{2}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right)'\widehat{\boldsymbol{\Delta}}_{gL,B_0}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right) \\
&\ge \lambda_{\min}\left(\widehat{\boldsymbol{\Delta}}_{gL,B_0}\right)\frac{n}{2}\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right)'\left(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right) \\
&= \lambda_{\min}\left(\widehat{\boldsymbol{\Delta}}_{gL,B_0}\right)\frac{n}{2}\left\|\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right\|_2^2,
\end{aligned} \tag{52}$$

where $\widehat{\boldsymbol{\Delta}}_{gL,B_0} = \widehat{\boldsymbol{\Xi}}^{-1}_{gL,B_0}/n$. Finally from what we learned in theorem 1 on the set $\mathscr{P}_\infty$, we have,

$$\left|\mathbf{Tr}\left(\boldsymbol{\Xi}_{\boldsymbol{\beta}}\left[\boldsymbol{\Xi}^{-1}_{B_0} - \widehat{\boldsymbol{\Xi}}^{-1}_{gL,B_0}\right]\right)\right| \ge k_n\boldsymbol{\Lambda}^{\boldsymbol{\beta}}_n\left\{n^{\gamma+1}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + n^2\left|\hat{\sigma}^2 - \sigma^2\right|\right\}. \tag{53}$$

Now recall by lemma 7 on $\mathscr{P}_n$, $\boldsymbol{\Lambda}^{\boldsymbol{\beta}}_n \ge \boldsymbol{O}_p\left(n^{-(3+4\alpha)\gamma}\right)$ and using $k_n^{-1} = \boldsymbol{O}(n^{(5+\alpha)\gamma/2})$ we have the following,

$$k_n\boldsymbol{\Lambda}^{\boldsymbol{\beta}}_n \ge \boldsymbol{O}_p\left(n^{-(3+4\alpha)\gamma}n^{-(5+\alpha)\gamma/2}\right) = \boldsymbol{O}_p\left(n^{-(11+9\alpha)\gamma/2}\right) \ge \boldsymbol{O}_p\left(n^{-\frac{11+9\alpha}{15+11\alpha}}\right),$$

and now if we follows the steps of equation (41) in theorem 1 on $\mathscr{P}_n$ we obtain,

$$\begin{aligned}
&n^{\gamma}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + n\left(1 - M\frac{n^{\frac{11+9\alpha}{15+11\alpha}}}{n}\right)\left|\hat{\sigma}^2 - \sigma^2\right| \\
&+\lambda_{\min}\left(\widehat{\boldsymbol{\Delta}}_{gL,B_0}\right)\frac{n^{\frac{11+9\alpha}{15+11\alpha}}}{2}\left\|\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right\|_2^2 \le \boldsymbol{O}_p\left(\frac{2Mn^{(5+2\alpha)\gamma}}{n^{-\frac{22+18\alpha}{15+11\alpha}}n^{\gamma+2}} + \frac{n^{\alpha\gamma}}{n^{-\frac{11+9\alpha}{15+11\alpha}}n}\right),
\end{aligned} \tag{54}$$

if we choose $\tau_n < Cn^{(4+\alpha)\gamma/2}$. Since $11 + 9\alpha < 15 + 9\alpha$, $n^{\frac{11+9\alpha}{15+11\alpha}}/n \longrightarrow 0$ we have,

$$\begin{aligned}
&n^{\gamma}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + n\left|\hat{\sigma}^2 - \sigma^2\right| + \frac{n^{\frac{11+9\alpha}{15+11\alpha}}}{2}\lambda_{\min}\left(\widehat{\boldsymbol{\Delta}}_{gL,B_0}\right)\left\|\boldsymbol{Z}\widehat{\boldsymbol{\beta}} - \boldsymbol{Z}\boldsymbol{\beta}\right\|_2^2 \\
&\le \boldsymbol{O}_p\left(\frac{2Mn^{(4+2\alpha)\gamma}n^{\frac{22+18\alpha}{15+11\alpha}}}{n^2} + \frac{n^{\alpha\gamma}n^{\frac{11+9\alpha}{15+11\alpha}}}{n}\right).
\end{aligned} \tag{55}$$

