Estimation of covariance matrix via shrinkage Cholesky factor

Naoto Chanohara, Tomoyuki Nakagawa and Hirofumi Wakaki

[*1]Department of Mathematics, Graduate School of Science, Hiroshima University

Abstract

We propose an estimator of the covariance matrix, the L_2 penalizing least squares esitmation of the Cholesky factor. We derive some theoretical properties of the estimator, consistency, asymptotic approximation of the Kullback-Leibler risk and optimal penalizing parameter. It is notable that these properties just hold for normal data, do not require any covariance structure constraint. Some simulation results show that the estimator using the optimal penalizing parameter is better than the others in high-dimensional setting.

keywords: covariance estimation, high dimensional, modified Cholesky decomposition, Ridge regression, asymptotic expansion.

1 Introduction

The purpose of this paper is to propose useful estimator of a covariance matrix. A $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ of a random vector $\boldsymbol{y} = (y_1, \cdots, y_p)'$ plays central role in multivariate analysis, time series analysis, and many applied problems. Some applications (e.g. linear discriminant analysis, cannonical correlation analysis) need to estimate $\boldsymbol{\Sigma}^{-1}$ rather than $\boldsymbol{\Sigma}$ itself. Therefore, positive-definiteness has been an important constraint for covariance matrix estimation. Usual estimator of $\boldsymbol{\Sigma}$ is the sample covariance matrix $\boldsymbol{S} = \sum_{i=1}^{n} \boldsymbol{y}_i' \boldsymbol{y}_i / n$, based on a sample of size n from a normal population with mean zero and covariance $\boldsymbol{\Sigma}$. Although $\boldsymbol{S}$ is unbiased and positive-definite when $n > p$, In high-dimensional data $n < p$, $\boldsymbol{S}$ is not invertible and the bias of the largest eigenvalue will be upward (Jonestone, 2001). For this

reason, many methods are proposed that are available in high-dimensional data (for the details in Pourahmadi, 2013).

Pourahmadhi (1999) employed the modified Cholesky decomposition of $\boldsymbol{\Sigma}$, which not only can guarantee positive-definiteness, but also provides statistical interpretation of the Cholesky factor as certain linear regression coefficients and the residual variances (Section 2). The most natural estimator of the coefficients and the variances are the least squares estimator. This estimator $\hat{\boldsymbol{\Sigma}}_{\mathrm{LSE}}$ is the modified Cholesky decomposition of the sample covariance matrix $\boldsymbol{S}$ under certain conditions. However, It cannot guarantee positive-definiteness in high-dimensional data owing to the singularity of the regression. Hence some regularization is necessary for high-dimensional data, Huang *et al.* (2006), Levina *et al.* (2008) and Chang and Tsay (2010) proposed estimation methods that penalize the log-likelihood for normal data.

The penalized likelihood estimator may be too complicated to study the theoretical properties without some covariance structure constraint (see Bickel and Levina, 2008). Thus, we propose a simple estimator $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda})$, the L_2 penalizing least squares esitmation (PLSE) of the Cholesky factor. We derive some theoretical properties of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda})$ in large sample asymptotic framework, consistency, asymptotic approximation of the Kullback-Leibler risk and optimal penalizing parameter $\boldsymbol{\lambda}_*$ (Section 3). It is notable that these properties just hold for normal data, do not require any covariance structure constraint. Since $\boldsymbol{\lambda}_*$ depends on the true covariance matrix $\boldsymbol{\Sigma}$, we also propose a cross-validation approach $\hat{\alpha}_{\mathrm{CV}*}$ that selects the scalar parameter for $\boldsymbol{\lambda}_*$.

Pourahmadhi (1999) assume mean vector zero for the reason of explaining statistical properties of the modified Cholesky decomposition of covariance matrix. We also study the modified Cholesky decomposition and the performance of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda})$ without the assumption of mean vector in Section 4.

In order to compare the performance of the penalizing parameter selection methods and the other existing methods, we run some Monte-Carlo simulations (Section 5). Although our method $\boldsymbol{\lambda}_*$ is derived in large sample asymptotic framework, some results show that the estimator using $\boldsymbol{\lambda}_*$ is better than the others in high-dimensional setting. Section 6 concludes the paper with discussion.

2 The modified Cholesky decomposition

In this section, we briefly review the modified Cholesky decomposition by Pourahmadi (1999), and provide two estimator of Cholesky factor, the sample

covariance matirix $\boldsymbol{S}$ and the L_2 penalized maximum likelihood estimator proposed by Hunag *et al.* (2006).

2.1 The modified Cholesky decomposition

Let $\boldsymbol{y} = (y_1, \ldots, y_p)'$ be distributed as a certain distribution with mean $\mathbf{0}$ and covariance $\boldsymbol{\Sigma}$ which is assumed to be positive-definite. The standard Cholesky decomposition of $\boldsymbol{\Sigma}$ is $\boldsymbol{CC}'$ where $\boldsymbol{C} = (c_{ij})$ is the lower-trianglar matrix with positive diangonal entries. The modified Cholesky decomposition of $\boldsymbol{\Sigma}$ is

$$\boldsymbol{\Sigma} = \boldsymbol{L}\boldsymbol{D}^2\boldsymbol{L}', \tag{2.1}$$

where $\boldsymbol{D} = \mathrm{diag}(c_{11}, \ldots, c_{pp})$ and $\boldsymbol{L} = \boldsymbol{C}\boldsymbol{D}^{-1}$ is an unit lower-triangular matrix whose diagonal elements are all one. Note that positive-definiteness of $\boldsymbol{\Sigma}$ is necessary and sufficient for uniquely defined $\boldsymbol{L}$ and $\boldsymbol{D}^2$. Pourahmadi (1999) shows that the elements of $\boldsymbol{L}$ and $\boldsymbol{D}$ have statistical interpretations as the following linear regression coefficients $\boldsymbol{\phi}_j$'s and the residual variances σ_j^2's:

$$y_j = \sum_{k=1}^{j-1} \phi_{j,k} y_k + \varepsilon_j, \quad (j = 2, \cdots, p) \tag{2.2}$$

where

$$\begin{aligned} \boldsymbol{\phi}_j &= (\phi_{j,1}, \ldots, \phi_{j,j-1})' \\ &= \underset{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_{j-1})' \in \mathbb{R}^{j-1}}{\mathrm{argmin}} E\left[\left(y_j - \sum_{i=1}^{j-1} y_i \beta_i\right)^2\right]. \end{aligned}$$

Let $\boldsymbol{T} = (t_{ij})$ be a unit lower-trianglar matrix with $t_{ij} = -\phi_{ij}$ for $i < j$, $\varepsilon_1 = y_1$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_p)'$, then (2.2) becomes

$$\boldsymbol{T}\boldsymbol{y} = \boldsymbol{\epsilon}. \tag{2.3}$$

Since ϵ_j's are uncorrelated, taking covariance of both side of (2.3) gives the modified Cholesky decomposition (2.1),

$$\boldsymbol{\Sigma} = \boldsymbol{T}^{-1}\boldsymbol{D}^2\left(\boldsymbol{T}'\right)^{-1}, \tag{2.4}$$

where $\boldsymbol{D}^2 = \mathrm{diag}(\mathrm{Var}(\epsilon_1), \ldots, \mathrm{Var}(\epsilon_p)) = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. Therefore the modified Cholesky decomposition can transform the problem of the covariance matrix estimation into that of the linear regression coefficients and the residual variances estimation. Note that the regression coefficients $\boldsymbol{\phi}_j$'s are unconstrained, however, the residual variances σ_j^2's must be positive. We call $\boldsymbol{\phi}_j$'s and σ_j^2's the Cholesky factors for short.

2.2 Some estimators using the modified Cholesky decomposition

Suppose that we observe a random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)' = (y_{ij})$ where $\boldsymbol{y}_i$'s are mutually independently and identically distributed as $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$. The most natural estimator of $\boldsymbol{\phi}_j$'s and σ_j^2's are the least squares estimator

$$\begin{aligned}\hat{\boldsymbol{\phi}}_{j,\mathrm{LSE}} &= \left(\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\right)^{-1}\boldsymbol{Y}_{j-1}'\boldsymbol{y}_{(j)},\\ \hat{\sigma}^2_{j,\mathrm{LSE}} &= \frac{1}{n}(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\hat{\boldsymbol{\phi}}_{j,\mathrm{LSE}})'(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\hat{\boldsymbol{\phi}}_{j,\mathrm{LSE}}),\end{aligned} \tag{2.5}$$

for $j = 2, \ldots, p$ where $\boldsymbol{y}_{(k)} = (y_{1k}, \ldots, y_{nk})'$, $\boldsymbol{Y}_k = (\boldsymbol{y}_{(1)}, \ldots, \boldsymbol{y}_{(k)})$, and $\hat{\sigma}^2_{1,\mathrm{LSE}} = \boldsymbol{y}_{(1)}'\boldsymbol{y}_{(1)}/n$. This estimator is derived from minimizing the squared error

$$(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\boldsymbol{\beta})'(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\boldsymbol{\beta}), (j = 2, \cdots, p)$$

that is, the same problem of maximizing the log-likelihood function $l(\boldsymbol{\Sigma}; \boldsymbol{Y})$ under the assumption of normality. The log-likelihood function is

$$\begin{aligned}-2l(\boldsymbol{\Sigma}; \boldsymbol{Y}) &= \sum_{i=1}^{n} \log|\boldsymbol{D}^2| + \boldsymbol{y}_i'\boldsymbol{T}'\boldsymbol{D}^{-2}\boldsymbol{T}\boldsymbol{y}_i\\ &= n\sum_{j=1}^{p}\log\sigma_j^2 + \frac{1}{\sigma_1^2}\boldsymbol{y}_{(1)}'\boldsymbol{y}_{(1)}\\ &\quad + \sum_{j=2}^{p}\frac{1}{\sigma_1^2}\left(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\boldsymbol{\phi}_j\right)'\left(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\boldsymbol{\phi}_j\right),\end{aligned} \tag{2.6}$$

ignoring constant. Therefore, the estimator $\hat{\boldsymbol{\Sigma}}_{\mathrm{LSE}}$ that replaces $\boldsymbol{\phi}_j$'s and σ_j^2's with $\hat{\boldsymbol{\phi}}_{j,\mathrm{LSE}}$'s and $\hat{\sigma}^2_{j,\mathrm{LSE}}$'s in (2.4), is the sample covariance matrix $\boldsymbol{S} = \boldsymbol{Y}'\boldsymbol{Y}/n$. If $n < p$ (for $j \geq n+1$), $\hat{\boldsymbol{\Sigma}}_{\mathrm{LSE}}$ cannot guarantee positive-definiteness because $\hat{\boldsymbol{\phi}}_{j,\mathrm{LSE}}$ is not unique and $\hat{\sigma}^2_{j,\mathrm{LSE}}$ can become zero, hence some regularization is necessary for $\hat{\boldsymbol{\phi}}_j$ to be unique.

