PREDICTIVE ESTIMATION OF A COVARIANCE MATRIX AND ITS STRUCTURAL PARAMETERS

Haruhiko Ogasawara*

ABSTRACT

Methods of estimating an unconstrained covariance matrix are derived using future data as well as current data in the likelihood. The estimators are obtained by optimizing coefficient(s) for adjusting the usual maximum likelihood estimators based on current data. The optimization is given by maximizing the expected log-likelihood over the distributions of future and current data. Under the Wishart and normal distributions, the coefficients in their adjusted estimators are obtained using known quantities. When a covariance matrix is structured with structural parameters, asymptotic adjustments of the Wishart maximum likelihood estimators are obtained. Similar estimators of an unconstrained covariance matrix derived by minimizing the mean square error are also given. Numerical illustrations with simulations are shown using factor analysis models. Methods of overcoming the problem of the dependence of the optimal values on unknown population values are discussed.

1. Introduction

For estimation of parameters in statistical models, methods of maximum likelihood (ML), penalized ML, Bayes and weighted score (WS) are typically used when given current data, where the WS method includes the remaining ones when the posterior mode is used in the case of Bayes estimation (see, e.g., Ogasawara, 2014, 2016a). Additional information other than current data is also used when an informative prior distribution is used in Bayes estimation. Even in this case, data other than current ones are not explicitly used although we have prior distributions based on historical data (Ibrahim & Chen, 2000; Ibrahim, Chen & Sinha, 2003). When an informative prior is used, the prior may typically be based on past data or experiences.

In cross-validation, two sets of data are used, i.e., one for estimation, training or calibration and the other for evaluation, testing or validation although single-sample cross-validation indexes are also available (see, e.g., Browne, 2000). From a predictive viewpoint, the data for validation are seen as future data to be predicted. The cross-validation method is used primarily for selecting appropriate models. Stone (1977) showed that a cross-validation index is asymptotically equal to the Takeuchi information criterion (TIC; Takeuchi, 1976), which is an extension of the Akaike information criterion (AIC; Akaike, 1973). That is, the AIC and TIC can also be seen as single-sample counterparts of cross-validation indexes.

The AIC and TIC are bias-corrected estimators of minus 2 times the two-fold expected log-likelihood, where the two-fold expectation indicates one for current data and the other for independent future data. The bias-corrected estimators are seen as those corrected from the

*Otaru University of Commerce, 3-5-21, Midori, Otaru 047-8501, Japan

E-mail: emt-hogasa@emt.otaru-uc.ac.jp.

Key words: covariance structure; factor analysis; predictive estimators; expected likelihood; mean square error; Kullback-Leibler distance.

predictive viewpoint. This type of parameter estimators in statistical distributions has been given by Takezawa (2012a, 2015a, b, c) for the variance in a univariate normal distribution and the parameters in the exponential, binomial and geometric distributions, respectively, which are called predictive estimators.

The predictive estimators are obtained by maximizing the likelihood, where data are given by future data and parameter estimators are given by the usual ML estimator (MLE) based on current data with a tuning parameter. The expectation of the log-likelihood over the future and current data is taken, followed by maximizing the expectation with respect to the tuning parameter. Ogasawara (2017) derived a general expression of the predictive estimators maximizing the asymptotically expected predictive log-likelihood corresponding to the above expectation, and gives various asymptotic properties of the predictive estimators including a comparison with estimators minimizing mean square errors (MSEs).

A class of estimators of a covariance matrix under Stein's loss (James and Stein, 1961, Equation (72)) with the assumption of normality is different from those by ML and MSEs, and has been given by James and Stein (1961, Section 5), Dey and Srinivasan (1985), Sheena and Takemura (1992), and Ye and Wang (2009). Note, however, that in this class of estimators, a predictive viewpoint is not employed as in the predictive estimators mentioned above.

The purposes of this paper are to derive predictive estimators of a covariance matrix, which is seen as an extension of Takezawa (2012a), and predictive estimators of parameters in a covariance structure model. Numerical examples for these estimators including simulations are also shown.

2. Predictive estimation under the assumption of normality

2.1. Predictive estimation of an unconstrained covariance matrix based on the Wishart distribution

Let l be the log-likelihood of a $q \times 1$ parameter vector $\boldsymbol{\theta}$ based on N independent observations for p observable variables summarized by an $N \times p$ matrix $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)'$, where $\mathbf{x}_i$ is the i -th vector variable or its realization for simplicity of notation ($i = 1, \dots, N$), and let $\mathbf{U} = (\mathbf{u}_1, \dots, \mathbf{u}_N)'$ be an $N \times r$ fixed design matrix for r covariates:

$$l = \sum_{i=1}^N \ln f(\mathbf{x}_i | \boldsymbol{\theta}, \mathbf{u}_i), \quad (2.1)$$

where $f(\mathbf{x}_i | \boldsymbol{\theta}, \mathbf{u}_i)$ is the density or probability mass for $\mathbf{x}_i$ when $\boldsymbol{\theta}$ and $\mathbf{u}_i$ are given. Let $\bar{l} = N^{-1}l$, which is the log-likelihood averaged over observations.

Assume that $\mathbf{x}_i \sim N(\mathbf{B}_0 \mathbf{u}_i, \boldsymbol{\Sigma}_0)$, i.e., $\mathbf{x}_i$ follows the normal distribution with $E(\mathbf{x}_i) = \mathbf{B}_0 \mathbf{u}_i$ and $\text{cov}(\mathbf{x}_i) = \boldsymbol{\Sigma}_0$ ($i = 1, \dots, N$), where $\mathbf{B}_0$ is the $p \times r$ matrix of population regression coefficients. Let $\mathbf{S} = n^{-1} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mathbf{x}}_i)(\mathbf{x}_i - \hat{\mathbf{x}}_i)'$, where $n = N - r$, $\hat{\mathbf{x}}_i = \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{u}_i$ is the i -column of $\hat{\mathbf{X}}' = \mathbf{X}'\mathbf{U}(\mathbf{U}'\mathbf{U})^{-1}\mathbf{U}'$ projected on the column space of $\mathbf{U}$. Note that when $\mathbf{x}_i \sim N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ with $\boldsymbol{\mu}_0 = E(\mathbf{x}_i)$ ($i = 1, \dots, N$), we have $n = N - 1$ and $\hat{\mathbf{x}}_i = \bar{\mathbf{x}}$ with $\bar{\mathbf{x}} = N^{-1} \sum_{i=1}^N \mathbf{x}_i$. That is, $\mathbf{S}$ with $n = N - r$ is an unbiased estimator of $\boldsymbol{\Sigma}_0$. Then, $\mathbf{S} \sim W_p(n^{-1}\boldsymbol{\Sigma}_0, n)$, which indicates that $\mathbf{S}$ follows the Wishart distribution with a scale parameter $n^{-1}\boldsymbol{\Sigma}_0$ and n degrees of freedom, whose density is

$$f(\mathbf{S} | \boldsymbol{\Sigma}, n) = \frac{|\mathbf{S}|^{(n-p-1)/2} \exp\{-\text{tr}(n\boldsymbol{\Sigma}^{-1}\mathbf{S})/2\}}{(2/n)^{pn/2} |\boldsymbol{\Sigma}|^{n/2} \Gamma_p(n/2)}, \quad (2.2)$$

where $\Gamma_p(\cdot)$ is the p -variate gamma function given by

$$\Gamma_p(t) = \pi^{p(p-1)/4} \prod_{i=1}^p \Gamma\{t - (1/2)(i-1)\} \quad (2.3)$$

and $\Gamma(\cdot)$ is the usual gamma function.

Let us temporarily change the definition of $\bar{l} = N^{-1}l$ to $\bar{l} = n^{-1}l$ in the case of the Wishart distribution. Note that n is the effective number of observations in this distribution since (2.2) holds when $\mathbf{S}$ is replaced by $n^{-1} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu}_0)(\mathbf{x}_i - \boldsymbol{\mu}_0)'$ using n observations when $\boldsymbol{\mu}_0$ is known. Then,

$$-2\bar{l} = -2n^{-1}l = \ln |\boldsymbol{\Sigma}| + \text{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}) + C(\mathbf{S}) \quad (2.4)$$

with $C(\mathbf{S}) = -\frac{n-p-1}{n} \ln |\mathbf{S}| + p \ln(2/n) + \frac{2}{n} \ln \Gamma_p(n/2)$.

Let $\boldsymbol{\theta} = \boldsymbol{\sigma} = \mathbf{v}(\boldsymbol{\Sigma})$, where $\mathbf{v}(\cdot)$ is the vectorizing operator taking the non-duplicated elements of a symmetric matrix with its population value $\boldsymbol{\theta}_0 = \boldsymbol{\sigma}_0 = \mathbf{v}(\boldsymbol{\Sigma}_0)$. A corresponding estimator of $\boldsymbol{\theta}_0$ based on current data or $\mathbf{S}$ is denoted by $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\sigma}} = \mathbf{v}(\hat{\boldsymbol{\Sigma}})$. Note that the Wishart MLE of $\boldsymbol{\Sigma}_0$ is given by $\hat{\boldsymbol{\theta}}_{\text{WML}} = \mathbf{s} = \mathbf{v}(\mathbf{S})$. Let $\mathbf{T}$ be an independent copy of $\mathbf{S}$ corresponding to future data. Then, $\hat{l}^*$ is defined as

$$\hat{l}^* = \mathbf{E}^{(\mathbf{T})}(\hat{l}) = \mathbf{E}^{(\mathbf{T})}\{\bar{l}(\hat{\boldsymbol{\theta}}|\mathbf{T})\} = \mathbf{E}^{(\mathbf{T})}[\bar{l}\{\boldsymbol{\theta}(\mathbf{S})|\mathbf{T}\}], \quad (2.5)$$

where $\mathbf{E}^{(\mathbf{T})}(\cdot)$ indicates that the expectation is taken over the distribution of $\mathbf{T}$ independent of $\mathbf{S}$. In the case of the Wishart distribution, (2.5) gives

$$-2\hat{l}^* = \mathbf{E}^{(\mathbf{T})}\{\ln |\hat{\boldsymbol{\Sigma}}| + \text{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{T}) + C(\mathbf{T})\} = \ln |\hat{\boldsymbol{\Sigma}}| + \text{tr}(\hat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}_0) + \mathbf{E}^{(\mathbf{T})}\{C(\mathbf{T})\}, \quad (2.6)$$

where

$$\begin{aligned} \mathbf{E}^{(\mathbf{T})}\{C(\mathbf{T})\} &= \mathbf{E}^{(\mathbf{S})}\{C(\mathbf{S})\} \\ &= -\frac{n-p-1}{n} \mathbf{E}^{(\mathbf{T})}(\ln |\mathbf{T}|) + p \ln(2/n) + \frac{2}{n} \ln \Gamma_p(n/2) \\ &= -\frac{n-p-1}{n} \{\ln |n^{-1}\boldsymbol{\Sigma}_0| + p \ln 2 + \psi_p(n/2)\} + p \ln(2/n) + \frac{2}{n} \ln \Gamma_p(n/2) \\ &= -\frac{n-p-1}{n} \{\ln |\boldsymbol{\Sigma}_0| + p \ln(2/n) + \psi_p(n/2)\} + p \ln(2/n) + \frac{2}{n} \ln \Gamma_p(n/2) \end{aligned} \quad (2.7)$$

and $\psi_p(\cdot)$ is the p -variate digamma function or the first derivative of $\ln \Gamma_p(\cdot)$. In (2.7), $\mathbf{E}^{(\mathbf{T})}(\ln |\mathbf{T}|)$ is derived by the property of a natural parameter $n/2$ in the Wishart distribution that belongs to the exponential family of distributions.