If possible, let us assume that there exists $\delta > 0$, such that for large enough N_δ, $\forall\ n \ge N_\delta$, $\mathbb{P}(\mathscr{P}_n) > \delta$ and hence we have, $\forall\ n \ge N_\delta$, $\delta < \mathbb{P}(\mathscr{P}_n) < \mathbb{P}(\mathscr{P}_{\infty,n})$. So on $\mathscr{P}_{\infty,n}$,

$$n^{\gamma}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0} - \widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2^2 + 2n\left|\hat{\sigma}^2 - \sigma^2\right| = \boldsymbol{O}_p\left(\frac{n^{(4+2\alpha)\gamma}n^{\frac{22+18\alpha}{15+11\alpha}}}{n^2} + \frac{n^{\alpha\gamma}n^{\frac{11+9\alpha}{15+11\alpha}}}{n}\right),$$

and,

$$\frac{n^{\frac{11+9\alpha}{15+11\alpha}}}{2}\lambda_{\min}\left(\widehat{\boldsymbol{\Delta}}_{gL,B_0}\right)\left\|\boldsymbol{Z}\widehat{\boldsymbol{\beta}}-\boldsymbol{Z}\boldsymbol{\beta}\right\|_2^2=\boldsymbol{O}_p\left(\frac{n^{(4+2\alpha)\gamma}n^{\frac{22+18\alpha}{15+11\alpha}}}{n^2}+\frac{n^{\alpha\gamma}n^{\frac{11+9\alpha}{15+11\alpha}}}{n}\right).$$

Since, $\gamma < 2/(15+11\alpha)$,

$$\boldsymbol{O}_p\left(\frac{n^{(4+2\alpha)\gamma}n^{\frac{22+18\alpha}{15+11\alpha}}}{n^2}\right)<\boldsymbol{O}_p\left(\frac{n^{(4+2\alpha)\frac{2}{15+11\alpha}}n^{\frac{22+18\alpha}{15+11\alpha}}}{n^2}\right)\downarrow 0,$$

and,

$$\boldsymbol{O}_p\left(\frac{n^{\alpha\gamma}n^{\frac{22+18\alpha}{15+11\alpha}}}{n}\right)<\boldsymbol{O}_p\left(\frac{n^{\alpha\frac{2}{15+11\alpha}}n^{\frac{11+9\alpha}{15+11\alpha}}}{n}\right)=\boldsymbol{O}_p\left(\frac{n^{\frac{11+11\alpha}{15+11\alpha}}}{n}\right)\downarrow 0.$$

$i.e.$, both $\left\{n^{\frac{\gamma}{2}}\left\|\widehat{\widetilde{\boldsymbol{\varphi}}}_{gL,B_0}-\widetilde{\boldsymbol{\varphi}}_{B_0}\right\|_2\right\}^2$ and, $\left\{n^{\frac{1}{2}}\left|\hat{\sigma}^2-\sigma^2\right|^{\frac{1}{2}}\right\}^2$ goes to 0, which is a contradiction. Hence our assumption, that there exists $n \geq N_\delta$ such that $\mathbb{P}(\mathscr{P}_{\infty,n}) > \delta$ is violated and we can conclude that, $\mathbb{P}(\mathscr{P}^c_{\infty,n}) \overset{n\to\infty}{\longrightarrow} 1$. Hence due to (42) we have theorem 3. □

References

[1] Achlioptas, D., McSherry, F. (2007), “Fast computation of low-rank matrix approximations,” *Journal of the ACM*, 54(2), 9.

[2] Arbenz, P., Drmac, Z. (2002), “On positive semidefinite matrices with known null space,” *SIAM Journal on Matrix Analysis and Applications*, 24(1), 132-149.

[3] Bai, Z. D., Yin, Y. Q. (1993), “Limit of the smallest eigenvalue of a large dimensional sample covariance matrix,” *The annals of Probability*, 1275-1294.