Huang et al. (2006), Levina et al. (2008), and Chang and Tsay (2010) proposed a regularzation method based on penalizing $l(\boldsymbol{\Sigma}; \boldsymbol{Y})$, impose a penalty on the Cholesky factor $\boldsymbol{\phi}_j's$. Huang et al. (2006) proposed adding L_q penalty to (2.6)

$$-2l(\boldsymbol{\Sigma}; \boldsymbol{Y}) + \delta\sum_{j=2}^{p}|\boldsymbol{\phi}_j|^q, \tag{2.7}$$

where $q > 0$ and $\delta \geq 0$. The penalized likelihood estimator minimising (2.7) for $q = 2$ can be written as

$$\begin{aligned}\hat{\boldsymbol{\phi}}_{j,\text{Huang}} &= \left(\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1} + \delta\hat{\sigma}^2_{j,\text{Huang}}\boldsymbol{I}_{j-1}\right)^{-1}\boldsymbol{Y}_{j-1}'\boldsymbol{y}_{(j)} \\ \hat{\sigma}^2_{j,\text{Huang}} &= \frac{1}{n}(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\hat{\boldsymbol{\phi}}_{j,\text{Huang}})'(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\hat{\boldsymbol{\phi}}_{j,\text{Huang}}),\end{aligned} \tag{2.8}$$

for $j = 2, \ldots, p$ where $\boldsymbol{I}_j$ is the $j \times j$ identity matrix and $\hat{\sigma}^2_{1,\text{Huang}} = \boldsymbol{y}_{(1)}'\boldsymbol{y}_{(1)}/n$. The estimator of $\boldsymbol{\Sigma}$ using $\hat{\boldsymbol{\phi}}_{j,\text{Huang}}$'s and $\hat{\sigma}^2_{j,\text{Huang}}$'s with $\delta > 0$ is positive-definite. However, iterative algorithm is necessary for the estimation.

3 The L_2 penalized least squares estimator

In this section, we propose $\hat{\boldsymbol{\Sigma}}_{\text{PLSE}}(\boldsymbol{\lambda})$: *the L_2 penalized least squares estimator* of $\boldsymbol{\Sigma}$, and study some asymptotic properties. Moreover, we clarify the optimal penalizing parameter $\boldsymbol{\lambda}_*$ with respect to a certain risk, and suggest how to estimate the parameter.

The penalized likelihood estimator may be too complicated to study theoretical properties. Therefore, we propose a simple estimator $\hat{\boldsymbol{\Sigma}}_{\text{PLSE}}(\boldsymbol{\lambda})$ that the estimator of $\boldsymbol{\phi}_j$'s and σ_j^2's has the following closed-form,

$$\begin{aligned}\hat{\boldsymbol{\phi}}_{j,\text{PLSE}} &= \left(\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1} + \lambda_j\boldsymbol{I}_{j-1}\right)^{-1}\boldsymbol{Y}_{j-1}'\boldsymbol{y}_{(j)}, \\ \hat{\sigma}^2_{j,\text{PLSE}} &= \frac{1}{n}(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\hat{\boldsymbol{\phi}}_{j,\text{PLSE}})'(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\hat{\boldsymbol{\phi}}_{j,\text{PLSE}}),\end{aligned} \tag{3.1}$$

for $j = 2, \ldots, p$ where $\hat{\sigma}^2_{1,\text{PLSE}} = \boldsymbol{y}_{(1)}'\boldsymbol{y}_{(1)}/n$, $\boldsymbol{\lambda} = (\lambda_2, \cdots, \lambda_p)'$, $\lambda_j \geq 0$. This estimator is derived from *Ridge regression* (Hoerl and Kennard, 1970), minimizing the L_2 penalized squares error $(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\boldsymbol{\beta})'(\boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\boldsymbol{\beta}) + \lambda_j\|\boldsymbol{\beta}\|_2^2$ for $j = 2, \cdots, p$. (2.8) can be considered as a special case of (3.1) in the sense that $\lambda_j = \delta\hat{\sigma}^2_{j,\text{Huang}}$. If $\delta = 0$ and $\boldsymbol{\lambda} = \boldsymbol{0}$, then the estimator (2.8) and (3.1) become the sample covariance matrix (2.5). If λ_j's are fixed constants, (3.1) does not need any iterative algorithm.

We also propose the following estimator by selecting the penalizing pa-

rameter:

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*}) := \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}\left(\boldsymbol{\lambda}_*\left(\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\hat{\alpha}_{CV*}\mathbf{1}_{p-1})\right)\right),$$

$$\boldsymbol{\lambda}_*(\boldsymbol{\Sigma}) := (\lambda_{2*}, \cdots, \lambda_{p*})', \ \lambda_{j*} = \mathrm{tr}\boldsymbol{\Sigma}_{j-1}^{-1}\frac{\sigma_j^2}{\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j},$$

$$\hat{\alpha}_{CV*} = \underset{\alpha\geq 0}{\mathrm{argmin}}\frac{1}{K}\sum_{\nu=1}^{K}\left(n_\nu \log\left|\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \alpha)_{-\nu}\right| + \sum_{i\in I_\nu}\boldsymbol{y}_i'\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}^{-1}(\boldsymbol{\lambda}_* : \alpha)_{-\nu}\boldsymbol{y}_i\right).$$

The detail and the derivation are given in the following subsection.

3.1 Asymptotic properties

$\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda})$ is derived by *Ridge regression*, shrink the estimators of regression coefficients $\boldsymbol{\phi}_j$'s. Shrinkage the elements of the sample covariance is studied by Stein (1975), Ledoit and Wolf (2003) etc. By contrast, Few studies have focused on the shrinkage estimation of the Cholesky factor for covariance estimation. Therefore, we study some theoretical properties of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda})$, consistency and asymptotic approximation of the Kullback-Leibler risk.

In this subsection, we assume that $\{\boldsymbol{y_i}\}_{i=1}^n$ are independently and identically distributed as $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ (Normal distribution), p is fixed and $n \to \infty$ (Large sample asymptotic framework), and $\lambda_j = O(1)$ for all $j = 2, \cdots, p$ (Constant order).

First, we represent the estimator (3.1) using the true residual ε_j's in (2.2). Let $\boldsymbol{\varepsilon}_{(j)} = \boldsymbol{y}_{(j)} - \boldsymbol{Y}_{j-1}\boldsymbol{\phi}_j$ for $j = 2, \ldots, p$, then

$$\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} = \left(\frac{1}{n}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1} + \frac{1}{n}\lambda_j\boldsymbol{I}_{j-1}\right)^{-1}\left(\frac{1}{n}\boldsymbol{Y}_{j-1}'\boldsymbol{\varepsilon}_{(j)} + \frac{1}{n}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\phi}_j\right), \quad (3.2)$$

$$\hat{\sigma}_{j,\mathrm{PLSE}}^2 = \frac{1}{n}\boldsymbol{\varepsilon}_{(j)}'\boldsymbol{\varepsilon}_{(j)} + \frac{2}{n}\left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}\right)'\boldsymbol{Y}_{j-1}'\boldsymbol{\varepsilon}_{(j)} + \left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}\right)'\frac{1}{n}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}\right)'. \quad (3.3)$$

Here, the distribution of the random vector or matrix, $\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}$, $\boldsymbol{Y}_{j-1}'\boldsymbol{\varepsilon}_{(j)}$, and $\boldsymbol{\varepsilon}_{(j)}'\boldsymbol{\varepsilon}_{(j)}$ can be expressed as the following lemma.

Lemma 3.1. *It holds for* $j = 2, \ldots, p$ *that*

$$n^{-1}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2} = \boldsymbol{I}_{j-1} + \boldsymbol{Z}n^{-1/2},$$

$$n^{-1/2}\sigma_j^{-1}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{\varepsilon}_{(j)} = \left\{\boldsymbol{I}_{j-1} + \frac{1}{2}\boldsymbol{Z}n^{-1/2} - \frac{1}{8}\boldsymbol{Z}^2n^{-1} + \frac{1}{16}\boldsymbol{Z}^3n^{-3/2}\right\}\boldsymbol{V} + O_p(n^{-5/2}),$$

$$n^{-1}\sigma_j^{-2}\boldsymbol{\varepsilon}_{(j)}'\boldsymbol{\varepsilon}_{(j)} = 1 + \sqrt{2}Xn^{-1/2} + (\boldsymbol{V}'\boldsymbol{V} - j + 1)n^{-1} - \frac{1}{\sqrt{2}}(j-1)Xn^{-3/2} + O_p(n^{-5/2})$$

where $\boldsymbol{\Sigma}_j$ *is the submatrix of* $\boldsymbol{\Sigma} = (\sigma_{ij})$*: the entries* $(\boldsymbol{\Sigma}_j)_{kl} = \sigma_{kl}$ *for* $1 \le k, l \le j$*,*

$$\boldsymbol{Z} := \sqrt{n}\left(\frac{1}{n}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2} - \boldsymbol{I}_{j-1}\right) \xrightarrow{d} N_{j-1\times j-1}(\boldsymbol{O}_{j-1\times j-1}, \boldsymbol{\Omega}),$$

$$\boldsymbol{V} := \sigma_j^{-1}\left(\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2}\right)^{-1/2}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{\varepsilon}_{(j)} \sim N_{j-1}(\boldsymbol{0}, \boldsymbol{I}_{j-1})$$

$$X := \sqrt{\frac{n-j+1}{2}}\left(\frac{1}{n-j+1}\boldsymbol{U}'\boldsymbol{U} - 1\right) \xrightarrow{d} N(0,1),$$

and $\boldsymbol{U} \sim N_{n-j+1}(\boldsymbol{0}, \boldsymbol{I}_{n-j+1})$*,* $\boldsymbol{\Omega}$ *:* $\mathrm{Cov}(Z_{ij}, Z_{kl}) = \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}$*, and* $\boldsymbol{Z}$*,* $\boldsymbol{V}$*,* $X(or\boldsymbol{U})$ *are mutually independent.*

The proof of the lemma is given in Appendix. Note that $\boldsymbol{\Sigma}_j^{-1}$ is not any submatrix of $\boldsymbol{\Sigma}^{-1}$, but the inverse matrix of $\boldsymbol{\Sigma}_j$.