The problem of constructing a predictive estimator, denoted by $\hat{\boldsymbol{\theta}}_{\text{W}}^{(\text{P})}$ based on the Wishart MLE $\hat{\boldsymbol{\theta}}_{\text{WML}} = \mathbf{s}$, is reduced to maximizing $\mathbf{E}^{(\mathbf{S})}(\hat{l}^*)$, where $\hat{\boldsymbol{\Sigma}}$ in $\hat{l}^*$ is replaced by $\hat{\boldsymbol{\Sigma}}_{\text{Wk}}^{(\text{P})} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_{\text{Wk}}^{(\text{P})})$, with respect to a tuning parameter k in $\hat{\boldsymbol{\Sigma}}_{\text{Wk}}^{(\text{P})} = k\mathbf{S}$. This gives the following result.

Theorem 1. *The optimal multiplicative constant k in the predictive estimator $\hat{\boldsymbol{\Sigma}}_{\text{Wk}}^{(\text{P})} = k\mathbf{S}$ maximizing $\mathbf{E}^{(\mathbf{S})}(\hat{l}^*)$ based on the Wishart distribution is*

$$k = \frac{n}{n-p-1} \quad (n > p+1). \quad (2.8)$$

The proof is given in the supplement to this paper (Ogasawara, 2018, Subsection A.1). From (2.8), the optimal predictive estimator is derived as

$$\hat{\boldsymbol{\Sigma}}_{\text{Wk}}^{(\text{P})} = \frac{n}{n-p-1} \mathbf{S} = \frac{1}{n-p-1} \sum_{i=1}^N (\mathbf{x}_i - \hat{\mathbf{x}}_i)(\mathbf{x}_i - \hat{\mathbf{x}}_i)'. \quad (2.9)$$

When $p = 1$ and $\hat{\mathbf{x}}_i = \bar{\mathbf{x}}$ ($i = 1, \dots, N$) with $n = N - 1$, we have

$$\hat{\sigma}_{\mathbf{W}k}^{2(\text{P})} = \frac{1}{N-3} \sum_{i=1}^N (x_i - \bar{x})^2. \quad (2.10)$$

Note that Takezawa's (2012a) "the third variance" (see also Takezawa, 2012b, Section 5.5) or the predictive estimator based on the univariate normal distribution is defined as

$$\frac{1}{N-4} \sum_{i=1}^N (x_i - \bar{x})^2. \quad (2.11)$$

It is seen that (2.10) is smaller than (2.11). Note also that for general p , (2.8) depends on p as well as n , which stems from the dependence of each element of $\mathbf{E}^{(\text{S})}(\mathbf{S}^{-1})$ on p (see Subsection A.1 of the appendix), while each element of $\mathbf{E}^{(\text{S})}(\mathbf{S})$ is independent of p .

Stein's loss mentioned in Section 1 is defined by

$$L(\hat{\Sigma}, \Sigma_0) = \text{tr}(\hat{\Sigma}\Sigma_0^{-1}) - \ln |\hat{\Sigma}\Sigma_0^{-1}| - p. \quad (2.12)$$

where $\hat{\Sigma}$ is an estimator of Σ_0 . Consider an estimator $k\mathbf{S}$ as $\hat{\Sigma}_{\mathbf{W}k}^{(\text{P})} = k\mathbf{S}$ in Theorem 1. It is known that under normality $\mathbf{E}^{(\text{S})}\{L(k\mathbf{S}, \Sigma_0)\}$ is minimized when $k = 1$ (e.g., Ye and Wang, 2009, p.716). Noting that under non-normality

$$\mathbf{E}^{(\text{S})}\{L(k\mathbf{S}, \Sigma_0)\} = p(k-1) - \ln k^p - \mathbf{E}^{(\text{S})}\{\ln |\mathbf{S}\Sigma_0^{-1}|\} \quad (2.13)$$

and taking the derivative of (2.13) with respect to k , we find that $k = 1$ is optimal under arbitrary distributions. The class of estimators minimizing the expectation of Stein's loss under normality is different from that given by $k\mathbf{S}$ and tends to be complicated.

2.2. Predictive estimation of an unconstrained covariance matrix based on the multivariate normal distribution

Assume the condition in Subsection 2.1 without $\mathbf{U}$, i.e., $\mathbf{x}_i \sim N(\boldsymbol{\mu}_0, \Sigma_0)$ ($i = 1, \dots, N$). For the additional parameter vector $\boldsymbol{\mu}$ in the multivariate normal density, we use the usual estimator $\bar{\mathbf{x}}$, giving

$$f(\mathbf{X}|\Sigma, \hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}) = \frac{1}{(2\pi)^{Np/2} |\Sigma|^{N/2}} \exp \left\{ - \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})' \Sigma^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}) / 2 \right\}. \quad (2.14)$$

Let $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)'$ be an independent copy of $\mathbf{X}$. Then,

$$\begin{aligned} -2\hat{l}^* &= p \ln(2\pi) + \ln |\hat{\Sigma}| + \text{tr}[\hat{\Sigma}^{-1} \mathbf{E}^{(\text{Z})}\{N^{-1} \sum_{i=1}^N (\mathbf{z}_i - \bar{\mathbf{x}})(\mathbf{z}_i - \bar{\mathbf{x}})'\}] \\ &= p \ln(2\pi) + \ln |\hat{\Sigma}| + \text{tr}[\hat{\Sigma}^{-1}\{\Sigma_0 + (\bar{\mathbf{x}} - \boldsymbol{\mu}_0)(\bar{\mathbf{x}} - \boldsymbol{\mu}_0)'\}]. \end{aligned} \quad (2.15)$$

Let

$$\hat{\Sigma}_{\text{N}k}^{(\text{P})} = k\hat{\Sigma}_{\text{NML}} = kN^{-1} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \quad (2.16)$$

where N and NML indicate the normal distribution and MLE using this distribution, respectively. Then,

Theorem 2. *The optimal k in $\hat{\Sigma}_{\text{N}k}^{(\text{P})} = k\hat{\Sigma}_{\text{NML}}$ maximizing $\mathbf{E}^{(\text{X})}(\hat{l}^*)$, when $\hat{\Sigma}$ in (3.2) is $k\hat{\Sigma}_{\text{NML}}$, is given by*

$$k = \frac{N+1}{N-p-2} = \frac{N}{N-p-3} + O(N^{-2}) \quad (N > p+3). \quad (2.17)$$

The proof is given in Ogasawara (2018, Subsection A.2). Theorem 2 gives

$$\hat{\Sigma}_{Nk}^{(P)} = \frac{N+1}{N-p-2} \hat{\Sigma}_{NML} = \frac{N+1}{N(N-p-2)} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})' \quad (2.18)$$

and its approximation up to order $O_p(N^{-1})$ is

$$\frac{N}{N-p-3} \hat{\Sigma}_{NML} = \frac{1}{N-p-3} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'. \quad (2.19)$$

Note that when $p = 1$, (2.19) gives Takezawa's (2012a) predictive estimator or "the third variance" mentioned earlier. The exact version of (2.18) when $p = 1$ can also be given by Takezawa (2012a, Equation (20)) though he did not use or define it. Again, we find that each element of (2.18) and (2.19) depends on p as well as N .

From Theorem 1 without $\mathbf{U}$ and Theorem 2, it is found that

$$\hat{\Sigma}_{Wk}^{(P)} = \frac{N}{N+1} \hat{\Sigma}_{Nk}^{(P)} < \hat{\Sigma}_{Nk}^{(P)}, \quad (2.20)$$

where the inequality sign is used in Löwner's sense. It is of interest to see that $\hat{\Sigma}_{WML} = \mathbf{S} > \hat{\Sigma}_{NML}$, where the direction of the inequality sign is reversed.

In Theorems 1 and 2, predictive estimators are derived using a single multiplicative constant k for all parameters. Except when $p = 1$ with $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$, this can be relaxed to distinct constants for the parameters. Ogasawara (2017, Equation (3.1)) defined a predictive estimator of a parameter vector:

$$\hat{\boldsymbol{\theta}}_{Ak}^{(P)} = \hat{\boldsymbol{\theta}}_{ML} - N^{-1} \text{diag}(\mathbf{k}) \hat{\boldsymbol{\alpha}}, \quad (2.21)$$

where $\hat{\boldsymbol{\theta}}_{ML}$ is the MLE of $\boldsymbol{\theta}_0$, $\mathbf{k}$ is a $q \times 1$ vector of coefficients, $\text{diag}(\mathbf{k})$ is the diagonal matrix whose diagonal elements are given by $\mathbf{k}$ and $\hat{\boldsymbol{\alpha}}$ is a $q \times 1$ stochastic vector typically given by $\hat{\boldsymbol{\theta}}_{ML}$ and a sample version $\hat{\boldsymbol{\beta}}_1$ of $\boldsymbol{\beta}_1$ with $N^{-1}\boldsymbol{\beta}_1$ being the asymptotic bias of $\hat{\boldsymbol{\theta}}_{ML}$.

When the elements of $\mathbf{k}$ are positive and $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$, (2.21) gives a shrinkage estimator of $\boldsymbol{\theta}_0$. In these cases (2.21) has properties common to ridge regression (Hoerl & Kennard, 1970) and Lasso (Tibshirani, 1996), where shrinkage is explicitly obtained. A general formula for the optimal $\mathbf{k}$ maximizing the asymptotic predictive expected log-likelihood up to order $O_p(N^{-2})$ denoted by, e.g., $E(\mathbf{X})(\hat{l}^*)_{\rightarrow O(N^{-2})}$, when $\mathbf{X}$ is a current data set, is given by Ogasawara (2017, Result 3) as

$$\mathbf{k} = \text{diag}^{-1}(\boldsymbol{\alpha}_0) \mathbf{I}_0^{-1} \left\{ NE \left(\frac{\partial \bar{l}}{\partial \boldsymbol{\theta}_0'} \otimes \mathbf{M} \right) \text{vec}(\mathbf{I}_0^{-1}) + \text{diag}^{-1}(\boldsymbol{\alpha}_0) \text{Diag} \left(\frac{\partial \boldsymbol{\alpha}'}{\partial \boldsymbol{\theta}_0} \right) \mathbf{1}_{(q)} \right\}, \quad (2.22)$$

where $\text{diag}^{-1}(\boldsymbol{\alpha}_0) = \{\text{diag}(\boldsymbol{\alpha}_0)\}^{-1}$; $\boldsymbol{\alpha}_0$ is the population counterpart of $\hat{\boldsymbol{\alpha}}$; $E(\cdot)$ is the expectation over current data, e.g., $E(\mathbf{X})(\cdot)$ in the case of the normal distribution, and is used for simplicity when confusion does not occur; $\partial \bar{l} / \partial \boldsymbol{\theta}_0 = \partial \bar{l} / \partial \boldsymbol{\theta} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$; $\otimes$ stands for the Kronecker product;

$$\mathbf{M} = \frac{\partial^2 \bar{l}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} - E \left(\frac{\partial^2 \bar{l}}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right) \equiv \frac{\partial^2 \bar{l}}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'} + \mathbf{I}_0; \quad (2.23)$$

$\text{vec}(\cdot)$ is the vectorizing operator stacking the columns of a matrix sequentially; $\text{Diag}(\cdot)$ is the diagonal matrix whose diagonal elements are those of the matrix in parentheses;

$\partial \boldsymbol{\alpha}' / \partial \boldsymbol{\theta}_0 = \partial \boldsymbol{\alpha}' / \partial \boldsymbol{\theta} |_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$; $\boldsymbol{\alpha}$ corresponding to $\boldsymbol{\alpha}_0$ is a vector variable; and $\mathbf{1}_{(q)}$ is the $q \times 1$ vector of 1's.