[4] Banerjee, A., Dunson, D. B., Tokdar, S. T. (2012), “Efficient Gaussian process regression for large datasets,” *Biometrika*, ass068.

[5] Bühlmann, P., Van De Geer, S. (2011), “Statistics for high-dimensional data: methods, theory and applications,” Springer Science Business Media.

[6] Cressie, N., (1990), “The origins of kriging,” *Mathematical geology*, 22(3), 239-252.

[7] Cressie, N. (1993), “Statistics for spatial data,” *Wiley series in probability and statistics. Wiley-Interscience New York*, 15, 16.

[8] Cressie, N., Johannesson, G. (2008), “Fixed rank kriging for very large spatial data sets,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1), 209-226.

[9] Datta, A., Banerjee, S., Finley, A. O., & Gelfand, A. E. (2014), “Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets,” *arXiv preprint* arXiv:1406.7343.

[10] Fan, Y., Li, R. (2012), “Variable selection in linear mixed effects models,” *Annals of statistics*, 40(4), 2043.

[11] Frieze, A., Kannan, R., & Vempala, S. (2004), “Fast Monte-Carlo algorithms for finding low-rank approximations,” *Journal of the ACM*, 51(6), 1025-1041.

[12] Furrer, R., Genton, M. G., & Nychka, D. (2006), “Covariance Tapering for Interpolation of Large Spatial Datasets,” *Journal of Computational and Graphical Statistics*, 15(3) 502-523.

[13] Marshall, A. W., Olkin, I. (1979), “Theory of Majorization and its Applications,” *Academic, New York,* 16, 4-93.

[14] Nychka, D., Ellner, S., Haaland, P. D., & O’Connell, M. A. (1996), “FUNFITS, data analysis and statistical tools for estimating functions,” *Raleigh: North Carolina State University*.

[15] Nychka, D., Wikle, C., & Royle, J. A. (2002), "Multiresolution models for nonstationary spatial covariance functions," Statistical Modelling, 2(4), 315-331.

[16] Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., & Sain, S. (2015), "A multiresolution gaussian process model for the analysis of large spatial datasets," *Journal of Computational and Graphical Statistics*, 24(2), 579-599.

[17] Peng, C., Wu, C. F. J., (2013), "On the choice of nugget in kriging modeling for deterministic computer experiments," *Journal of Computational and Graphical Statistics*, 23(1), 151-168.

[18] Schott, J. R. (2005), "Matrix analysis for statistics."

[19] Simon, B., (1979), "Trace ideals and their applications (Vol. 35)," *Cambridge: Cambridge University Press*.

[20] Stein, M. L. (2013), "Statistical properties of covariance tapers," *Journal of Computational and Graphical Statistics*, 22(4), 866-885.

[21] Stein, M. L. (2014), " Limitations on low rank approximations for covariance matrices of spatial data. Spatial Statistics," *Spatial Statistics*, 8 (2014): 1-19.

[22] Tzeng, S., Huang, H. C. (2015), "Non-stationary Multivariate Spatial Covariance Estimation via Low-Rank Regularization," *Statistical Sinica*, 26, 151-172.

[23] Journ$\acute{e}$e, M., Bach, F., Absil, P. A., & Sepulchre, R. (2010), "Low-rank optimization on the cone of positive semidefinite matrices," *SIAM Journal on Optimization*, 20(5), 2327-2351.

[24] Xu, Y., Yin, W. (2013), "A block coordinate descent method for multi-convex optimization with applications to nonnegative tensor factorization and completion," *SIAM Journal on Imaging Sciences*, 6(3), 1758-1789.

[25] Yuan, M. and Lin, Y. (2006), "Model selection and estimation in regression with grouped variables", *Journal of the Royal Statistical Society. Series B (Methodological)*, 68, 49-67.

[26] Zhao, Y. B. (2012), "An approximation theory of matrix rank minimization and its application to quadratic equations," *Linear Algebra and its Applications*, 437(1), 77-93.