Applying Lemma 3.1 to (3.2) and (3.3), we can derive a perturbation expansion of $\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}$'s and $\hat{\sigma}^2_{j,\mathrm{PLSE}}$'s.

Lemma 3.2.

$$\begin{aligned}\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} = \boldsymbol{\Sigma}_{j-1}^{-1/2}\Bigg\{&\boldsymbol{\Sigma}_{j-1}^{1/2}\boldsymbol{\phi}_j + \sigma_j\boldsymbol{V}n^{-1/2} + \left(-\frac{1}{2}\sigma_j\boldsymbol{Z}\boldsymbol{V} - \lambda_j\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j\right)n^{-1}\\ &+\left(\frac{3}{8}\sigma_j\boldsymbol{Z}^2\boldsymbol{V} - \lambda_j\sigma_j\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{V} + \lambda_j\boldsymbol{Z}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j\right)n^{-3/2}\\ &+\Bigg(-\frac{5}{16}\sigma_j\boldsymbol{Z}^3\boldsymbol{V} + \frac{1}{2}\lambda_j\sigma_j\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{Z}\boldsymbol{V} + \lambda_j\sigma_j\boldsymbol{Z}\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{V}\\ &\quad+\lambda_j^2\boldsymbol{\Sigma}_{j-1}^{-3/2}\boldsymbol{\phi}_j - \lambda_j\boldsymbol{Z}^2\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j\Bigg)n^{-2}\Bigg\} + O_p(n^{-5/2}),\end{aligned}$$

$$\begin{aligned}\hat{\sigma}^2_{j\ PLSE} = \sigma_j^2\Bigg(&1 + \sqrt{2}Xn^{-1/2} - (j-1)n^{-1}\\ &-\frac{1}{\sqrt{2}}(j-1)Xn^{-3/2} + \frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_jn^{-2}\Bigg) + O_p(n^{-5/2}),\end{aligned}$$

From Lemma 3.2 and the consistency of $\hat{\sigma}^2_{1,\mathrm{PLSE}} = \boldsymbol{y}_{(1)}'\boldsymbol{y}_{(1)}/n$, the consistency of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}$ holds:

Theorem 3.1.

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}} \xrightarrow{p} \boldsymbol{\Sigma}.$$

Note that this theorem holds for arbitrary positive value $\boldsymbol{\lambda}$ and covariance structure.

In order to estimate the optimal penalizing parameter and theoretical comparison with the sample covariance matrix, we approximate the Kullback-Leibler risk with using some recurrence formula. Let $\hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}) = \hat{\boldsymbol{\Sigma}}$ be an estimator of $\boldsymbol{\Sigma}$ using a sample $\boldsymbol{Y}$, then the Kullback-Leibler loss is defined as

$$KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) = \mathrm{tr}\left(\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\right) - \log\left|\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\right| - p. \tag{3.4}$$

The loss is used as the measure for an estimator of inverse covariance matrix. We say an estimator is considered better than another estimator, if the risk of the estimator

$$E_{\boldsymbol{Y}}\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}))\right],$$

is smaller than that of another one. For a fixed p, the Kullback-Leibler loss can be obtained recursively as in the following lemma.

Lemma 3.3. *For $j = 2, \ldots, p$, it holds that*

$$\begin{aligned} KL(\boldsymbol{\Sigma}_j, \hat{\boldsymbol{\Sigma}}_j) = {} & KL(\boldsymbol{\Sigma}_{j-1}, \hat{\boldsymbol{\Sigma}}_{j-1}) \\ & + \frac{1}{\hat{\sigma}_j^2}\left(\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j\right)' \boldsymbol{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j\right) + \frac{\sigma_j^2}{\hat{\sigma}_j^2} - \log\frac{\sigma_j^2}{\hat{\sigma}_j^2} - 1, \end{aligned}$$

where $\hat{\boldsymbol{\Sigma}}_j$ is the $j \times j$ submatrix of $\hat{\boldsymbol{\Sigma}}$.

The proof is given in Appendix. Note that the lemma do not need normal and large sample assumption. The following Lemma gives an approximation of the expected value using $\hat{\boldsymbol{\Sigma}}_{PLSE}(\boldsymbol{\lambda})$.

Lemma 3.4. *It holds for $j = 2, \ldots, p$ that*

$$\begin{aligned} & E\left[\frac{1}{\hat{\sigma}_{j,\mathrm{PLSE}}^2}\left(\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} - \boldsymbol{\phi}_j\right)' \boldsymbol{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} - \boldsymbol{\phi}_j\right) + \frac{\sigma_j^2}{\hat{\sigma}_{j,\mathrm{PLSE}}^2} - \log\frac{\sigma_j^2}{\hat{\sigma}_{j,\mathrm{PLSE}}^2} - 1\right] \\ & = jn^{-1} + \left\{\frac{5}{2}j^2 + j + \frac{1}{6} + \frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j - 2\mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)\lambda_{j-1}\right\} n^{-2} + O(n^{-5/2}) \end{aligned}$$

The proof is given in Appendix. Accordingly, we obtain the approximation of the Kullback-Leibler risk.

Theorem 3.2.

$$E\left[KL(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}))\right] = \frac{p}{2}(p+1)n^{-1} + \left\{\frac{p}{12}\left(10p^2+21p+13\right)\right. \tag{3.5}$$
$$\left. + \sum_{j=2}^{p} \frac{\boldsymbol{\phi}_j' \mathbf{\Sigma}_{j-1}^{-1} \boldsymbol{\phi}_j}{\sigma_j^2}\lambda_j^2 - 2\mathrm{tr}\mathbf{\Sigma}_{j-1}^{-1}\lambda_j\right\} n^{-2} + O(n^{-5/2}).$$

Proof. For simplicity, we write $\hat{\mathbf{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}) = \hat{\mathbf{\Sigma}}$, $\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} = \hat{\boldsymbol{\phi}}_j$ and $\hat{\sigma}^2_{j,\mathrm{PLSE}} = \hat{\sigma}^2_j$. From Lemma 3.3 and Lemma 3.4, we evaluate the first term of the following equation,

$$E\left[KL(\mathbf{\Sigma}, \hat{\mathbf{\Sigma}})\right] = E\left[KL(\Sigma_1, \hat{\Sigma}_1)\right]$$
$$+ \sum_{j=2}^{p} E\left[\frac{1}{\hat{\sigma}_j^2}\left(\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j\right)' \mathbf{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_j - \boldsymbol{\phi}_j\right) + \frac{\sigma_j^2}{\hat{\sigma}_j^2} - \log\frac{\sigma_j^2}{\hat{\sigma}_j^2} - 1\right].$$

Since $\hat{\Sigma}_1 = \hat{\sigma}_1^2 = \boldsymbol{y}_{(1)}'\boldsymbol{y}_{(1)}/n$, $\Sigma_1 = \sigma_1^2$, and $n\hat{\sigma}_1^2/\sigma_1^2 \sim \chi_n^2$,

$$\begin{aligned} E\left[KL(\Sigma_1, \hat{\Sigma}_1)\right] &= E\left[\frac{\sigma_1^2}{\hat{\sigma}_1^2} - \log\frac{\sigma_1^2}{\hat{\sigma}_1^2} - 1\right] \\ &= \frac{n}{n-2} - \log n + \log 2 + \psi\left(\frac{n}{2}\right) - 1 \\ &= n^{-1} + \frac{11}{3}n^{-2} + O\left(n^{-5/2}\right), \end{aligned}$$

where ψ is the digamma function. Here, we used the asymptotic expansion, $\psi(x) = \log x - 1/(2x) - \sum_{k=1}^{\infty} B_{2k}(2kx^{2k})^{-1}$ where B_{2k} is the k-th Bernoulli number. □

From Theorem 2, we can compare $\hat{\mathbf{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda})$ with the sample covariance matrix $\hat{\mathbf{\Sigma}}_{\mathrm{PLSE}}(\mathbf{0})$ for a fixed $\boldsymbol{\lambda}$. Furthermore, (4.4) can be considered a quadratic function of $\boldsymbol{\lambda}$ ignoring the higher order term in the expansion. These discussion is in Section 3.3 due to the constant constraint of penalizing parameters.

3.2 How to select the penalizing parameter

The performance of penalized methods depends on the choice of the penalizing parameter. Huang et al. (2006) used two cross-validation methods,

K-fold cross-validation and generalized cross-varidation. The penalizing parameter of the proposed method in (2.8) is a scalar δ, whereas that of our method is $p-1$ dimensional vector $\boldsymbol{\lambda} = (\lambda_2, \cdots, \lambda_p)'$. The optimization with respect to $p-1$ variables in the cross-validation criterion should be avoided for computational cost. Therefore, we give the optimal penalizing parameter $\boldsymbol{\lambda}_*$ under the Kullback-Leibler risk, and theoretical comparison $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_*)$ with the sample covariance matrix $\boldsymbol{S}$. Since $\boldsymbol{\lambda}_*$ depends on the true covariance matrix $\boldsymbol{\Sigma}$, we also propose a cross-validation approach $\hat{\alpha}_{CV*}$ that selects a scalar parameter for $\boldsymbol{\lambda}_*$.

3.2.1 Optimal penalizing parameter under the Kullback-Leibler risk

From theorem 2, the approximated risk can be considered as a quadratic function of $\boldsymbol{\lambda}$ ignoring the higher order term in the expansion. Thus, the optimal penalizing parameter and the risk can be approximated by minimizing the quadratic function.

Theorem 3.3. *Suppose that* $(\sigma_{1j}, \ldots, \sigma_{j-1,j})' \neq \mathbf{0}$ *for all* $j = 2, \ldots, p$. *If we use the penalizing parameter* $\boldsymbol{\lambda}_* = (\lambda_{2*}, \ldots, \lambda_{p*})'$, *where*

$$\lambda_{j*} = \mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right) \frac{\sigma_j^2}{\boldsymbol{\phi}_j' \boldsymbol{\Sigma}_{j-1}^{-1} \boldsymbol{\phi}_j}, \quad (j = 2, \ldots, p),$$

then the risk can be written as

$$E\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{PLSE}(\boldsymbol{\lambda}_*))\right] =$$
$$\frac{p}{2}(p+1)n^{-1} + \left\{\frac{p}{12}\left(10p^2 + 21p + 13\right) - \sum_{j=2}^{p} \lambda_{j*} \mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)\right\} n^{-2} + O(n^{-5/2}).$$

For all $j = 2, \ldots, p$, λ_{j*} must be a positive value because $\boldsymbol{\Sigma}_j$ is positive-definite. The risk of the sample covariance matrix is also approximated by $\boldsymbol{\lambda} = \mathbf{0}$ in (4.4), that is the first and second term of (4.4). Therefore, the approximated risk of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_*)$ is smaller than that of the sample covariance matrix by $\sum_{j=2}^{p} \lambda_{j*} \mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)$.