In the case of the p -variate normal distribution with $\boldsymbol{\theta} = (\boldsymbol{\mu}', \boldsymbol{\sigma}')'$, $q = p + \{p(p+1)/2\}$. Then, we have the following result:

Theorem 3. *Under the assumption of normality, the constant vector $\mathbf{k}$ in $\hat{\boldsymbol{\theta}}_{Ak}^{(P)}$ of (2.21) maximizing the asymptotic predictive expected log-likelihood, when $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\theta}}_{ML}$, is given by*

$$\mathbf{k} = (\mathbf{k}_{\boldsymbol{\mu}}', \mathbf{k}_{\boldsymbol{\sigma}}')', \quad (2.24)$$

where the $p \times 1$ vector $\mathbf{k}_{\boldsymbol{\mu}}$ is

$$\mathbf{k}_{\boldsymbol{\mu}} = \text{diag}^{-1}(\boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0 \text{diag}^{-1}(\boldsymbol{\mu}_0) \mathbf{1}_{(p)}, \quad (2.25)$$

and the $p^* \times 1$ vector $\mathbf{k}_{\boldsymbol{\sigma}}$ with $p^* = p(p+1)/2$ is

$$\mathbf{k}_{\boldsymbol{\sigma}} = -2(p+2)\mathbf{1}_{(p^*)} + 2\text{diag}^{-1}(\boldsymbol{\sigma}_0) \mathbf{D}_p^+ (\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0) \mathbf{D}_p^{+'} \text{diag}^{-1}(\boldsymbol{\sigma}_0) \mathbf{1}_{(p^*)}, \quad (2.26)$$

where $\mathbf{D}_p^+ = (\mathbf{D}_p' \mathbf{D}_p)^{-1} \mathbf{D}_p'$ is the Moore-Penrose g -inverse or the left inverse of $\mathbf{D}_p$, and $\mathbf{D}_p$ is the duplication matrix with $\text{vec}(\boldsymbol{\Sigma}_0) = \mathbf{D}_p \mathbf{v}(\boldsymbol{\Sigma}_0)$.

The proof is given in Ogasawara (2018, Subsection A.3). The vector $\mathbf{k}_{\boldsymbol{\sigma}}$ in (2.26) is rewritten as

$$\mathbf{k}_{\boldsymbol{\sigma}} = -2(p+2)\mathbf{1}_{(p^*)} + \text{diag}^{-1}(\boldsymbol{\sigma}_0) \mathbf{I}_{\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0'}^{-1} \text{diag}^{-1}(\boldsymbol{\sigma}_0) \mathbf{1}_{(p^*)}, \quad (2.27)$$

where $\mathbf{I}_{\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0'}$ is the $p^* \times p^*$ submatrix of $\mathbf{I}_0$ whose rows and columns correspond to $\boldsymbol{\sigma}_0$ and $\boldsymbol{\sigma}_0'$, respectively. The element of the second term on the right-hand side of (2.27) corresponding to $\sigma_{0ab} = (\boldsymbol{\Sigma}_0)_{ab}$, i.e., the (a, b) -th element of $\boldsymbol{\Sigma}_0$ ($p \geq a \geq b \geq 1$) is

$$\sigma_{0ab}^{-1} \sum_{c \geq d} (\sigma_{0ac} \sigma_{0bd} + \sigma_{0ad} \sigma_{0bc}) \sigma_{0cd}^{-1} \quad (p \geq a \geq b \geq 1). \quad (2.28)$$

where $\sum_{c \geq d} (\cdot) = \sum_{p \geq c \geq d \geq 1} (\cdot) = \sum_{c=d}^p \sum_{d=1}^c (\cdot)$. When $p = 1$, (2.28) becomes 2 though when $p > 1$, (2.28) does not generally become 2. When $p = 1$, (2.26) becomes $k_{\boldsymbol{\sigma}} = -6 + 2 = -4$, which reduces to the univariate case of Ogasawara (2017, Example 4.1). Note that when $p = 1$, the value $k_{\boldsymbol{\sigma}} = -4$ gives another asymptotic approximation $1 + N^{-1}4$ to $(N+1)/(N-p-2)$ in (2.17).

In $\mathbf{k}_{\boldsymbol{\mu}}$ of (2.25), $\text{diag}^{-1}(\boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0 \text{diag}^{-1}(\boldsymbol{\mu}_0) \equiv \mathbf{C}_{V\boldsymbol{\mu}_0}^{(2)}$ is seen as a matrix squared coefficient of variation of $\mathbf{x}_i$, which becomes the squared usual coefficient of variation $c_{V\boldsymbol{\mu}_0}^2$ when $p = 1$ for x_i ($i = 1, \dots, N$). Similarly, in $\mathbf{k}_{\boldsymbol{\sigma}}$ of (2.27), $\text{diag}^{-1}(\boldsymbol{\sigma}_0) \mathbf{I}_{\boldsymbol{\sigma}_0, \boldsymbol{\sigma}_0'}^{-1} \text{diag}^{-1}(\boldsymbol{\sigma}_0) \equiv \mathbf{C}_{V\boldsymbol{\sigma}_0}^{(2)}$ is seen as an asymptotic matrix squared coefficient of variation for $N^{1/2} \hat{\boldsymbol{\sigma}}_{NML}$. It is known that $k = 1/(1 + N^{-1}c_{V\boldsymbol{\mu}_0}^2)$ gives the least MSE of $k\bar{x}$ with $\bar{x}$ and $c_{V\boldsymbol{\mu}_0}$ being a sample mean and the coefficient of variation of x_i ($i = 1, \dots, N$), respectively under arbitrary distributions as long as the mean and variance of $\bar{x}$ exist (Gruber, 1998, Section 1.6; see also Ogasawara, 2015, Subsection 3.1). Note that $1/(1 + N^{-1}c_{V\boldsymbol{\mu}_0}^2) = 1 - N^{-1}c_{V\boldsymbol{\mu}_0}^2 + O(N^{-2})$.

The optimal vector $\mathbf{k}$ in Theorem 3 depends on unknown $\mathbf{C}_{V\boldsymbol{\mu}_0}^{(2)}$ and $\mathbf{C}_{V\boldsymbol{\sigma}_0}^{(2)}$ defined earlier as well as p , which is different from the results in Theorems 1 and 2, where the results were given only from known quantities. This problem with methods overcoming the difficulty will be addressed in the later discussion section.

2.3. Predictive estimation of structural parameters in a general covariance structure

In many cases of interest, a covariance matrix is structured with a $q \times 1$ structural parameter vector $\boldsymbol{\theta}$, which is generally written as $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}(\boldsymbol{\theta})$ with its population counterpart $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\theta}_0)$. A typical case of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is the factor analysis model, which will be numerically illustrated later. In this section, we obtain a similar result as in Theorem 3 using a $q \times 1$ coefficient vector $\mathbf{k}$ based on the Wishart distribution.

Theorem 4. *Let $\hat{\boldsymbol{\theta}}_{Ak}^{(P)} = \hat{\boldsymbol{\theta}}_{\text{WML}} - n^{-1} \text{diag}(\mathbf{k}) \hat{\boldsymbol{\theta}}_{\text{WML}}$, where $\hat{\boldsymbol{\theta}}_{\text{WML}}$ is a Wishart MLE of $\boldsymbol{\theta}_0$. Then, the optimal $\mathbf{k}$ maximizing the asymptotic predictive expected log-likelihood under normality is given by*

$$\mathbf{k} = \text{diag}^{-1}(\boldsymbol{\theta}_0) \left[\frac{1}{2} \sum_{a,b,c=1}^q (\mathbf{I}_0^{-1})_{\cdot b} \text{tr} \left\{ \boldsymbol{\Sigma}_0^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0a}} \boldsymbol{\Sigma}_0^{-1} \left(-2 \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0b}} \boldsymbol{\Sigma}_0^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0c}} + \frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_{0b} \partial \theta_{0c}} \right) \right\} (\mathbf{I}_0^{-1})_{ac} + \mathbf{I}_0^{-1} \text{diag}^{-1}(\boldsymbol{\theta}_0) \mathbf{1}_{(q)} \right], \quad (2.29)$$

where $(\cdot)_{\cdot b}$ is the b -th column of a matrix, $\theta_{0a} = (\boldsymbol{\theta}_0)_a$ the a -th element of $\boldsymbol{\theta}_0$, and $(\mathbf{I}_0)_{ij} = \frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}_0^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0i}} \boldsymbol{\Sigma}_0^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0j}} \right)$ ($i, j = 1, \dots, q$).

The proof is given in Ogasawara (2018, Subsection A.4). When an inverse structure $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}(\boldsymbol{\theta}) = \{\boldsymbol{\Sigma}(\boldsymbol{\theta})\}^{-1}$, i.e., a structure of the inverse of $\boldsymbol{\Sigma}$ is used (for the inverse structure see Ogasawara, 2016b), $\bar{l}$ becomes

$$\bar{l} = \frac{1}{2} \ln |\boldsymbol{\Gamma}| - \frac{1}{2} \text{tr}(\boldsymbol{\Gamma} \mathbf{S}) - \frac{1}{2} C(\mathbf{S}). \quad (2.30)$$

When $\boldsymbol{\Gamma}$ with its population counterpart $\boldsymbol{\Gamma}_0$ is linear in terms of $\boldsymbol{\theta}$, $\boldsymbol{\theta}$ is a vector of natural parameters in the exponential family of distributions and $\mathbf{M}$ vanishes. Then, in this case, (2.29) becomes

$$\mathbf{k} = \text{diag}^{-1}(\boldsymbol{\theta}_0) \mathbf{I}_0^{-1} \text{diag}^{-1}(\boldsymbol{\theta}_0) \mathbf{1}_{(q)}, \quad (2.31)$$

where $(\mathbf{I}_0)_{ij} = \frac{1}{2} \text{tr} \left(\boldsymbol{\Gamma}_0^{-1} \frac{\partial \boldsymbol{\Gamma}}{\partial \theta_{0i}} \boldsymbol{\Gamma}_0^{-1} \frac{\partial \boldsymbol{\Gamma}}{\partial \theta_{0j}} \right) = \frac{1}{2} \text{tr} \left(\boldsymbol{\Sigma}_0 \frac{\partial \boldsymbol{\Gamma}}{\partial \theta_{0i}} \boldsymbol{\Sigma}_0 \frac{\partial \boldsymbol{\Gamma}}{\partial \theta_{0j}} \right)$ ($i, j = 1, \dots, q$). In (4.3), $\text{diag}^{-1}(\boldsymbol{\theta}_0) \mathbf{I}_0^{-1} \text{diag}^{-1}(\boldsymbol{\theta}_0) \equiv \mathbf{C}_{\mathbf{V}\boldsymbol{\theta}_0}^{(2)}$ is seen as an asymptotic matrix squared coefficient of variation of $n^{1/2} \hat{\boldsymbol{\theta}}_{\text{WML}}$ as $\mathbf{C}_{\mathbf{V}\boldsymbol{\sigma}_0}^{(2)}$ (see the discussion in the second last paragraph of Subsection 2.2).

An application of Theorem 4 is shown for an unconstrained covariance matrix as follows.

Corollary 1. *For an unconstrained covariance matrix with $\boldsymbol{\theta} = \mathbf{v}(\boldsymbol{\Sigma})$, let $\mathbf{s}_{Ak}^{(P)} = \mathbf{s} - n^{-1} \text{diag}(\mathbf{k}) \mathbf{s}$ with $\mathbf{s} = \mathbf{v}(\mathbf{S}) = \mathbf{v}(\hat{\boldsymbol{\Sigma}}_{\text{WML}})$. Then, the optimal $\mathbf{k}$ maximizing the asymptotic predictive expected likelihood under normality based on the Wishart MLE, i.e., $\mathbf{s}$ is given by*

$$(\mathbf{k})_{(gh)} = -2(p+1) + \frac{1}{\sigma_{0gh}} \sum_{c \geq d} (\sigma_{0gc} \sigma_{0hd} + \sigma_{0gd} \sigma_{0hc}) \frac{1}{\sigma_{0cd}}, \quad (2.32)$$

where $(\cdot)_{(gh)}$ indicates the element of a vector corresponding to σ_{0gh} ($p \geq g \geq h \geq 1$) using a double subscript notation.

The proof is given in Ogasawara (2018, Subsection A.5). In (2.32), the second term on the right-hand side is written as $(\mathbf{C}_{\mathbf{V}\boldsymbol{\theta}_0}^{(2)})_{(gh)} \cdot \mathbf{1}_{(q)} = (\mathbf{C}_{\mathbf{V}\boldsymbol{\theta}_0}^{(2)} \mathbf{1}_{(q)})_{(gh)}$, where $(\mathbf{C}_{\mathbf{V}\boldsymbol{\theta}_0}^{(2)})_{(gh)} \cdot$ is the (gh) -th row of $\mathbf{C}_{\mathbf{V}\boldsymbol{\theta}_0}^{(2)}$. Note that the second term on the right-hand side of (2.32) for $\mathbf{k}$ is equal to that in (2.26) for $\mathbf{k}_{\boldsymbol{\sigma}}$. When $p = 1$, (2.32) becomes -2 , giving $\mathbf{s}_{Ak}^{(P)} = (1 + N^{-1}) \mathbf{s}$

with $s = (\mathbf{S})_{11}$, which can be seen as an approximation to $\frac{n}{n-2}s = \frac{1}{N-3} \sum_{i=1}^N (x_i - \bar{x})^2$ up to $O_p(N^{-1})$ with $x_i = (\mathbf{x}_i)_1$ and $\bar{x} = (\bar{\mathbf{x}})_1$ in Theorem 1 without covariates when $p = 1$.

3. Predictive estimation of an unconstrained covariance matrix under possible non-normality

In practice, data are more or less non-normally distributed. In this section, an asymptotic result corresponding to Theorem 1 for an unconstrained covariance matrix under possible non-normality is obtained using the Wishart likelihood. The result is given as follows.

Theorem 5. *The optimal multiplicative constant k in $\hat{\Sigma}_{\mathbf{W}k}^{(P)} = k\mathbf{S}$ maximizing the asymptotic predictive expected likelihood based on the Wishart distribution under possible non-normality is given by*

$$k = 1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\mathbf{\Omega}_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}, \quad (3.1)$$

where $\mathbf{\Omega}_{\text{NT}} = \text{ncov}(\mathbf{s})$ is the normal-theory (NT) exact covariance of $\mathbf{s}$, and $\mathbf{K}_{(4)}$ is the $p^* \times p^*$ matrix of multivariate fourth cumulants for $\mathbf{x}_i$, i.e., $(\mathbf{K}_{(4)})_{ab,cd} = \kappa_{abcd}(p \geq a \geq b \geq 1; p \geq c \geq d \geq 1)$ and κ_{abcd} is the multivariate fourth cumulant for $(\mathbf{x}_i)_a$, $(\mathbf{x}_i)_b$, $(\mathbf{x}_i)_c$ and $(\mathbf{x}_i)_d$ ($i = 1, \dots, N$; $a, b, c, d = 1, \dots, p$).

The proof is given in Ogasawara (2018, Subsection A.6). Note that under normality $\mathbf{K}_{(4)} = \mathbf{O}$ a zero matrix, which yields $k = 1 + n^{-1}(p + 1)$. This multiplicative constant is an asymptotic approximation to $n/(n - p - 1)$ in Theorem 1 up to order $O(n^{-1})$.

4. Evaluation of the predictive estimator by the Kullback-Leibler distance

The quantity $E_f^{(\mathbf{S})}(-2\hat{\tilde{l}}^*)$, which is equal to $E^{(\mathbf{S})}(-2\hat{\tilde{l}}^*)$ used earlier when the density $f(\mathbf{S}|\boldsymbol{\theta})$ holds, is minimized when $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$. Let non-stochastic $\bar{l}_0^*$ be $\hat{\tilde{l}}^*$, when $\hat{\boldsymbol{\theta}}$ is replaced by $\boldsymbol{\theta}_0$. Then, the difference $E_f^{(\mathbf{S})}(-2\hat{\tilde{l}}^* + 2\bar{l}_0^*)$ is

$$E_f^{(\mathbf{S})}(-2\hat{\tilde{l}}^* + 2\bar{l}_0^*) = 2 \int_{R(\mathbf{S})} f(\mathbf{S}|\boldsymbol{\theta}_0) \ln \frac{L^*(\boldsymbol{\theta}_0)}{L^*(\hat{\boldsymbol{\theta}})} d\mathbf{S}, \quad (4.1)$$

where $R(\mathbf{S})$ is the range or region of $\mathbf{S}$,

$$L^*(\boldsymbol{\theta}_0) = \exp(\bar{l}_0^*) \text{ and } L^*(\hat{\boldsymbol{\theta}}) = \exp(\hat{\tilde{l}}^*). \quad (4.2)$$

Since (4.1) is seen as two times the Kullback-Leibler distance for $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, (4.1) can be an index of the evaluation of $\hat{\boldsymbol{\theta}}$. In the case of the Wishart distribution under correct model specification, (4.1) becomes

$$\begin{aligned} E_f^{(\mathbf{S})}(-2\hat{\tilde{l}}^* + 2\bar{l}_0^*) &= E_f^{(\mathbf{S})}(-2\hat{\tilde{l}}^*) + 2\bar{l}_0^* \\ &= E_f^{(\mathbf{S})}(\ln |\hat{\Sigma}|) - \ln |\Sigma_0| + E_f^{(\mathbf{S})}\{\text{tr}(\hat{\Sigma}^{-1}\Sigma_0)\} - p. \end{aligned} \quad (4.3)$$

4.1. The case of $\hat{\Sigma} = k\mathbf{S}$

When $\hat{\Sigma} = k\mathbf{S}$ under normality, (4.3) becomes

$$\begin{aligned} E_f^{(\mathbf{S})}(-2\hat{\tilde{l}}^* + 2\bar{l}_0^*) &= p \ln k + E_f^{(\mathbf{S})}(\ln |\mathbf{S}|) - \ln |\Sigma_0| + k^{-1} E_f^{(\mathbf{S})}\{\text{tr}(\mathbf{S}^{-1}\Sigma_0)\} - p \\ &= p \ln k + \ln |\Sigma_0| + p \ln(2/n) + \psi_p(n/2) - \ln |\Sigma_0| + k^{-1} \frac{np}{n-p-1} - p \\ &= p \ln k - p \ln(n/2) + \psi_p(n/2) + k^{-1} \frac{np}{n-p-1} - p. \end{aligned} \quad (4.4)$$

When k is optimal under normality, i.e., $k = n/(n - p - 1)$, (4.4) is

$$\begin{aligned} E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\bar{l}_0^*) &= p \ln \frac{n}{n-p-1} - p \ln(n/2) + \psi_p(n/2) \\ &= p \ln \frac{2}{n-p-1} + \psi_p(n/2). \end{aligned} \quad (4.5)$$

For a general k under possible non-normality with the density of $\mathbf{S}$ being $g(\mathbf{S}|\cdot)$, using (A6.5) and (A6.8) in Ogasawara (2018), we have

$$\begin{aligned} E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\bar{l}_0^*) &= p \ln k + E_g^{(\mathbf{S})}(\ln |\mathbf{S}|) - \ln |\boldsymbol{\Sigma}_0| + k^{-1} E_g^{(\mathbf{S})}\{\text{tr}(\mathbf{S}^{-1} \boldsymbol{\Sigma}_0)\} - p \\ &= p \ln k - n^{-1}\{p^* + \text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\} + k^{-1}[p + n^{-1}2\{p^* + \text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\}] - p \\ &\quad + O(n^{-2}) \\ &= p \ln k + k^{-1}p - p + n^{-1}[(2k^{-1} - 1)\{p^* + \text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\}] + O(n^{-2}). \end{aligned} \quad (4.6)$$

It is of some interest to compare (4.4) and (4.6) under normality with $\mathbf{K}_{(4)} = \mathbf{O}$. Equating the two expressions, it is found that $-p \ln(n/2) + \psi_p(n/2)$ in (4.4) should be equal to $-n^{-1}p^* + O(n^{-2})$. When k is asymptotically optimal, i.e., $k = 1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\}$, it can be shown that

$$E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\bar{l}_0^*) = n^{-1}\{p^* + \text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\} + O(n^{-2}). \quad (4.7)$$

Note that the term of order $O(n^{-1})$ in (4.7) is unchanged even when $k = 1$. That is, the difference between the results using the asymptotically optimal k and $k = 1$ is of order $O(n^{-2})$. The actual expression of order $O(n^{-2})$ is too complicated to derive and is not given.

On the other hand, we can have $E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*)$ up to order $O(n^{-2})$, where $\hat{l}^*$ is given by $\hat{\boldsymbol{\Sigma}} = k\mathbf{S}$ with $k = 1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\}$ and $\hat{l}_{\text{WML}}^*$ is $\hat{l}^*$ when a Wishart MLE $\hat{\boldsymbol{\Sigma}}_{\text{WML}} = \mathbf{S}$ or $k = 1$ is used for $\hat{\boldsymbol{\Sigma}}$. That is, $E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*)$ is the negative amount of the reduction of $E_g^{(\mathbf{S})}(-2\hat{l}^*)$ when the asymptotically optimal predictive estimator is used over that when the usual $\mathbf{S}$ is used.