Note that $\boldsymbol{\lambda}_*$ is expected to be useful without the covariance structure assumption, $(\sigma_{1j}, \ldots, \sigma_{j-1,j})' \neq \mathbf{0}$ for all $j = 2, \ldots, p$. If $(\sigma_{1j}, \ldots, \sigma_{j-1,j}) = \mathbf{0}$ (e.g. $\boldsymbol{\Sigma}$ is a diagonal matrix), then the parameter constant assumption $\boldsymbol{\lambda} = O(1)$ is not satisfied because $\boldsymbol{\phi}_j = \boldsymbol{\Sigma}_{j-1}^{-1}(\sigma_{1j}, \ldots, \sigma_{j-1,j})' = \mathbf{0} \Rightarrow \lambda_{j*} = \infty$. For this reason, the approximation of the risk may be poor in these covariance structure. However, if $\lambda_j = \infty$, the estimator takes the right value $\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} = \mathbf{0}$. To see the performance in such cases, we run the simulations in Section 4.

3.2.2 The cross-validated estimator

Since the optimal penalizing parameter $\boldsymbol{\lambda}_*$ includes the true covariance matrix $\boldsymbol{\Sigma}$, we propose a cross-validation method that replaces it with $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\hat{\alpha}_{CV*}\mathbf{1}_{p-1})$.

We define $\boldsymbol{\lambda}_*$ as a function of $\boldsymbol{\Sigma}$, $\boldsymbol{\lambda}_*(\boldsymbol{\Sigma}) = (\lambda_{2*}(\boldsymbol{\Sigma}), \ldots, \lambda_{p*}(\boldsymbol{\Sigma}))'$,

$$\lambda_{j*}(\boldsymbol{\Sigma}) = \frac{\sigma_j^2 \mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)}{\boldsymbol{\phi}_j' \boldsymbol{\Sigma}_{j-1}^{-1} \boldsymbol{\phi}_j}, \quad (j = 2 \ldots, p).$$

We consider to replace $\boldsymbol{\Sigma}$ in $\boldsymbol{\lambda}_*(\boldsymbol{\Sigma})$ with $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\alpha\mathbf{1}_{p-1})$, where $\alpha \geq 0$ is a scalar parameter, and $\mathbf{1}_{p-1} = (1, \cdots, 1)'$. Therefore, for a fixed α, the estimator of $\boldsymbol{\Sigma}$ can be expressed as

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \alpha) := \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_*(\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\alpha\mathbf{1}_{p-1}))).$$

Hence we just consider how to select a scalar $\alpha \geq 0$. According to K-fold cross-validation in Huang et al. (2006), we can select the parameter $\hat{\alpha}_{CV}$ that minimize the criterion

$$\hat{\alpha}_{CV} = \underset{\alpha \geq 0}{\mathrm{argmin}} \frac{1}{K} \sum_{\nu=1}^{K} \left(n_\nu \log \left| \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\alpha\mathbf{1}_{p-1})_{-\nu} \right| + \sum_{i \in I_\nu} \boldsymbol{y}_i' \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}^{-1}(\alpha\mathbf{1}_{p-1})_{-\nu} \boldsymbol{y}_i \right), \tag{3.6}$$

where K is the number of folding the sample, D is the full sample, D_ν is the ν-th test sample, $D - D_\nu$ is the training sample, I_ν is the index set of D_ν, n_ν is the size of I_ν, and $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\alpha\mathbf{1}_{p-1})_{-\nu}$ is the estimator using $D - D_\nu$ for $\boldsymbol{\lambda} = \alpha\mathbf{1}_{p-1}$. The function of α in (3.6) can be considered as an estimator of the log-lokelihood for normally distributed sample (see Huang *et al.*, 2006, Levina *et al.*, 2008). However, we can also consider it as a criterion of the Kullback-Leibler risk

$$\begin{aligned} KL(\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}) &= \mathrm{tr}\left(\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\right) - \log\left|\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\right| - p \\ &\propto \mathrm{tr}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1} + \log\left|\hat{\boldsymbol{\Sigma}}\right| \approx \mathrm{tr}\frac{1}{n_\nu}\sum_{i \in I_\nu} \boldsymbol{y}_i \boldsymbol{y}_i' \hat{\boldsymbol{\Sigma}}_{-\nu}^{-1} + \log\left|\hat{\boldsymbol{\Sigma}}_{-\nu}^{-1}\right|. \end{aligned} \tag{3.7}$$

Although the estimation target of $\hat{\alpha}_{CV}$ can be considered as

$$\alpha_{CV} = \mathrm{argmin}\, E\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\alpha\mathbf{1}_{p-1}))\right],$$

in generally

$$E\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \alpha_{CV}))\right] \neq \min E\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \alpha))\right]. \tag{3.8}$$

Therefore, the estimation target that we actually want is the right side in (3.8),

$$\alpha_{CV*} = \underset{\alpha \geq 0}{\operatorname{argmin}} E\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \alpha))\right]. \tag{3.9}$$

Hence we propose the cross-validation method to choose the scalar $\alpha = \hat{\alpha}_{CV*}$ that minimize the Kullback-Leibler criterion with $\boldsymbol{\lambda}_*(\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\alpha \mathbf{1}_{p-1}))$

$$\hat{\alpha}_{CV*} = \underset{\alpha \geq 0}{\operatorname{argmin}} \frac{1}{K} \sum_{\nu=1}^{K} \left(n_\nu \log \left| \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \alpha)_{-\nu} \right| + \sum_{i \in I_\nu} \boldsymbol{y}_i' \hat{\boldsymbol{\Sigma}}^{-1}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \alpha)_{-\nu} \boldsymbol{y}_i \right).$$

Note that this cross-validation method is not an optimization of $p-1$ variables function. The computational cost of this cross-validation method is expected as same as or a little more expensive than the scalar cross-validation (3.6).

4 The estimator when non-zero mean

In this section, we study the modified Cholesky decomposition of $\boldsymbol{\Sigma}$ and the estimator $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}$ without the assumption of mean vector $\boldsymbol{\mu} = \mathbf{0}$. The differences between the result under the assumption $\boldsymbol{\mu} = \mathbf{0}$ (Section 3) and that of $\boldsymbol{\mu} \neq \mathbf{0}$ are centering the data before the estimation, and the order (degree of freedom) n and $N = n - 1$.

Let $\boldsymbol{y} = (y_1, \ldots, y_p)'$ be distributed as a certain distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)'$ and covariance $\boldsymbol{\Sigma} = (\sigma_{ij})$. Then, the model corresponding in (2.2) can be expressed as

$$y_j = \alpha_j + \sum_{k=1}^{j-1} \phi_{kj} y_k + \varepsilon_j, \quad (j = 2, \ldots, p), \tag{4.1}$$

where

$$\begin{aligned} \begin{pmatrix} \alpha_j \\ \boldsymbol{\phi}_j \end{pmatrix} &= \underset{\boldsymbol{\beta}=(\beta_0, \beta_1, \ldots, \beta_{j-1})' \in R^j}{\operatorname{argmin}} E\left[(y_j - \beta_0 - \sum_{i=1}^{j-1} y_i \beta_i)^2 \right] \\ &= \begin{pmatrix} \mu_j - \boldsymbol{\mu}_{j-1}' \boldsymbol{\phi}_j \\ \boldsymbol{\Sigma}_{j-1}^{-1} (\sigma_{1j}, \cdots, \sigma_{j-1,j})' \end{pmatrix}. \end{aligned}$$

Therefore, (4.1) can be written as

$$(y_j - \mu_j) = \sum_{k=1}^{j-1} \phi_{kj}(y_k - \mu_k) + \epsilon_j, \quad (j = 2, \ldots, p). \tag{4.2}$$

Thus, we can consider the model (4.2) is the usually model (2.2) because $(y_1 - \mu_1, \ldots, y_p, -\mu_p)'$ is mean vector zero.

4.1 The asymptotic properties when non-zero mean

Suppose that we observe a random sample $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)' = (y_{ij})$ where $\boldsymbol{y}_i$'s are mutually independently and identically distributed as $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

From (4.1), the model of n sample can be written for $j = 2, \ldots, p$,

$$\tilde{\boldsymbol{y}}_{(j)} = \tilde{\boldsymbol{Y}}_{j-1}\boldsymbol{\phi}_j + \tilde{\boldsymbol{\varepsilon}}_{(j)}$$

where $\tilde{\boldsymbol{y}}_{(j)} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{y}_{(j)}$, $\tilde{\boldsymbol{Y}}_{j-1} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{Y}_{j-1}$, $\tilde{\boldsymbol{\varepsilon}}_{(j)} = (\boldsymbol{I}_n - \boldsymbol{J}_n)\boldsymbol{\varepsilon}_{(j)}$, $\boldsymbol{J}_n = n^{-1}\mathbf{1}_n\mathbf{1}_n'$, $\mathbf{1}_n$ is n dimensional vector that all elements are one. $(\boldsymbol{I}_n - \boldsymbol{J}_n)$ is the column centered operator matrix. Here, the estimator can be derived that minimize $\tilde{\boldsymbol{\varepsilon}}_{(j)}'\tilde{\boldsymbol{\varepsilon}}_{(j)} + \lambda_j\|\boldsymbol{\phi}_j\|_2^2$,

$$\begin{aligned}
\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} &= \left(\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{Y}}_{j-1} + \lambda_j I_{j-1}\right)^{-1}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{y}}_{(j)} \\
\hat{\sigma}^2_{j,\mathrm{PLSE}} &= \frac{1}{n-1}(\tilde{\boldsymbol{y}}_{(j)} - \tilde{\boldsymbol{Y}}_{j-1}\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}})'(\tilde{\boldsymbol{y}}_{(j)} - \tilde{\boldsymbol{Y}}_{j-1}\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}),
\end{aligned} \tag{4.3}$$

for $j = 2, \ldots, p$ where $\hat{\sigma}^2_{1,\mathrm{PLSE}} = \tilde{\boldsymbol{y}}_{(1)}'\tilde{\boldsymbol{y}}_{(1)}/(n-1)$. Therefore, one can use the estimator $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda})$ as in the case of $\boldsymbol{\mu} = \mathbf{0}$ (3.1) after columm centering the data. Note that (4.3) can also be derived as $\boldsymbol{\varepsilon}_{(j)}'\boldsymbol{\varepsilon}_{(j)} + \lambda_j\|\boldsymbol{\phi}_j\|_2^2$ minimization problem, where

$$\begin{aligned}
\boldsymbol{y}_{(j)} &= \mathbf{1}_n\alpha_j + \boldsymbol{Y}_{j-1}\boldsymbol{\phi}_j + \boldsymbol{\varepsilon}_{(j)} \\
&= \boldsymbol{W}_j\boldsymbol{\psi}_j + \boldsymbol{\varepsilon}_{(j)},
\end{aligned}$$

where $\boldsymbol{W}_j = (\mathbf{1}_n, \boldsymbol{Y}_{j-1})$, $\boldsymbol{\psi}_j = (\alpha_j, \boldsymbol{\phi}_j')'$.