Theorem 6. *Under possible non-normality when $k = 1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\}$, we have*

$$E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*) = -\frac{n^{-2}}{2}p\{p + 1 + 2p^{-1}\text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\}^2 + O(n^{-3}). \quad (4.8)$$

Under normality, (4.8) becomes

$$E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*) = -\frac{n^{-2}}{2}p(p + 1)^2 + O(n^{-3}) \quad (4.9)$$

and in particular when the exactly optimal $k = n/(n - p - 1)$ is used, (4.8) is

$$E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*) = p \ln \frac{n}{n-p-1} - \frac{p(p+1)}{n-p-1}. \quad (4.10)$$

The proof is given in Ogasawara (2018, Subsection A.7). Note that the equality of (4.10) and (4.9) up to order $O(n^{-2})$ is seen by expanding (4.10) as

$$\begin{aligned} &p[n^{-1}(p + 1) + n^{-2}\{1 - (1/2)\}(p + 1)^2 + O(n^{-3})] - p(p + 1)\{n^{-1} + n^{-2}(p + 1) + O(n^{-3})\} \\ &= -\frac{n^{-2}}{2}p(p + 1)^2 + O(n^{-3}). \end{aligned}$$

4.2. General covariance structure $\Sigma = \Sigma(\theta)$

In the case of a general covariance structure $\Sigma = \Sigma(\theta)$ using $\hat{\theta}_{Ak}^{(P)} = \hat{\theta}_{WML} - n^{-1} \text{diag}(\mathbf{k}) \hat{\theta}_{WML}$ based on the Wishart distribution with the asymptotically optimal $\mathbf{k}$ (see (2.29)), as in (4.7) it can be shown that $E_f^{(S)}(-2\hat{l}^* + 2\bar{l}_0^*) = E_f^{(S)}(\ln |\hat{\Sigma}|) - \ln |\Sigma_0| + E_f^{(S)}\{\text{tr}(\hat{\Sigma}^{-1} \Sigma_0)\} - p$ (see (4.3)) with $\hat{\Sigma} = \Sigma(\hat{\theta}_{Ak}^{(P)})$, is equal to that when $\hat{\Sigma} = \hat{\Sigma}_{WML} = \Sigma(\hat{\theta}_{WML})$ up to order $O(n^{-1})$. The corresponding result up to order $O(n^{-2})$ becomes too complicated and is not derived. The result when $\hat{\Sigma} = \hat{\Sigma}_{WML}$ up to order $O(n^{-1})$ under possible non-normality is given in Ogasawara (2018, Subsection A.8).

On the other hand, $E_f^{(S)}(-2\hat{l}^* + 2\hat{l}_{WML}^*)$ under correct model specification and normality can be given by a special case of Ogasawara (2017, Result 3) as follows:

$$\begin{aligned} & E_f^{(S)}(-2\hat{l}^* + 2\hat{l}_{WML}^*) \\ &= -n^{-2} \left\{ nE_f \left(\frac{\partial \bar{l}}{\partial \theta_0} \otimes \mathbf{M} \right) \text{vec}(\mathbf{I}_0^{-1}) + \text{diag}^{-1}(\theta_0) \mathbf{1}_{(q)} \right\}' \\ & \quad \times \mathbf{I}_0^{-1} \left\{ nE_f \left(\frac{\partial \bar{l}}{\partial \theta_0} \otimes \mathbf{M} \right) \text{vec}(\mathbf{I}_0^{-1}) + \text{diag}^{-1}(\theta_0) \mathbf{1}_{(q)} \right\} + O(n^{-3}). \end{aligned} \quad (4.11)$$

When the inverse structure $\Gamma(\theta)$ is linear with respect to θ with $\mathbf{M} = \mathbf{O}$, (4.11) becomes

$$\begin{aligned} E_f^{(S)}(-2\hat{l}^* + 2\hat{l}_{WML}^*) &= -n^{-2} \mathbf{1}_{(q)}' \text{diag}^{-1}(\theta_0) \mathbf{I}_0^{-1} \text{diag}^{-1}(\theta_0) \mathbf{1}_{(q)} + O(n^{-3}) \\ &= -n^{-2} \mathbf{1}_{(q)}' \mathbf{C}_{V\theta_0}^{(2)} \mathbf{1}_{(q)} + O(n^{-3}) = -n^{-2} \mathbf{1}_{(q)}' \mathbf{k} + O(n^{-3}), \end{aligned} \quad (4.12)$$

where $\mathbf{I}_0$ is the information matrix for the inverse structure and $\mathbf{C}_{V\theta_0}^{(2)}$ was defined earlier (see the result after (2.31)).

While the results in (4.11) and (4.12) are given by the difference of $E_f^{(S)}(-2\hat{l}^*)$ and $E_f^{(S)}(-2\hat{l}_{WML}^*)$, $E_f^{(S)}(-2\hat{l}_{WML}^*)$ and an associated result are given as follows:

Theorem 7. *Under correct model specification and normality, $E_f^{(S)}(-2\hat{l}_{WML}^*)$ using $\hat{\theta}_{WML}$ and $E_f^{(S)}(-2\hat{l}^*)$ using $\hat{\theta}_{Ak}^{(P)}$ based on the Wishart distribution are*

$$E_f^{(S)}(-2\hat{l}_{WML}^*) = -2\bar{l}_0^* + n^{-1}q + O(n^{-2}), \quad (4.13)$$

which is known (see e.g. Ogasawara, 2017, Appendix), and

$$E_f^{(S)}(-2\hat{l}^*) = -2\bar{l}_0^* + n^{-1}q + O(n^{-2}). \quad (4.14)$$

The proof is given in Ogasawara (2018, Subsection A.9). Note that q of $n^{-1}q$ is a half of the correction term $2q$ of the AIC.

5. Numerical illustration with simulations

In this section, numerical illustrations of the predictive estimators for unconstrained covariance matrices and structural parameters in exploratory factor analysis are shown with simulations in the former cases under normality and non-normality.

Table 1 shows theoretical and simulated results for predictive estimators of a covariance matrix when $\Sigma_0 = \lambda_0 \lambda_0' + \Psi_0$, where $\lambda_0 = (.2, .3, .4, .5, .6, .7)'$ and $\Psi_0 = \text{diag}(.96, .91, .84, .75, .64, .51)$. In the simulations, non-normal data are generated by independently distributed common and unique factors following the uniform and chi-square distributions with 10 and 1 degrees of freedom, which are standardized with unit variances

to give Σ_0 . Note that in the uniform distribution, the excess kurtosis is -1.2 , while that of the chi-square with 10 degrees of freedom is 1.2 . The excess kurtosis of the chi-square with 1 degree of freedom is 12 , showing strong non-normality.

In the table, $A = 1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}$ is the asymptotically optimal value under arbitrary distributions (see (3.1)). The NT-exact A corresponding to $1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}$ is $n/(n - p - 1)$ (see (2.8)). The theoretical $B = E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\bar{l}_0^*)_{\rightarrow O(n^{-1})} = n^{-1}\{p^* + \text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}$ (see (4.7)) is given when $\hat{\Sigma} = \mathbf{AS}$ is used. The corresponding simulated B is the simulated mean of $-2\hat{l}^* + 2\bar{l}_0^*$ under normality and non-normality using 10^5 replications. The NT-exact B is $E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\bar{l}_0^*) = p \ln\{2/(n - p - 1)\} + \psi_p(n/2)$ (see (4.5)) when $\{n/(n - p - 1)\}\mathbf{S}$ is used. The corresponding simulated value is the sample mean of $-2\hat{l}^* + 2\bar{l}_0^*$ over 10^5 replications when $\{n/(n - p - 1)\}\mathbf{S}$ is used under normality and non-normality. Under normality, since the exact B is available, the corresponding simulated mean is unnecessary, but is included to show the validity of the simulations. The theoretical $C = E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*)_{\rightarrow O(n^{-2})} = -(n^{-2}/2)p\{p + 1 + 2p^{-1}\text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}^2$ (see (4.8)) is the value when $\hat{\Sigma} = \mathbf{AS}$ is used. The corresponding simulated value is the sample mean of $-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*$ when $\hat{\Sigma} = \mathbf{AS}$ for $-2\hat{l}^*$ and $\hat{\Sigma}_{\text{WML}} = \mathbf{S}$ for $2\hat{l}_{\text{WML}}^*$ are used under normality and non-normality. The NT-exact $C = E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*) = p \ln\{n/(n - p - 1)\} - \{p(p + 1)/(n - p - 1)\}$ (see (4.10)) when $\{n/(n - p - 1)\}\mathbf{S}$ is used. The corresponding simulated values are similarly given.

In the table, A depends on $\mathbf{K}_{(4)}$. In the case of the uniform distribution with a negative excess kurtosis, A is smaller than that under normality. However, it is found that A under the uniform distribution is still substantially larger than 1 especially when $N = 200$. On the other hand, the values of A in the chi-square distributions with positive excess kurtoses are larger than that under normality. It is found that A under normality is reasonably close to the NT-exact A .

The values of the theoretical and simulated B are similar under normality and non-normality. Their values are much larger than the absolute values of the theoretical and simulated C and NT-exact C . When the excess kurtosis becomes larger, the theoretical and simulated values of B become larger.

Table 2 gives results corresponding to those in Table 1, where a real correlation matrix of Harman's (1976, p.22; $N = 305$) eight physical variables is used. The fitted correlation matrix with two common factors is used as a population covariance matrix. Table 2 shows tendencies similar to those in Table 1.

Table 3 gives the values of the asymptotically optimal $1 - n^{-1}(\mathbf{k})_i$ ($i = 1, \dots, q$) in $\hat{\Theta}_{\text{Ak}}^{(\text{P})} = \hat{\Theta}_{\text{WML}} - n^{-1}\text{diag}(\mathbf{k})\hat{\Theta}_{\text{WML}}$ (see Theorem 4) in the case of Tables 1 and 2 under the assumption of normality, where $\hat{\Theta}_{\text{WML}}$ is the Wishart MLE of the parameters in factor analysis models. Note that $(\Lambda_0)_{12}$ is set to 0 to remove rotational indeterminacy, which is shown by 0* in the table. The corresponding coefficient denoted by * in the table is not defined since the fixed value is unchanged. It is found that $1 - n^{-1}(\mathbf{k})_i$ ($i = 1, \dots, p$) for $\hat{\Psi}_{\text{WML}}$ are relatively similar. In contrast, the values of $1 - n^{-1}(\mathbf{k})_i$ ($i = p + 1, \dots, q$) for $\hat{\Lambda}_{\text{WML}}$ or $\hat{\Lambda}_{\text{WML}}$ have a relatively large variance. The extreme case of $1 - n^{-1}(\mathbf{k})_i$ for $(\Lambda_0)_{42} = -.031$ is $-.748$, which happened due to the finite sample size. In this case $1 - n^{-1}(\mathbf{k})_i$ should be 0, which is close to the population value $-.031$.

Table 1: Theoretical and simulated values for predictive estimators of a covariance matrix following the one-factor model (for Λ_0 and Ψ_0 see Table 3)

Data		$N = n + 1 = 200$		$N = n + 1 = 400$	
		Th.	Sim.	Th.	Sim.
Normal	A	1.0352		1.0175	
	NT-exact A	1.0365		1.0179	
	B	.10553	.10818	.05263	.05324
	NT-exact B	.10819	.10817	.05328	.05324
	C	-.00371	-.00389	-.00092	-.00094
	NT-exact C	-.00389	-.00389	-.00095	-.00094
Uniform	A	1.0299		1.0149	
	B	.08974	.09250	.04476	.04551
	NT-exact B	*	.09257	*	.04553
	C	-.00268	-.00284	-.00067	-.00069
	NT-exact C	*	-.00277	*	-.00068
Chi-square (10 df)	A	1.0404		1.0202	
	B	.12131	.12336	.06051	.06094
	NT-exact B	*	.12343	*	.06096
	C	-.00491	-.00503	-.00122	-.00123
	NT-exact C	*	-.00497	*	-.00121
Chi-square (1 df)	A	1.0878		1.0438	
	B	.26340	.25632	.13137	.12928
	NT-exact B	*	.26382	*	.13120
	C	-.02313	-.02252	-.00575	-.00560
	NT-exact C	*	-.01502	*	-.00368

Note. Th.= Theoretical values, Sim.= Simulated values, $A = 1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}$, NT-exact $A = n/(n - p - 1)$, Theoretical $B = E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_0^*)_{\rightarrow O(n-1)} = n^{-1}\{p^* + \text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}$ using $\hat{\Sigma} = \mathbf{AS}$, NT-exact $B = E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_0^*) = p \ln\{2/(n - p - 1)\} + \psi_p(n/2)$ using $\hat{\Sigma} = \{n/(n - p - 1)\}\mathbf{S}$, Theoretical $C = E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*)_{\rightarrow O(n-2)} = -(n^{-2}/2)p\{p + 1 + 2p^{-1}\text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}^2$ using $\hat{\Sigma} = \mathbf{AS}$, NT-exact $C = E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*) = p \ln\{n/(n - p - 1)\} - \{p(p + 1)/(n - p - 1)\}$ using $\hat{\Sigma} = \{n/(n - p - 1)\}\mathbf{S}$. An asterisk indicates the corresponding NT value.

Table 2: Theoretical and simulated values for predictive estimators of a covariance matrix following the two-factor model based on Harman's (1976, p.22, $N = 305$) eight physical variables (for Λ_0 and Ψ_0 see Table 3)

Data		Th.	Sim.	Th.	Sim.
		Normal		Uniform	
Normal	A	1.0296		1.0263	
	NT-exact A	1.0305			
	B	.11842	.12085	.10516	.10763
	NT-exact B	.12089	.12085	*	.10767
	C	-.00351	-.00364	-.00276	-.00290
	NT-exact C	-.00365	-.00364	*	-.00286
Chi-square (10 df)	A	1.0329		1.0628	
	B	.13169	.13393	.25106	.24751
	NT-exact B	*	.13397	*	.25154
	C	-.00434	-.00446	-.01576	-.01552
	NT-exact C	*	-.00443	*	-.01148

Note. Th.= Theoretical values, Sim.= Simulated values, $A = 1 + n^{-1}\{p + 1 + 2p^{-1}\text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}$, NT-exact $A = n/(n - p - 1)$, Theoretical $B = E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_0^*)_{\rightarrow O(n-1)} = n^{-1}\{p^* + \text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}$ using $\hat{\Sigma} = \mathbf{AS}$, NT-exact $B = E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_0^*) = p \ln\{2/(n - p - 1)\} + \psi_p(n/2)$ using $\hat{\Sigma} = \{n/(n - p - 1)\}\mathbf{S}$, Theoretical $C = E_g^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*)_{\rightarrow O(n-2)} = -(n^{-2}/2)p\{p + 1 + 2p^{-1}\text{tr}(\Omega_{\text{NT}}^{-1}\mathbf{K}_{(4)})\}^2$ using $\hat{\Sigma} = \mathbf{AS}$, NT-exact $C = E_f^{(\mathbf{S})}(-2\hat{l}^* + 2\hat{l}_{\text{WML}}^*) = p \ln\{n/(n - p - 1)\} - \{p(p + 1)/(n - p - 1)\}$ using $\hat{\Sigma} = \{n/(n - p - 1)\}\mathbf{S}$. An asterisk indicates the corresponding NT value.

Table 3: Population values and coefficients $1 - n^{-1}(\mathbf{k})_i (i = 1, \dots, q)$ under normality for the predictive estimators based on the Wishart MLEs in factor analysis

Artificial one-factor model				Two-factor model based on Harman's		
$N=200$		$N=400$		(1976, p.22, $N=305$) eight physical variables		
Ψ_0	λ_0	Ψ_0	λ_0	Ψ_0	Λ_0	
Values of Ψ_0 and $\lambda_0(\Lambda_0)$						
.96	.2	(the same as when $N=200$)		.170	.911	0*
.91	.3		.107	.937	-.120	
.84	.4		.166	.906	-.112	
.75	.5		.199	.894	-.031	
.64	.6		.089	.513	.805	
.51	.7		.364	.428	.673	
				.416	.364	.672
				.537	.459	.502
Coefficients $1 - n^{-1}(\mathbf{k})_i (i = 1, \dots, q)$						
1.025	.821	1.013	.911	1.013	1.003	*
1.026	.923	1.013	.961	1.027	1.010	.699
1.028	.958	1.014	.979	1.026	1.009	.666
1.030	.975	1.015	.987	1.028	1.003	-.748
1.031	.986	1.016	.993	.960	.917	1.033
1.024	.998	1.012	.999	1.032	.910	1.022
				1.030	.891	1.018
				1.025	.934	1.027

Note. The zero with an asterisk 0* is a fixed value to remove rotational indeterminacy.

The corresponding coefficient denoted by * is not given since the fixed value is unchanged.

6. Predictive estimation using mean square errors

6.1. Predictive estimation of variances and covariances

Predictive estimators can also be derived using associated mean square errors. In this section, predictive estimation of unconstrained variances and covariances is considered. Let $\mathbf{T}$ be an independent copy of $\mathbf{S}$ as before. Let

$$F = \sum_{i \geq j} w_{ij} (k_{ij} s_{ij} - t_{ij})^2, \quad (6.1)$$

where w_{ij} is a given weight, e.g., $w_{ij} = 1$ and $w_{ij} = 2 - \delta_{ij}$, where δ_{ij} is the Kronecker delta; k_{ij} is a constant to be derived; and $s_{ij} = (\mathbf{S})_{ij}$ and $t_{ij} = (\mathbf{T})_{ij} (p \geq i \geq j \geq 1)$. Then, the expectation of F over the distribution of $\mathbf{T}$ under possible non-normality is

$$\begin{aligned} E_g^{(\mathbf{T})}(F) &= E_g^{(\mathbf{T})} \left[\sum_{i \geq j} w_{ij} \{k_{ij} s_{ij} - \sigma_{0ij} - (t_{ij} - \sigma_{0ij})\}^2 \right] \\ &= \sum_{i \geq j} w_{ij} (k_{ij} s_{ij} - \sigma_{0ij})^2 + E_g^{(\mathbf{T})} \sum_{i \geq j} w_{ij} (t_{ij} - \sigma_{0ij})^2. \end{aligned} \quad (6.2)$$

We minimize $E_g^{(\mathbf{S})}\{E_g^{(\mathbf{T})}(F)\}$ with respect to k_{ij} . The results are given as follows:

Theorem 8. *The optimal constant k_{ij} minimizing $E_g^{(\mathbf{S})}\{E_g^{(\mathbf{T})}(F)\}$ under possible non-normality is*

$$k_{ij} = \frac{1}{1 + c_V^2(s_{ij})}, \quad (6.3)$$

where $c_V^2(s_{ij})$ is the squared coefficient of variation of s_{ij} , i.e.,

$$c_V^2(s_{ij}) = \text{var}(s_{ij})/\sigma_{0ij}^2 \quad (p \geq i \geq j \geq 1) \quad (6.4)$$

The proof is given in Ogasawara (2018, Subsection A.10). Since $\text{var}(s_{ij}) = \frac{\kappa_{iiij}}{N} + \frac{\sigma_{0ii}\sigma_{0jj} + \sigma_{0ij}^2}{N-1}$ (Kaplan, 1952, Equation (3)), k_{ij} in (6.3) is written as

$$k_{ij} = \frac{1}{1 + \frac{\kappa_{iiij}/\sigma_{0ij}^2}{N} + \frac{\rho_{0ij}^{-2} + 1}{N-1}}, \quad (6.5)$$

where $\kappa_{iiij}/\sigma_{0ij}^2$ is seen as a multivariate excess kurtosis of $(\mathbf{x}_m)_i$, $(\mathbf{x}_m)_i$, $(\mathbf{x}_m)_j$ and $(\mathbf{x}_m)_j$ ($m = 1, \dots, N$); $\rho_{0ij} = \sigma_{0ij}/(\sigma_{0ii}\sigma_{0jj})^{1/2}$ is the population correlation coefficient of $(\mathbf{x}_m)_i$ and $(\mathbf{x}_m)_j$ ($p \geq i \geq j \geq 1$; $m = 1, \dots, N$).

Under normality

$$k_{ij} = \frac{1}{1 + n^{-1}(\rho_{0ij}^{-2} + 1)} \quad (p \geq i \geq j \geq 1), \quad (6.6)$$

with $n = N - 1$, which depends on unknown ρ_{0ij} when $i \neq j$.

When $i = j$, (6.5) becomes

$$k_{ii} = \frac{1}{1 + \frac{\kappa_{iiii}/\sigma_{0ii}^2}{N} + \frac{2}{N-1}} \quad (i = 1, \dots, p), \quad (6.7)$$

which is known (see e.g., Ogasawara, 2015, Subsection 3.2), and depends on the unknown excess kurtosis of $(\mathbf{x}_m)_i$ ($m = 1, \dots, N$). Under normality, (6.7) becomes

$$k_{ii} = \frac{1}{1 + \frac{2}{N-1}} = \frac{N-1}{N+1} \quad (i = 1, \dots, p), \quad (6.8)$$

which does not depend on unknown values and gives the known estimator of σ_{0ii} ,

$$k_{ii}s_{ii} = \frac{1}{N+1} \sum_{m=1}^N (x_m - \bar{x})^2 \quad (i = 1, \dots, p) \quad (6.9)$$

(DeGroot & Schervish, 2002, p.431).

When $k_{ij} = k$ ($p \geq i \geq j \geq 1$) is used, we minimize

$$\mathbb{E}_g^{(\mathbf{S})} \left\{ \sum_{i \geq j} w_{ij} (k s_{ij} - \sigma_{0ij})^2 \right\} = \sum_{i \geq j} w_{ij} \text{var}(s_{ij}) k^2 + \sum_{i \geq j} w_{ij} (k-1)^2 \sigma_{0ij}^2 \quad (6.10)$$

with respect to k . The result is easily given as follows.