$$\begin{aligned}
\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} &= \left(\frac{1}{N}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{Y}}_{j-1} + \frac{1}{N}\lambda_j\boldsymbol{I}_{j-1}\right)^{-1}\left(\frac{1}{N}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{\varepsilon}}_{(j)} + \frac{1}{N}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{Y}}_{j-1}\boldsymbol{\phi}_j\right), \\
\hat{\sigma}^2_{j,\mathrm{PLSE}} &= \frac{1}{N}\tilde{\boldsymbol{\epsilon}}_{(j)}'\tilde{\boldsymbol{\varepsilon}}_{(j)} + \frac{2}{N}\left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}\right)'\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{\varepsilon}}_{(j)} \\
&\quad + \left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}\right)'\frac{1}{N}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{Y}}_{j-1}\left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}}\right),
\end{aligned}$$

where $n = N - 1$.

$$\boldsymbol{I}_n - \boldsymbol{J}_n = (\boldsymbol{Q}_1\ \boldsymbol{Q}_2) \begin{pmatrix} \boldsymbol{I}_N & \boldsymbol{0} \\ \boldsymbol{0}' & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{Q}_1' \\ \boldsymbol{Q}_2' \end{pmatrix} = \boldsymbol{Q}_1\boldsymbol{Q}_1',$$

We understand that $\boldsymbol{Q}_1'\boldsymbol{Y}_{j-1} \sim N_{N\times j-1}(\boldsymbol{0}, \boldsymbol{\Sigma}_{j-1} \otimes I_N)$, $\boldsymbol{Q}_1'\boldsymbol{\epsilon}_{(j)} \sim N_N(\boldsymbol{0}, \sigma_j^2 I_N)$

By proof like Lemma 3.1, we obtain the following lemma.

Lemma 4.1. *It holds for $j = 2, \ldots, p$ that*

$$\frac{1}{N}\boldsymbol{\Sigma}_{j-1}^{-1/2}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{Y}}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2} = \boldsymbol{I}_{j-1} + \boldsymbol{Z}N^{-1/2},$$

$$\sigma_j^{-1}N^{-1/2}\boldsymbol{\Sigma}_{j-1}^{-1/2}\tilde{\boldsymbol{Y}}'\tilde{\boldsymbol{\varepsilon}}_{(j)} = \left\{\boldsymbol{I}_{j-1} + \frac{1}{2}\boldsymbol{Z}N^{-1/2} - \frac{1}{8}\boldsymbol{Z}^2N^{-1} + \frac{1}{16}\boldsymbol{Z}^3N^{-3/2}\right\}\boldsymbol{V} + O_p(N^{-5/2}),$$

$$\frac{1}{\sigma_j^2 N}\tilde{\boldsymbol{\varepsilon}}_{(j)}'\tilde{\boldsymbol{\varepsilon}}_{(j)} = 1 + \sqrt{2}XN^{-1/2} + (\boldsymbol{V}'\boldsymbol{V} - j + 1)N^{-1} - \frac{\sqrt{2}}{2}(j-1)XN^{-3/2} + O_p(N^{-5/2})$$

where $\boldsymbol{\Sigma}_j$ is the submatrix of $\boldsymbol{\Sigma} = (\sigma_{ij})$: the entries $(\boldsymbol{\Sigma}_j)_{kl} = \sigma_{kl}$ for $1 \le k, l \le j$,

$$\boldsymbol{Z} := \sqrt{N}\left(\frac{1}{N}\boldsymbol{\Sigma}_{j-1}^{-1/2}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{Y}}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2} - \boldsymbol{I}_{j-1}\right) \xrightarrow{d} N_{j-1\times j-1}(\boldsymbol{0}_{j-1\times j-1}, \boldsymbol{\Omega}),$$

,

$$\boldsymbol{V} := \sigma_j^{-1}\left(\boldsymbol{\Sigma}_{j-1}^{-1/2}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{Y}}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2}\right)^{-1/2}\boldsymbol{\Sigma}_{j-1}^{-1/2}\tilde{\boldsymbol{Y}}_{j-1}'\tilde{\boldsymbol{\epsilon}}_{(j)} \sim N_{j-1}(\boldsymbol{0}, \boldsymbol{I}_{j-1}),$$

$$X := \sqrt{\frac{N-j+1}{2}}\left(\frac{1}{N-j+1}\boldsymbol{U}'\boldsymbol{U} - 1\right) \xrightarrow{d} N(0,1),$$

and $\boldsymbol{U} \sim N_{N-j+1}(\boldsymbol{0}, \boldsymbol{I}_{N-j+1})$, $\boldsymbol{\Omega}$: $\mathrm{Cov}(Z_{ij}, Z_{kl}) = \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}$ and $\boldsymbol{Z}$, $\boldsymbol{V}$, $X(or\boldsymbol{U})$ are mutually independent.

Similarly, we have the following lemma's and theorem's.

Lemma 4.2.

$$\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} = \boldsymbol{\Sigma}_{j-1}^{-1/2}\left\{\boldsymbol{\Sigma}_{j-1}^{1/2}\boldsymbol{\phi}_j + \sigma_j \boldsymbol{V} N^{-1/2} + \left(-\frac{1}{2}\sigma_j \boldsymbol{Z}\boldsymbol{V} - \lambda_j \boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j\right) N^{-1} \right.$$
$$+ \left(\frac{3}{8}\sigma_j \boldsymbol{Z}^2\boldsymbol{V} - \lambda_j\sigma_j\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{V} + \lambda_j \boldsymbol{Z}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j\right) N^{-3/2}$$
$$+ \left(-\frac{5}{16}\sigma_j\boldsymbol{Z}^3\boldsymbol{V} + \frac{1}{2}\lambda_j\sigma_j\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{Z}\boldsymbol{V} + \lambda_j\sigma_j\boldsymbol{Z}\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{V}\right.$$
$$\left.\left. + \lambda_j^2\boldsymbol{\Sigma}_{j-1}^{-3/2}\boldsymbol{\phi}_j - \lambda_j\boldsymbol{Z}^2\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j\right) N^{-2}\right\} + O_p(N^{-5/2}),$$

$$\hat{\sigma}_{j\ PLSE}^2 = \sigma_j^2\left(1 + \sqrt{2}XN^{-1/2} - (j-1)N^{-1}\right.$$
$$\left. - \frac{1}{\sqrt{2}}(j-1)XN^{-3/2} + \frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j N^{-2}\right) + O_p(N^{-5/2}),$$

From Lemma 4.2 and the consistency of $\hat{\sigma}_{1,\mathrm{PLSE}}^2 = \boldsymbol{y}_{(1)}'\boldsymbol{y}_{(1)}/n$, the consistency of $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}$ holds:

Theorem 4.1.

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}} \xrightarrow{p} \boldsymbol{\Sigma}.$$

Note that this theorem holds for arbitrary positive value $\boldsymbol{\lambda}$ and covariance structure.

Theorem 4.2.

$$E\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}))\right] = \frac{p}{2}(p+1)N^{-1} + \left\{\frac{p}{12}\left(10p^2+21p+13\right)\right. \tag{4.4}$$
$$\left. + \sum_{j=2}^{p}\frac{\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j}{\sigma_j^2}\lambda_j^2 - 2\mathrm{tr}\boldsymbol{\Sigma}_{j-1}^{-1}\lambda_j\right\} N^{-2} + O(N^{-5/2}).$$

Theorem 4.3. *Suppose that* $(\sigma_{1j}, \ldots, \sigma_{j-1,j})' \neq \boldsymbol{0}$ *for all* $j = 2, \ldots, p$. *If we use the penalizing parameter* $\boldsymbol{\lambda}_* = (\lambda_{2*}, \ldots, \lambda_{p*})'$, *where*

$$\lambda_{j*} = \mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)\frac{\sigma_j^2}{\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j}, \quad (j = 2, \ldots, p),$$

then the risk can be written as

$$E\left[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_*))\right] = \frac{p}{2}(p+1)N^{-1}$$
$$+ \left\{\frac{p}{12}\left(10p^2+21p+13\right) - \sum_{j=2}^{p}\lambda_{j*}\mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)\right\} N^{-2} + O(N^{-\frac{5}{2}}).$$
$$= \frac{p}{2}(p+1)n^{-1} + \left\{\frac{p}{12}\left(10p^2+27p+19\right) - \sum_{j=2}^{p}\lambda_{j*}\mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)\right\} n^{-2} + O(n^{-\frac{5}{2}}).$$

5 Numerical study

In this section, we compare the performance of the penalizing parameter selection methods, $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\hat{\alpha}_{CV}\mathbf{1}_{p-1})$, $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV})$, and $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*})$. As a benchmark, we also compare them with two existing methods, the sample covariance matrix, and the L_2 penalized maximum likelihood estimator of Huang et al. (2006).

To evaluate the performance, we use $M = 1000$ repeated Monte-Carlo simulations to approximate the mean and the corresponding standard deviation of the the Kullback-Leibler loss

$$E_{\boldsymbol{Y}}[KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}})] \approx \frac{1}{M}\sum_{m=1}^{M} KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}(\boldsymbol{Y}_m)), \tag{5.1}$$

where $\boldsymbol{Y}_m$'s are $n \times p$ observations from a multivariate distribution. The sample size is $n = 100$, the dimensions are $p = 10$, 30, 50, 80 and 120, and two different distributions, multivariate normal (§4.1) and multivariate t with 5 degrees of freedom (§4.2).