Corollary 2. *The optimal $k = k_{ij}$ ($p \geq i \geq j \geq 1$) minimizing $\mathbb{E}_g^{(\mathbf{S})}\{\mathbb{E}_g^{(\mathbf{T})}(F)\}$ under possible non-normality is given by*

$$k = \frac{\sum_{i \geq j} w_{ij} \sigma_{0ij}^2}{\sum_{i \geq j} w_{ij} \text{var}(s_{ij}) + \sum_{i \geq j} w_{ij} \sigma_{0ij}^2} = \frac{1}{1 + \left\{ \sum_{i \geq j} w_{ij} \text{var}(s_{ij}) / \sum_{i \geq j} w_{ij} \sigma_{0ij}^2 \right\}}. \quad (6.11)$$

In (6.11), $\sum_{i \geq j} w_{ij} \text{var}(s_{ij}) / \sum_{i \geq j} w_{ij} \sigma_{0ij}^2$ is a weighted mean squared coefficient of variation for s_{ij} ($p \geq i \geq j \geq 1$).

Recall that the result in Theorem 8 does not depend on w_{ij} while that in Corollary 2 depends on w_{ij} ($p \geq i \geq j \geq 1$). When $w_{ij} = w$ ($p \geq i \geq j \geq 1$) under possible non-normality, we have

$$k = \frac{1}{1 + \frac{\sum_{i \geq j} \text{var}(s_{ij})}{\sum_{i \geq j} \sigma_{0ij}^2}} = \frac{1}{1 + \frac{\sum_{i \geq j} (N^{-1} \kappa_{iijj} + n^{-1} \sigma_{0ii} \sigma_{0jj})}{\sum_{i \geq j} \sigma_{0ij}^2} + n^{-1}}, \quad (6.12)$$

while under normality

$$k = 1 / \left[1 + n^{-1} \left\{ \left(\sum_{i \geq j} \sigma_{0ii} \sigma_{0jj} / \sum_{i \geq j} \sigma_{0ij}^2 \right) + 1 \right\} \right]. \quad (6.13)$$

Both of (6.12) and (6.13) depend on unknown population values.

As mentioned earlier, in (6.1) when $w_{ij} = 2 - \delta_{ij}$ ($p \geq i \geq j \geq 1$),

$$F = \sum_{i \geq j} w_{ij} (k_{ij} s_{ij} - t_{ij})^2 = \sum_{i,j=1}^p (k_{ij} s_{ij} - t_{ij})^2 = \text{tr}\{(\mathbf{K} \odot \mathbf{S} - \mathbf{T})^2\}, \quad (6.14)$$

where $\odot$ indicates the Hadamard product and $(\mathbf{K})_{ij} = k_{ij}$ ($i, j = 1, \dots, p$). In (6.14), the weights for the non-duplicated off-diagonal elements of $\mathbf{S}$ are two times those for the diagonal elements. On the other hand, when $w_{ij} = 1$ ($p \geq i \geq j \geq 1$),

$$F = \sum_{i \geq j} w_{ij} (k_{ij} s_{ij} - t_{ij})^2 = \mathbf{v}'(\mathbf{K} \odot \mathbf{S} - \mathbf{T})\mathbf{v}(\mathbf{K} \odot \mathbf{S} - \mathbf{T}), \quad (6.15)$$

where the weights for the non-duplicated elements of $\mathbf{S}$ are equal.

When $w_{11} = 1$ and the other w_{ij} 's are 0, this case under non-normality and normality reduces to (6.7) and (6.8), respectively. When $w_{21} = 1$ and the other w_{ij} 's are 0, this case under non-normality and normality reduces to (6.5) and (6.6) when $i \neq j$, respectively.

6.2. Predictive estimation of an unconstrained mean vector based on the multivariate normal distribution

In Theorem 2, the optimal k for predictive estimation of a covariance matrix when $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ under normality was given, while in Theorem 3, simultaneous estimation of $\boldsymbol{\mu}_0$ and $\boldsymbol{\Sigma}_0$ under normality was shown. Since a sample mean vector is closely associated with these estimators, a predictive estimator of $\boldsymbol{\mu}_0$, when $\hat{\boldsymbol{\Sigma}}$ is given by e.g., $\hat{\boldsymbol{\Sigma}}_{\text{NML}}$ and $\mathbf{S}$ under normality, is shown in this section. This is a reversed problem of Theorem 2.

Let $\mathbf{Z} = (\mathbf{z}_1, \dots, \mathbf{z}_N)'$ be an independent copy of $\mathbf{X}$ as in (2.15). Define

$$\hat{\boldsymbol{\Sigma}} = c_N^{-1} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})', \quad (6.16)$$

where c_N is a fixed quantity depending on N . For instance, when $\hat{\boldsymbol{\Sigma}}$ is $\hat{\boldsymbol{\Sigma}}_{\text{NML}}$, $\mathbf{S}$, $\hat{\boldsymbol{\Sigma}}_{\text{Wk}}^{(\text{P})}$ (see Theorem 1) and $\hat{\boldsymbol{\Sigma}}_{\text{Nk}}^{(\text{P})}$ (see Theorem 2), c_N is N , $n(= N - 1)$, $N - p - 2$ and $N(N - p - 2)/(N + 1)$, respectively. Then, the conditional density of $\mathbf{X}$ is

$$f(\mathbf{X} | \boldsymbol{\mu}, \hat{\boldsymbol{\Sigma}}) = \frac{1}{(2\pi)^{Np/2} |\hat{\boldsymbol{\Sigma}}|^{N/2}} \exp \left\{ - \sum_{i=1}^N (\mathbf{x}_i - \boldsymbol{\mu})' \hat{\boldsymbol{\Sigma}}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) / 2 \right\} \quad (6.17)$$

(compare with (2.14)). As in (2.15), it follows that

$$\begin{aligned} -2\hat{l}^* &= p \ln(2\pi) + \ln |\hat{\Sigma}| + \text{tr}[\hat{\Sigma}^{-1} \text{E}(\mathbf{Z}) \{N^{-1} \sum_{i=1}^N (\mathbf{z}_i - \hat{\boldsymbol{\mu}})(\mathbf{z}_i - \hat{\boldsymbol{\mu}})'\}] \\ &= p \ln(2\pi) + \ln |\hat{\Sigma}| + \text{tr}[\hat{\Sigma}^{-1} \{\Sigma_0 + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)'\}], \end{aligned} \quad (6.18)$$

where $\hat{\boldsymbol{\mu}}$ is an estimator of $\boldsymbol{\mu}_0$. Using a $p \times 1$ fixed vector $\mathbf{k}$, let

$$\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(\text{P})} = \mathbf{k} \odot \bar{\mathbf{x}} = \text{diag}(\bar{\mathbf{x}})\mathbf{k} = \text{diag}(\mathbf{k})\bar{\mathbf{x}}. \quad (6.19)$$

Then, we have the following result.

Theorem 9. *The optimal $\mathbf{k}$ in $\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(\text{P})} = \mathbf{k} \odot \bar{\mathbf{x}}$ maximizing $\text{E}^{(\mathbf{X})}(\hat{l}^*)$ under normality, when $\hat{\boldsymbol{\mu}}$ in (6.18) is $\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(\text{P})}$ and $\hat{\Sigma} = c_N^{-1} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'$ with c_N being a fixed quantity depending on N , is given by*

$$\mathbf{k} = \{N^{-1} \Sigma_0 \odot \Sigma_0^{-1} + \text{diag}(\boldsymbol{\mu}_0) \Sigma_0^{-1} \text{diag}(\boldsymbol{\mu}_0)\}^{-1} \text{diag}(\boldsymbol{\mu}_0) \Sigma_0^{-1} \boldsymbol{\mu}_0. \quad (6.20)$$

The proof is given in Ogasawara (2018, Subsection A.11). When $p = 1$, (6.20) becomes

$$k = \frac{\mu_{01}^2 / \sigma_{011}}{N^{-1} + (\mu_{01}^2 / \sigma_{011})} = \frac{1}{1 + c_V^2(\bar{x})} = \frac{1}{1 + N^{-1} c_{V\mu_0}^2}, \quad (6.21)$$

where $c_{V\mu_0}^2 = \sigma_{011} / \mu_{01}^2$ (see Subsection 2.2). The k in (6.21) is the same as that minimizing the MSE (see Subsection 2.2 and (6.3)).

When

$$F = \sum_{i=1}^p w_i \{k_i(\bar{\mathbf{x}})_i - (\bar{\mathbf{z}})_i\}^2, \quad (6.22)$$

where $(\bar{\mathbf{x}})_i$ and $(\bar{\mathbf{z}})_i (i = 1, \dots, p)$ are sample means of the i -th observable variables in $\mathbf{X}$ and $\mathbf{Z}$, respectively, it can be shown that $k_i (i = 1, \dots, p)$ minimizing F is equal to (6.21).

Let the predictive generalized MSE (PGMSE) be

$$\text{PGMSE} = \text{E}^{(\mathbf{X})}[\text{E}^{(\mathbf{Z})}\{(\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(\text{P})} - \bar{\mathbf{z}})' \Sigma_0^{-1} (\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(\text{P})} - \bar{\mathbf{z}})\}] \quad (6.23)$$

with $\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(\text{P})} = \mathbf{k} \odot \bar{\mathbf{x}}$ and $\bar{\mathbf{z}}$ being an independent copy of $\bar{\mathbf{x}}$. Then, we have

Corollary 3. *The optimal $\mathbf{k}$ in $\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(\text{P})} = \mathbf{k} \odot \bar{\mathbf{x}}$ minimizing PGMSE of (6.23) under normality is given by (6.20) of Theorem 9.*

The proof is given in Ogasawara (2018, Subsection A.12). The result of (6.23) is expected from the form of (6.17). Note, however, that the result of Corollary 3 is distribution-free in that the result holds under non-normality as long as the expectation in (6.23) exists and that no specific form of $\hat{\Sigma}$ is used. We find that $\text{GMSE} \equiv \text{E}^{(\mathbf{X})}\{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)' \Sigma_0^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}_0)\}$ for an estimator $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}_0$ is a special case of Ogasawara (2017, Definition 1).