We consider the following four covariance structures.

$\boldsymbol{\Sigma}(1) = \mathrm{diag}(1, 2, \ldots, p)$.

$\boldsymbol{\Sigma}(2)$: the compound symmetry $\sigma_{ij} = \rho, \quad i \neq j, \quad \sigma_{ii} = 1, \quad \rho = 0.5$.

$\boldsymbol{\Sigma}(3)$: AR(1), $\quad \sigma_{ij} = \dfrac{\sigma^2}{1-\gamma^2}\gamma^{i-j}, \quad \sigma^2 = 0.01 \quad \gamma = 0.8$.

$\boldsymbol{\Sigma}(4)$: random structure (Onion method).

$\boldsymbol{\Sigma}(1)$ is to check for usefulness of $\boldsymbol{\lambda}_*$ when the covariance structure assumption is not satisfied (§3.31). $\boldsymbol{\Sigma}(2)$ and $\boldsymbol{\Sigma}(3)$ are also considered in Huang et al. (2006). The first has a sparse Cholesky factor, another not sparse. To test the robustness of various covariance structures, $\boldsymbol{\Sigma}(4)$ is changed the structure at random every simulation run. We use *Onion method* for change by Ghosh and Henderson (2003), and Joe (2006) (*R*-package: clusterGeneration).

The penalizing parameter of the L_2 penalized maximum likelihood estimator of Huang et al. (2006) is selected by K-fold cross-validation for $K = 5$. Other estimator, $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\hat{\alpha}_{CV}\mathbf{1}_{p-1})$, $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV})$, and $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*})$ is also selected for $K = 5$. For the optimization about these criterion, we use *optimize* function in R for δ and $\alpha \in (0, 10^4]$.

5.1 Multivariate normal simulations

The results for $\boldsymbol{Y}_m = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$, $\boldsymbol{y}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ are given in Table 1. On $\boldsymbol{\Sigma}(1)$, as expected, $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV})$ and $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*})$ perform

Table 1: *The means and, in parentheses, standard deviations of the the Kullback-Leibler losses for multivariate normal sample.*

				$\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}$		
$\boldsymbol{\Sigma}$	p	$\boldsymbol{S}$	Huang *et al.*	$\hat{\alpha}_{CV}\mathbf{1}_{p-1}$	$\boldsymbol{\lambda}_* : \hat{\alpha}_{CV}$	$\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*}$
	10	0.675(0.005)	0.106(0.002)	0.117(0.002)	0.107(0.002)	0.116(0.002)
	30	8.295(0.030)	0.314(0.003)	0.393(0.003)	0.313(0.003)	0.321(0.003)
$\boldsymbol{\Sigma}(1)$	50	36.443(0.099)	0.522(0.004)	1.076(0.005)	0.526(0.004)	0.530(0.004)
	80	293.633(1.086)	0.841(0.004)	4.181(0.011)	0.903(0.005)	0.903(0.005)
	120	Inf(NA)	1.281(0.005)	17.554(0.026)	1.801(0.007)	1.801(0.007)
	10	0.661(0.005)	0.459(0.004)	0.431(0.004)	0.462(0.004)	0.446(0.004)
	30	8.307(0.030)	2.183(0.010)	1.901(0.009)	2.294(0.015)	1.771(0.010)
$\boldsymbol{\Sigma}(2)$	50	36.154(0.102)	4.196(0.017)	3.547(0.015)	5.323(0.032)	3.172(0.015)
	80	299.830(3.773)	7.516(0.080)	6.219(0.069)	12.261(0.172)	5.476(0.067)
	120	Inf(NA)	11.742(0.040)	9.571(0.035)	24.233(0.109)	8.293(0.033)
	10	0.678(0.005)	0.605(0.005)	0.600(0.005)	0.626(0.005)	0.627(0.005)
	30	8.282(0.029)	5.023(0.017)	4.834(0.016)	4.805(0.016)	4.790(0.016)
$\boldsymbol{\Sigma}(3)$	50	36.491(0.104)	13.888(0.033)	13.000(0.030)	12.752(0.029)	12.435(0.029)
	80	297.095(3.735)	34.273(0.162)	31.195(0.146)	31.511(0.150)	28.907(0.142)
	120	Inf(NA)	70.565(0.076)	63.856(0.069)	68.091(0.064)	58.260(0.066)
	10	0.667(0.005)	0.660(0.005)	0.664(0.005)	0.660(0.005)	0.662(0.005)
	30	8.290(0.029)	8.212(0.031)	8.430(0.032)	7.945(0.029)	7.968(0.029)
$\boldsymbol{\Sigma}(4)$	50	36.261(0.106)	32.881(0.079)	31.981(0.075)	32.291(0.075)	31.383(0.075)
	80	292.741(3.333)	68.895(0.283)	64.173(0.271)	68.503(0.269)	63.620(0.273)
	120	Inf(NA)	111.819(0.088)	106.597(0.087)	113.583(0.086)	106.009(0.088)

better than the sample covariance matrix and $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\hat{\alpha}_{CV}\mathbf{1}_{p-1})$. Althoght $\boldsymbol{\lambda}_*$ may be useful without the covariance structure assumption, the L_2 penalized maximum likelihood estimator of Huang et al. (2006) is the best estimator on $\boldsymbol{\Sigma}(1)$. On other structures, we can see the following trend. Although $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV})$ does good in large sample setting (e.g. $p = 10$), It is not good in high-dimensional setting. In contrast, $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*})$ is better than the others, and relatively stable as the dimension increase. Moreover, our method $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*})$ is the best estimator in high-dimensional setting. Thus, $\boldsymbol{\lambda}_*$ may be useful for high-dimensional estimation.

5.2 Multivariate t_5 simulations

To test the behavior of the methods with a non-normal and non-centered sample, we run the simulations with multivariate t distributions. The sample is taken from

$$\sqrt{\frac{d}{x}}\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{y} + \boldsymbol{\mu} \sim t_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{y} \sim N_p(\mathbf{0}, \boldsymbol{I}_p)$, $x \sim \chi^2_d$, and $\boldsymbol{y}$ and x are independent. After centering the sample, we similarly use the sample as is the case of normal distribution. Note that we must replace $\boldsymbol{\Sigma}$ with $d(d-2)^{-1}\boldsymbol{\Sigma}$ in loss (5.1). Table 2 gives the results for $d = 5$ and $\boldsymbol{\mu} = 3\mathbf{1}_p$. The results exhibit almost similar trends with multivariate normal sample. Thus, our proposed method $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*})$ may be well estimation for non-normal and non-centerd data.

6 Discussion

We propose an estimator of covariance matrix using modified Cholesky decomposition (Pourahmadi, 1999), that using *Ridge regression*, and a method of select a penalizing parameter. The estimator is shown to be better than the sample covariance matrix by simulations and the theoretical result, an asymptotic approximation of the Kullback-Leibler risk. The optimal penalizing parameter is also derived by the approximated risk in large sample asymptotic framework, and could be well estimated in high-dimensional simulations. Moreover, in comparison, the L_2 penalized likelihood estimator by Huang et al. (2006), our method does not need iterative algorithm. The derivation of penalizing parameter in high-dimensional asymptotic framework is a subject for future work.

Table 2: *The means and, in parentheses, standard deviations of the the Kullback-Leibler losses for multivariate t_5 sample.*

				$\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}$		
$\boldsymbol{\Sigma}$	p	$\boldsymbol{S}$	Huang *et al.*	$\hat{\alpha}_{CV}\mathbf{1}_{p-1}$	$\boldsymbol{\lambda}_* : \hat{\alpha}_{CV}$	$\boldsymbol{\lambda}_* : \hat{\alpha}_{CV*}$
	10	1.371(0.013)	0.317(0.007)	0.318(0.007)	0.295(0.006)	0.318(0.007)
	30	15.541(0.074)	0.864(0.013)	1.206(0.015)	0.833(0.012)	0.866(0.013)
$\boldsymbol{\Sigma}(1)$	50	66.285(0.256)	1.454(0.021)	3.789(0.029)	1.506(0.019)	1.516(0.019)
	80	527.750(2.269)	2.325(0.030)	14.499(0.064)	3.134(0.032)	3.134(0.032)
	120	Inf(NA)	3.590(0.046)	54.724(0.143)	8.805(0.059)	8.805(0.059)
	10	1.354(0.012)	0.912(0.010)	0.858(0.009)	0.867(0.009)	0.873(0.010)
	30	15.497(0.074)	3.832(0.028)	3.352(0.028)	3.296(0.028)	3.122(0.027)
$\boldsymbol{\Sigma}(2)$	50	66.466(0.251)	7.209(0.047)	6.170(0.046)	6.917(0.055)	5.555(0.045)
	80	531.129(2.306)	12.344(0.078)	10.411(0.078)	14.553(0.102)	9.164(0.078)
	120	Inf(NA)	19.287(0.111)	16.085(0.112)	28.432(0.173)	13.932(0.116)
	10	1.350(0.012)	1.187(0.011)	1.169(0.011)	1.214(0.011)	1.213(0.011)
	30	15.649(0.076)	8.667(0.046)	8.261(0.045)	8.132(0.044)	8.113(0.044)
$\boldsymbol{\Sigma}(3)$	50	66.427(0.256)	21.559(0.080)	20.111(0.079)	19.116(0.074)	19.142(0.079)
	80	527.065(2.253)	48.709(0.134)	44.916(0.136)	43.193(0.114)	41.655(0.137)
	120	Inf(NA)	91.844(0.178)	85.162(0.192)	85.078(0.137)	77.879(0.202)
	10	1.362(0.012)	1.400(0.017)	1.399(0.013)	1.358(0.012)	1.378(0.013)
	30	15.376(0.072)	15.631(0.080)	16.880(0.077)	16.138(0.078)	15.999(0.078)
$\boldsymbol{\Sigma}(4)$	50	65.541(0.256)	41.825(0.095)	40.096(0.097)	40.832(0.094)	39.890(0.101)
	80	529.655(2.324)	75.599(0.107)	73.036(0.119)	75.436(0.099)	72.878(0.128)
	120	Inf(NA)	117.762(0.112)	115.380(0.131)	118.264(0.100)	115.347(0.143)

Appendix

A.1 Proof of Lemma 3.1.