Table 4: Summary of the predictive estimators

Theorems and Corollaries	Assumption for data	Distribution or MSE	Predictive estimators and optimal coefficients
Theorem 1	Normal	Wishart	$\hat{\mathbf{\Sigma}}_{Wk}^{(P)} = k \mathbf{S} \quad k = n/(n - p - 1)$
Theorem 2	Normal	Multivariate normal with $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$	$\hat{\mathbf{\Sigma}}_{Nk}^{(P)} = k \hat{\mathbf{\Sigma}}_{\text{NML}} = k(n/N) \mathbf{S} \quad k = (N + 1)/(N - p - 2)$
Theorem 3	Normal	Multivariate normal	$\hat{\boldsymbol{\theta}}_{Ak}^{(P)} = (\hat{\boldsymbol{\mu}}_{Ak}^{(P)})', \hat{\boldsymbol{\sigma}}_{Ak}^{(P)'} = \hat{\boldsymbol{\theta}}_{\text{ML}} - N^{-1} \text{diag}\{(\mathbf{k}_{\boldsymbol{\mu}}', \mathbf{k}_{\boldsymbol{\sigma}}')'\} \hat{\boldsymbol{\theta}}_{\text{ML}},$ $\mathbf{k}_{\boldsymbol{\mu}} = \text{diag}^{-1}(\boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0 \text{diag}^{-1}(\boldsymbol{\mu}_0) \mathbf{1}_{(p)},$ $\mathbf{k}_{\boldsymbol{\sigma}} = -2(p + 2) \mathbf{1}_{(p^*)}$ $+ 2 \text{diag}^{-1}(\boldsymbol{\sigma}_0) \mathbf{D}_p^+ (\boldsymbol{\Sigma}_0 \otimes \boldsymbol{\Sigma}_0) \mathbf{D}_p^{+'} \text{diag}^{-1}(\boldsymbol{\sigma}_0) \mathbf{1}_{(p^*)}$
Theorem 4	Normal	Wishart	$\hat{\mathbf{\Sigma}} = \boldsymbol{\Sigma}(\hat{\boldsymbol{\theta}}_{Ak}^{(P)}), \quad \hat{\boldsymbol{\theta}}_{Ak}^{(P)} = \hat{\boldsymbol{\theta}}_{\text{WML}} - n^{-1} \text{diag}(\mathbf{k}) \hat{\boldsymbol{\theta}}_{\text{WML}},$ $\mathbf{k} = \text{diag}^{-1}(\boldsymbol{\theta}_0) \left[\frac{1}{2} \sum_{a,b,c=1}^q (\mathbf{I}_0^{-1})_{.b} \text{tr} \left\{ \boldsymbol{\Sigma}_0^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0a}} \boldsymbol{\Sigma}_0^{-1} \right. \right.$ $\left. \times \left(-2 \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0b}} \boldsymbol{\Sigma}_0^{-1} \frac{\partial \boldsymbol{\Sigma}}{\partial \theta_{0c}} + \frac{\partial^2 \boldsymbol{\Sigma}}{\partial \theta_{0b} \partial \theta_{0c}} \right) \right] (\mathbf{I}_0^{-1})_{ac} + \mathbf{I}_0^{-1} \text{diag}^{-1}(\boldsymbol{\theta}_0) \mathbf{1}_{(q)} \Big]$
Corollary 1	Normal	Wishart	$\mathbf{s}_{Ak}^{(P)} = \mathbf{s} - n^{-1} \text{diag}(\mathbf{k}) \mathbf{s},$ $(\mathbf{k})_{(gh)} = -2(p + 1) + \frac{1}{\sigma_{0gh}} \sum_{c \geq d} (\sigma_{0gc} \sigma_{0hd} + \sigma_{0gd} \sigma_{0hc}) \frac{1}{\sigma_{0cd}}$
Theorem 5	Non-normal	Wishart	$\hat{\mathbf{\Sigma}}_{Wk}^{(P)} = k \mathbf{S}, \quad k = 1 + n^{-1} \{p + 1 + 2p^{-1} \text{tr}(\boldsymbol{\Omega}_{\text{NT}}^{-1} \mathbf{K}_{(4)})\}$
Theorem 8	Non-normal	MSE	$k_{ij} s_{ij}, \quad k_{ij} = 1/\{1 + c_V^2(s_{ij})\}$
Equation (6.6)	Normal	MSE	$k_{ij} s_{ij}, \quad k_{ij} = 1/\{1 + n^{-1}(\rho_{0ij}^{-2} + 1)\}$
Corollary 2	Non-normal	MSE	$k s_{ij}, \quad k = 1/\left[1 + \left\{ \sum_{i \geq j} w_{ij} \text{var}(s_{ij}) / \sum_{i \geq j} w_{ij} \sigma_{0ij}^2 \right\}\right]$
Theorem 9	Normal	Multivariate normal	$\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(P)} = \mathbf{k} \odot \bar{\mathbf{x}}$ with $\hat{\mathbf{\Sigma}}$ given $\mathbf{k} = \{N^{-1} \boldsymbol{\Sigma}_0 \odot \boldsymbol{\Sigma}_0^{-1} + \text{diag}(\boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0^{-1} \text{diag}(\boldsymbol{\mu}_0)\}^{-1}$ $\times \text{diag}(\boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0$
Corollary 3	Non-normal	PGMSE	$\hat{\boldsymbol{\mu}}_{\mathbf{k}}^{(P)} = \mathbf{k} \odot \bar{\mathbf{x}}$ with $\hat{\mathbf{\Sigma}}$ given $\mathbf{k} = \{N^{-1} \boldsymbol{\Sigma}_0 \odot \boldsymbol{\Sigma}_0^{-1} + \text{diag}(\boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0^{-1} \text{diag}(\boldsymbol{\mu}_0)\}^{-1}$ $\times \text{diag}(\boldsymbol{\mu}_0) \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0$

7. Discussion

The predictive estimators derived earlier are summarized in Table 4. The optimal values of k and $\mathbf{k}$ derived earlier depend on unknown population values except those of Theorems 1, 2 and k_{ii} of (6.8). This problem can be solved in several ways as in Ogasawara (2016c, Section 7). Firstly, the optimal coefficients k and $\mathbf{k}$ can be estimated using current data, which gives some information about the population optimal values. However, the estimators cannot be substituted for k and $\mathbf{k}$ since the asymptotic optimality is generally lost by this replacement. When an additional independent data set of size $n^* = O(n)$ is available as in cross-validation, this can be used for estimation of k and $\mathbf{k}$ without losing the asymptotic optimality.

Secondly, in many cases we have prior information about associated statistics, where k and $\mathbf{k}$ are assumed to be monotonic functions of the statistics. For instance, consider $k_{ij} = 1/\{1 + c_V^2(s_{ij})\}$ in (6.3). If a lower bound of c_V^2 is available, the upper bound of k_{ij} is obtained, which can be used as a conservative version of the optimal k_{ij} .

Thirdly, we can obtain information about the optimal k and $\mathbf{k}$ by simulations, where typical population $\boldsymbol{\theta}_0$'s can be used. Based on this information, we can have lower or upper bounds for k and $\mathbf{k}$.

Theorems 1 and 2 give the predictive estimators $\{n/(n - p - 1)\} \mathbf{S}$ and $\{(N + 1)/(N - p - 2)\} \hat{\mathbf{\Sigma}}_{\text{NML}}$, respectively, which suggests the use of these predictive estimators in place of $\mathbf{S}$ and $\hat{\mathbf{\Sigma}}_{\text{NML}}$ when a parameter vector $\boldsymbol{\theta}$ in a covariance structure $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ is estimated. This

is intriguing since the usual method of estimation can be used only by replacing, e.g., $\mathbf{S}$ with $\{n/(n-p-1)\}\mathbf{S}$. However, this is not optimal since this method uniformly increases $\hat{\boldsymbol{\theta}}_{\text{WML}}$ when, e.g., $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}$. That is, by this replacement $\{n/(n-p-1)\}^{1/2}\hat{\boldsymbol{\Lambda}}_{\text{WML}}$ and $\{n/(n-p-1)\}\hat{\boldsymbol{\Psi}}_{\text{WML}}$ are obtained. Note that this result is contradictory to the asymptotic result of Theorem 4 and its illustration in Table 3.

REFERENCES

- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csáki (Eds.), *Proceedings of the 2nd international symposium on information theory* (pp. 267–281). Budapest: Akadémiai Kiado.
- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108–132.
- DeGroot, M. H., & Schervish, M. J. (2002). *Probability and statistics* (3rd ed.). Boston, MA: Addison-Wesley.
- Dey, D. K., & Srinivasan, C. (1985). Estimation of a covariance matrix under Stein’s loss. *The Annals of Statistics*, **13**, 1581–1591.
- Gruber, M. H. J. (1998). *Improving efficiency by shrinkage: The James-Stein and ridge regression estimators*. New York: Marcel Dekker.
- Harman, H. H. (1976). *Modern factor analysis* (3rd ed.). Chicago: University of Chicago Press.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.
- Ibrahim, J. G., & Chen, M.-H. (2000). Power prior distributions for regression models. *Statistical Science*, **15**, 46–60.
- Ibrahim, J. G., Chen, M.-H., & Sinha, D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association*, **98**, 204–213.
- James, W., & Stein, C. (1961). Estimation with quadratic loss. In J. Neyman (Ed.), *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability: Vol.1. Contributions to the theory of statistics* (pp. 361–379). Berkeley, CA: University of California Press.
- Kaplan, E. L. (1952). Tensor notation and the sampling cumulants of k -statistics. *Biometrika*, **39**, 319–323.
- Ogasawara, H. (2014). Optimization of the Gaussian and Jeffreys power priors with emphasis on the canonical parameters in the exponential family. *Behaviormetrika*, **41**, 195–223.
- Ogasawara, H. (2015). Bias adjustment minimizing the asymptotic mean square error. *Communications in Statistics — Theory and Methods*, **44**, 3503–3522.
- Ogasawara, H. (2016a). Asymptotic expansions for the estimators of Lagrange multipliers and associated parameters by the maximum likelihood and weighted score methods. *Journal of Multivariate Analysis*, **147**, 20–37.
- Ogasawara, H. (2016b). Bias correction of the Akaike information criterion in factor analysis. *Journal of Multivariate Analysis*, **149**, 144–159.
- Ogasawara, H. (2016c). Optimal information criteria minimizing their asymptotic mean square errors. *Sankhyā, B*, **78**, 152–182.
- Ogasawara, H. (2017). A family of the adjusted estimators maximizing the asymptotic predictive expected log-likelihood. *Behaviormetrika*, **44**, 57–95.
- Ogasawara, H. (2018). Supplement to the paper “Predictive estimation of a covariance matrix and its structural parameters”. To appear in *Economic Review*

- (*Otaru University of Commerce*), <http://www.res.otaru-uc.ac.jp/~emt-hogasa/>, <https://barrel.repo.nii.ac.jp/>.
- Sheena, Y., & Takemura, A. (1992). Inadmissibility of non-order-preserving orthogonality invariant estimators of the covariance matrix under Stein's loss. *Journal of Multivariate Analysis*, **41**, 117–131.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society, B*, **39**, 44–47.
- Takeuchi, K. (1976). Distributions of information statistics and criteria of the goodness of models. *Mathematical Science*, **153**, 12–18. (in Japanese)
- Takezawa, K. (2012a). A revision of *AIC* for normal error models. *Open Journal of Statistics*, **2**, 309–312.
- Takezawa, K. (2012b). *Learning regression analysis by simulation*. New York: Springer.
- Takezawa, K. (2015a). Estimation of the exponential distribution in the light of future data. *British Journal of Mathematics and Computer Science*, **5**(1), 128–132.
- Takezawa, K. (2015b). Estimation of binomial distribution in the light of future data. *British Journal of Mathematics and Computer Science*, **10**(2), 1–7.
- Takezawa, K. (2015c). Estimation of the geometric distribution in the light of future data. *British Journal of Mathematics and Computer Science*, **11**(4), 1–11.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, B*, **58**, 267–288.
- Ye, R.-D., & Wang, S.-G. (2009). Improved estimation of the covariance matrix under Stein's loss. *Statistics and Probability Letters*, **79**, 715–721.

(Received: June 24, 2017, Accepted: September 25, 2017)