Since $\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1} \sim W_{j-1}(n, \boldsymbol{\Sigma}_{j-1})$, $\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2} \sim W_{j-1}(n, \boldsymbol{I}_{j-1})$. We use the central limit theorem for Wishart distribution (see section 2.5 in Fujikoshi et al., 2010)

$$\boldsymbol{Z} := \sqrt{n}\left(\frac{1}{n}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2} - \boldsymbol{I}_{j-1}\right) \xrightarrow{d} N_{j-1\times j-1}(\boldsymbol{0}_{j-1\times j-1}, \boldsymbol{\Omega}),$$

where $\boldsymbol{\Omega}$: $\mathrm{cov}(z_{ij}, z_{kl}) = \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk},$, δ_{ij} is Kronecker's delta. From the normal assumption, $\boldsymbol{\varepsilon}_{(j)} \sim N_n(\boldsymbol{0}, \sigma_j^2\boldsymbol{I}_n)$, and $\boldsymbol{\varepsilon}_{(j)}$ and $\boldsymbol{Y}_{j-1}$ are independent. Here, we define a random vector

$$\boldsymbol{V} := \frac{1}{\sigma_j}\left(\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2}\right)^{-1/2}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{\varepsilon}_{(j)} \sim N_{j-1}(\boldsymbol{0}, \boldsymbol{I}_{j-1}),$$

that is independent of $\boldsymbol{Z}$. Thus, we obtain

$$\begin{aligned}\frac{1}{n}\boldsymbol{Y}_{j-1}'\boldsymbol{\varepsilon}_{(j)} &= \sigma_j n^{-1/2}\boldsymbol{\Sigma}_{j-1}^{1/2}\left(\boldsymbol{I}_{j-1} + \boldsymbol{Z}n^{-1/2}\right)^{1/2}\boldsymbol{V} \\ &= \sigma_j n^{-1/2}\boldsymbol{\Sigma}_{j-1}^{1/2}\left(\boldsymbol{I}_{j-1} + \frac{1}{2}\boldsymbol{Z}n^{-1/2} - \frac{1}{8}\boldsymbol{Z}^2n^{-1} + \frac{1}{16}\boldsymbol{Z}^3n^{-3/2}\right)\boldsymbol{V} + O_p(n^{-5/2}).\end{aligned}$$

Here, we consider an orthogonal transformation,

$$\tilde{\boldsymbol{\varepsilon}}_{(j)} := \frac{1}{\sigma_j}\boldsymbol{H}\boldsymbol{\varepsilon}_{(j)} = \begin{pmatrix}\boldsymbol{V}\\ \boldsymbol{U}\end{pmatrix}, \tag{A.1.1}$$

where

$$\boldsymbol{H} = \begin{pmatrix}\left(\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\boldsymbol{Y}_{j-1}\boldsymbol{\Sigma}_{j-1}^{-1/2}\right)^{-1/2}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{Y}_{j-1}'\\ \boldsymbol{H}_2\end{pmatrix},$$

and $\boldsymbol{H}_2$ is a $n-j+1\times n-j+1$ orthogonal matrix. From $\sigma_j^{-1}\boldsymbol{\epsilon}_{(j)} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$, $\tilde{\boldsymbol{\varepsilon}}_{(j)} \sim N_n(\boldsymbol{0}, \boldsymbol{I}_n)$. Then $\boldsymbol{U}$ and $\boldsymbol{V}$ are independent, and $\boldsymbol{U}'\boldsymbol{U} \sim \chi^2(n-j+1)$. Therefore, from the central limit theorem,

$$X := \sqrt{\frac{n-j+1}{2}}\left(\frac{1}{n-j+1}\boldsymbol{U}'\boldsymbol{U} - 1\right) \xrightarrow{d} N(0,1).$$

Then it holds that

$$\begin{aligned}
\frac{1}{n}\boldsymbol{\varepsilon}'_{(j)}\boldsymbol{\varepsilon}_{(j)} &= \frac{1}{n}\sigma_j^2\left(\boldsymbol{V}'\boldsymbol{V}+\boldsymbol{U}'\boldsymbol{U}\right), \\
&= \sigma_j^2\left\{\boldsymbol{V}'\boldsymbol{V}n^{-1}+\sqrt{2}Xn^{-1/2}\sqrt{1-\frac{j-1}{n}}-(j-1)n^{-1}+1\right\} \\
&= \sigma_j^2\left\{\boldsymbol{V}'\boldsymbol{V}n^{-1}+\sqrt{2}Xn^{-1/2}\left(1-\frac{1}{2}(j-1)n^{-1}\right)-(j-1)n^{-1}+1\right\}+O_p(n^{-5/2}).
\end{aligned}$$

A.2 Proof of Lemma 3.3.

We partition the modified Cholesky decomposition of covariance matrix or its estimator:

$$\begin{aligned}
\boldsymbol{\Sigma} &= \boldsymbol{T}^{-1}\boldsymbol{D}^2(\boldsymbol{T}')^{-1}, \\
\begin{pmatrix} \boldsymbol{\Sigma}_{p-1} & \boldsymbol{\sigma}_p \\ \boldsymbol{\sigma}'_p & \sigma_{pp} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{T}_{p-1} & \boldsymbol{0} \\ -\boldsymbol{\phi}'_p & 1 \end{pmatrix}^{-1}\begin{pmatrix} \boldsymbol{D}^2_{p-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \sigma_p^2 \end{pmatrix}\begin{pmatrix} \boldsymbol{T}'_{p-1} & -\boldsymbol{\phi}_p \\ \boldsymbol{0}' & 1 \end{pmatrix}^{-1} \\
&= \begin{pmatrix} \boldsymbol{T}^{-1}_{p-1} & \boldsymbol{0} \\ \boldsymbol{\phi}'_p\boldsymbol{T}^{-1}_{p-1} & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{D}^2_{p-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \sigma_p^2 \end{pmatrix}\begin{pmatrix} (\boldsymbol{T}'_{p-1})^{-1} & (\boldsymbol{T}'_{p-1})^{-1}\boldsymbol{\phi}_p \\ \boldsymbol{0}' & 1 \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{T}^{-1}_{p-1}\boldsymbol{D}^2_{p-1}(\boldsymbol{T}'_{p-1})^{-1} & \boldsymbol{T}^{-1}_{p-1}\boldsymbol{D}^2_{p-1}(\boldsymbol{T}'_{p-1})^{-1}\boldsymbol{\phi}_p \\ \boldsymbol{\phi}'_p\boldsymbol{T}^{-1}_{p-1}\boldsymbol{D}^2_{p-1}(\boldsymbol{T}'_{p-1})^{-1} & \boldsymbol{\phi}'_p\boldsymbol{T}^{-1}_{p-1}\boldsymbol{D}^2_{p-1}(\boldsymbol{T}'_{p-1})^{-1}\boldsymbol{\phi}_p+\sigma_p^2 \end{pmatrix} \\
&= \begin{pmatrix} \boldsymbol{\Sigma}_{p-1} & \boldsymbol{\Sigma}_{p-1}\boldsymbol{\phi}_p \\ \boldsymbol{\phi}'_p\boldsymbol{\Sigma}_{p-1} & \boldsymbol{\phi}'_p\boldsymbol{\Sigma}_{p-1}\boldsymbol{\phi}_p+\sigma_p^2 \end{pmatrix},
\end{aligned}$$

where $\boldsymbol{T}_{p-1}$ and $\boldsymbol{D}^2_{p-1}$ are the submatrix of $\boldsymbol{T}$ and $\boldsymbol{D}^2$, respectively. We also partition the modified Cholesky decomposition of the inverse covariance matrix.

$$\begin{aligned}
\boldsymbol{\Sigma}^{-1} &= \boldsymbol{T}'\boldsymbol{D}^{-2}\boldsymbol{T}, \\
\begin{pmatrix} \boldsymbol{\Sigma}^{p-1} & \boldsymbol{\sigma}^p \\ \boldsymbol{\sigma}^{p'} & \sigma^{pp} \end{pmatrix} &= \begin{pmatrix} \boldsymbol{T}'_{p-1} & -\boldsymbol{\phi}_p \\ \boldsymbol{0}' & 1 \end{pmatrix}\begin{pmatrix} \boldsymbol{D}^{-2}_{p-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \sigma_p^{-2} \end{pmatrix}\begin{pmatrix} \boldsymbol{T}_{p-1} & \boldsymbol{0} \\ -\boldsymbol{\phi}'_p & 1 \end{pmatrix}, \\
&= \begin{pmatrix} \boldsymbol{T}'_{p-1}\boldsymbol{D}^{-2}_{p-1}\boldsymbol{T}_{p-1}+\sigma_p^{-2}\boldsymbol{\phi}_p\boldsymbol{\phi}'_p & -\sigma_p^{-2}\boldsymbol{\phi}_p \\ -\sigma_p^{-2}\boldsymbol{\phi}'_p & \sigma_p^{-2} \end{pmatrix}, \\
&= \begin{pmatrix} \boldsymbol{\Sigma}^{-1}_{p-1}+\sigma_p^{-2}\boldsymbol{\phi}_p\boldsymbol{\phi}'_p & -\sigma_p^{-2}\boldsymbol{\phi}_p \\ -\sigma_p^{-2}\boldsymbol{\phi}'_p & \sigma_p^{-2} \end{pmatrix}.
\end{aligned}$$

The same argument can be considered for $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\Sigma}}^{-1}$. Therefore, it holds that

$$
\begin{aligned}
KL(\boldsymbol{\Sigma}, \hat{\boldsymbol{\Sigma}}) &= \mathrm{tr}\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1} - \log\left|\boldsymbol{\Sigma}\hat{\boldsymbol{\Sigma}}^{-1}\right| - p \\
&= \mathrm{tr}\left\{ \begin{pmatrix} \boldsymbol{\Sigma}_{p-1} & \boldsymbol{\Sigma}_{p-1}\boldsymbol{\phi}_p \\ \boldsymbol{\phi}_p'\boldsymbol{\Sigma}_{p-1} & \boldsymbol{\phi}_p'\boldsymbol{\Sigma}_{p-1}\boldsymbol{\phi}_p + \sigma_p^2 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{\Sigma}}_{p-1}^{-1} + \hat{\sigma}_p^{-2}\hat{\boldsymbol{\phi}}_p\hat{\boldsymbol{\phi}}_p' & -\hat{\sigma}_p^{-2}\hat{\boldsymbol{\phi}}_p \\ -\hat{\sigma}_p^{-2}\hat{\boldsymbol{\phi}}_p' & \hat{\sigma}_p^{-2} \end{pmatrix} \right\} \\
&\quad - \log\left| \begin{pmatrix} \boldsymbol{D}_{p-1}^2 & \mathbf{0} \\ \mathbf{0}' & \sigma_p^2 \end{pmatrix} \begin{pmatrix} \hat{\boldsymbol{D}}_{p-1}^2 & \mathbf{0} \\ \mathbf{0}' & \hat{\sigma}_p^2 \end{pmatrix}^{-1} \right| - (p-1) - 1 \\
&= KL(\boldsymbol{\Sigma}_{p-1}, \hat{\boldsymbol{\Sigma}}_{p-1}) + \frac{1}{\hat{\sigma}_p^2}\left(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}_p\right)'\boldsymbol{\Sigma}_{p-1}\left(\hat{\boldsymbol{\phi}}_p - \boldsymbol{\phi}_p\right) + \frac{\sigma_p^2}{\hat{\sigma}_p^2} - \log\frac{\sigma_p^2}{\hat{\sigma}_p^2} - 1.
\end{aligned}
$$

We can use the same argument for $j = p-1, p-2, \ldots, 2$.

A.3 Proof of Lemma 3.4

For simplicity, we write $\hat{\boldsymbol{\Sigma}}_{\mathrm{PLSE}}(\boldsymbol{\lambda}) = \hat{\boldsymbol{\Sigma}}$, $\hat{\boldsymbol{\phi}}_{j,\mathrm{PLSE}} = \hat{\boldsymbol{\phi}}_j$, and $\hat{\sigma}_{j,\mathrm{PLSE}}^2 = \hat{\sigma}_j^2$. From Lemma 3.2, it holds that

$$
\begin{aligned}
&\left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_j\right)'\boldsymbol{\Sigma}_{j-1}\left(\boldsymbol{\phi}_j - \hat{\boldsymbol{\phi}}_j\right) \qquad \text{(A.3.2)} \\
&= \sigma_j^2\left\{\boldsymbol{V}'\boldsymbol{V}n^{-1} + \left(-\boldsymbol{V}'\boldsymbol{Z}\boldsymbol{V} - \frac{2\lambda_j}{\sigma_j}\boldsymbol{V}'\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j\right)n^{-3/2}\right. \\
&\left. + \left(\boldsymbol{V}'\boldsymbol{Z}^2\boldsymbol{V} - 2\lambda_j\boldsymbol{V}'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{V} + \frac{3\lambda_j}{\sigma_j}\boldsymbol{V}'\boldsymbol{Z}\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_j + \frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j\right)n^{-2}\right\} + O_p(n^{-\frac{5}{2}}), \\
&\frac{\sigma_j^2}{\hat{\sigma}_j^2} = 1 - \sqrt{2}Xn^{-1/2} + (2X^2 + j - 1)n^{-1} + \left(-2\sqrt{2}X^3 - \frac{3\sqrt{2}}{2}(j-1)X\right)n^{-3/2} \\
&\qquad + \left(4X^4 + 4(j-1)X^2 + (j-1)^2 - \frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j\right)n^{-2} + O_p(n^{-5/2}), \qquad \text{(A.3.3)} \\
&\frac{\sigma_j^2}{\hat{\sigma}_j^2} - \log\frac{\sigma_j^2}{\hat{\sigma}_j^2} = 1 + X^2n^{-1} + \left\{-\sqrt{2}(j-1)X - \frac{4\sqrt{2}}{3}X^3\right\}n^{-3/2} \\
&\qquad + \left\{3X^4 + 3(j-1)X^2 + \frac{1}{2}(j-1)^2\right\}n^{-2} + O_p(n^{-5/2}). \qquad \text{(A.3.4)}
\end{aligned}
$$

Here, (A.3.3) and (A.3.4) are functions of X, and (A.3.2) is a function of $\boldsymbol{Z}$ and $\boldsymbol{V}$. Thus,

$$E\left[\frac{1}{\hat{\sigma}_j^2}\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)'\boldsymbol{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)+\frac{\sigma_j^2}{\hat{\sigma}_j^2}-\log\frac{\sigma_j^2}{\hat{\sigma}_j^2}-1\right]$$
$$=E_X\left[\frac{1}{\hat{\sigma}_j^2}\right]E_{\boldsymbol{V},\boldsymbol{Z}}\left[\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)'\boldsymbol{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)\right]+E_X\left[\frac{\sigma_j^2}{\hat{\sigma}_j^2}-\log\frac{\sigma_j^2}{\hat{\sigma}_j^2}-1\right].$$

Note that from the definition of X,

$$E[X]=0,\quad E[X^2]=1,$$
$$E[X^3]=2\sqrt{2}n^{-\frac{1}{2}}+O(n^{-1}),\quad E[X^4]=3+O(n^{-1}).$$

Hence

$$E_X\left[\frac{1}{\hat{\sigma}_j^2}\right]=\frac{1}{\sigma_j^2}\left\{1+(j+1)n^{-1}+\left((j+1)^2-\frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j\right)n^{-2}\right\}+O(n^{-\frac{5}{2}}),$$
$$E_X\left[\frac{\sigma_j^2}{\hat{\sigma}_j^2}-\log\frac{\sigma_j^2}{\hat{\sigma}_j^2}-1\right]=n^{-1}+\left(\frac{1}{2}j^2+2j+\frac{7}{6}\right)n^{-2}+O(n^{-\frac{5}{2}}).$$

Furthermore, from the distribution of $\boldsymbol{Z}$ and $\boldsymbol{V}$,

$$E[\boldsymbol{Z}]=\boldsymbol{0}_{j-1\times j-1},\quad E[\boldsymbol{Z}^2]=j\boldsymbol{I}_{j-1},$$
$$\boldsymbol{V}'\boldsymbol{V}\sim\chi^2(j-1),\quad \boldsymbol{V}\boldsymbol{V}'\sim W_{j-1}(1,\boldsymbol{I}_{j-1}).$$

Then

$$\begin{aligned}
&E_{\boldsymbol{V},\boldsymbol{Z}}\left[\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)'\boldsymbol{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)\right]\\
=\;&E_{\boldsymbol{V}}\left[E_{\boldsymbol{Z}}\left[\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)'\boldsymbol{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)\middle|\boldsymbol{V}\right]\right]\\
=\;&\sigma_j^2E_{\boldsymbol{V}}\left[\boldsymbol{V}'\boldsymbol{V}n^{-1}-\frac{2\lambda_j}{\sigma_j}\boldsymbol{V}'\boldsymbol{\Sigma}_{j-1}^{-1/2}\boldsymbol{\phi}_jn^{-3/2}\right.\\
&\left.+\left(j\boldsymbol{V}'\boldsymbol{V}-2\lambda_j\mathrm{tr}\left(\boldsymbol{V}\boldsymbol{V}'\boldsymbol{\Sigma}_{j-1}^{-1}\right)+\frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j\right)n^{-2}\middle|\boldsymbol{V}\right]+O(n^{-5/2})\\
=\;&\sigma_j^2\left\{(j-1)n^{-1}+\left(j(j-1)-2\lambda_j\mathrm{tr}\left(\boldsymbol{\Sigma}_{j-1}^{-1}\right)+\frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j\right)n^{-2}\right\}+O(n^{-5/2})
\end{aligned}$$

Therefore,

$$
\begin{aligned}
& E\left[\frac{1}{\hat{\sigma}_j^2}\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)'\boldsymbol{\Sigma}_{j-1}\left(\hat{\boldsymbol{\phi}}_j-\boldsymbol{\phi}_j\right)+\frac{\sigma_j^2}{\hat{\sigma}_j^2}-\log\frac{\sigma_j^2}{\hat{\sigma}_j^2}-1\right] \\
= & \left\{1+(j+1)n^{-1}+O(n^{-2})\right\}\left\{(j-1)n^{-1}+\left(j(j-1)-2\lambda_j\mathrm{tr}\boldsymbol{\Sigma}_{j-1}^{-1}+\frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j\right)n^{-2}\right\} \\
& +n^{-1}+\left(\frac{1}{2}j^2+2j+\frac{7}{6}\right)n^{-2}+O(n^{-5/2}) \\
= & jn^{-1}+\left(\frac{5}{2}j^2+j+\frac{1}{6}-2\lambda_j\mathrm{tr}\boldsymbol{\Sigma}_{j-1}^{-1}+\frac{\lambda_j^2}{\sigma_j^2}\boldsymbol{\phi}_j'\boldsymbol{\Sigma}_{j-1}^{-1}\boldsymbol{\phi}_j\right)n^{-2}+O(n^{-5/2})
\end{aligned}
$$

References

[1] Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.*, **36**, 199–227.

[2] Chang, C. and Tsay, R. S. (2010). Estimation of covariance matrix via the sparse Cholesky factor with lasso. *J. Statist. Plann. Inference.*, **140**, 3858–3873.

[3] Fujikoshi, Y., Ulyanv, V. V. and Shimizu, R. (2010). *Multivariate Statistics: High-dimensional and Large-Sample Approximations.* Wiley, New Jersey.

[4] Ghosh, S. and Henderson, S. G. (2003). Behavior of the norta method for correlated random vector generation as the dimension increases. *TOMACS* **13**, 3, 276–294.

[5] Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression. Applications to nonorthogonal problems. *Technometrics*, **12**, 55–67.

[6] Huang, J. Z., Liu, N., Pourahmadi, M. and Liu, L. (2006). Covariance matrix selection and estimation via penalised normal likelihood. *Biometrika*, **93**, 85–98.

[7] Joe, H. (2006) Generating Random Correlation Matrices Based on Partial Correlations. *J. Multivariate, Anal.* **97**, 2177–2189.

[8] Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *Ann. Statist.* **29** 295-327.

[9] Ledoit, O. and Wolf, M. (2003). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivariate, Anal.*, **88**, 365–411.

[10] Pourahmadi, M. (2013). *High-dimensional Covariance Estimation.* Wiley, New Jersey. Levina, E., Rothman, A. and Zhu, J. (2008). Sparse estimation of large covariance matrices via a nested Lasso penalty. *Ann. Appl. Statist.*, **2**, 245–263.

[11] Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, **86** 677–690.

[12] Pourahmadi, M. (2013). *High-dimensional Covariance Estimation.* Wiley, New Jersey.

[13] Stein, C. (1975). Estimation of a covariance matrix. In *Rietz Lecture. 39th Annual Meeting IMS. Atlanta, Georgia.*