An Investigation and Development of Three Estimators of Inverse Covariance Matrices with Applications to the Mahalanobis Distance

H.E.T. HOLGERSSON AND PETER S. KARLSSON
An Investigation and Development of Three Estimators of Inverse Covariance Matrices with Applications to the Mahalanobis Distance

H.E.T. Holgersson
Jönköping University
SE-551 11 Jönköping
Sweden

Peter S. Karlsson
Jönköping University
SE-551 11 Jönköping
Sweden

Abstract. This paper treats the problem of estimating the inverse covariance matrix in an increasing dimension context. Specifically, three ridge-type estimators are considered, of which two new are proposed by the authors and one has been considered previously. Risk functions for deciding an appropriate value of the ridge coefficient are developed and the finite sample properties of the estimators are investigated in a Monte Carlo simulation. Moreover, risk functions for the Mahalanobis distance are derived which, in turn, leads to three new estimators which has not been considered previously.

Keywords. Ridge estimators, inverse covariance matrix, increasing dimension, Mahalanobis distance, adaptive estimator.

Corresponding author: thomas.holgersson@ihh.hj.se, +46 36 - 10 17 48
I Introduction

Statistical analyses frequently involve co-variations between a large number of variables, e.g. cluster analysis, principal component analysis or discriminant analysis which requires estimators of covariance matrix. The sampling properties of such estimators, however, are known to be poor if the dimension of the statistic \((p)\) is comparable to the sample size \((n)\). Moreover, in cases when \(p\) grows asymptotically with \(n\) it frequently happens that the statistic does not converge to its true parameter – a well-known example is the non-convergence of the sample periodogram. While such increasing dimension asymptotics (IDA) have long been considered in time series contexts, e.g. White and Domowitz (1984), Newey and West (1987), they have been given less attention in the cross-section context. In 1967, however, Marchenko and Pastur published a pioneering paper on the convergence of spectral functions of sample covariance matrices for zero-mean and unit-variance variables, where \(p\) and \(n\) were allowed to grow simultaneously in such a way that their quotient limits a constant. Later on, a number of important contributions were made to IDA, many of which involved relaxing the moment restrictions. Some important references include Jonsson (1982), Yin, Bai and Krishnaiah (1988) and Silverstein and Bai (1995) who derived limits for the eigenvalues. Girko (1975, 1990) used spectral theory (basically, traces of resolvents of random matrices) to derive more general IDA properties of eigenvalues and determinants of covariance matrices. Many of these results were also refined and put in a more applied context. In particular, Serdobolskii (1985, 2000, 2008) used spectral theory to derive IDA properties of the inverse covariance matrix which in turn was used in problems concerning discriminant analysis, estimation of high-dimensional expectation vectors and regression analysis. These contributions involve estimators which, unlike traditional estimators, have risk functions that are bounded even when the dimension of the parameter grows asymptotically and they therefore provide highly interesting alternatives to traditional methods. Little, however, is known about these estimators’ behavior in finite, given combinations of \(n\) and \(p\). Moreover, they are all shrinkage estimators in the sense that they depend on a coefficient that pulls the estimator towards a certain point. For these reasons it is of great interest to investigate the proposed IDA estimators relative to traditional estimators and other shrinkage estimators. Most of the research of ridge estimators
conducted so far has focused on the search of an optimal ridge coefficient within a given family of estimators. Little is therefore known about the superiority of one family of one estimator to another. One purpose of this paper is therefore to investigate the properties of resolvent-type estimators proposed in the literature and to compare it to other shrinkage estimators of different functional form. Specifically, three types of shrinkage estimators of the inverse covariance matrix are considered: (i) the multiplicative resolvent-type estimator proposed by Girko (1995) and Serdobolskii (1985, 2000, 2008) whose finite sample properties have only been investigated for some highly restrictive cases previously, (ii) an additive resolvent-type estimator which does not seem to have been given any attention in terms of estimating inverse covariance matrices but has been frequently used in regression contexts (Hoerl and Kennard 1970a, 1970b) and (iii) a new type of shrinkage estimator that may be considered as a mixture of the first two. The IDA properties of these estimators are examined with respect to two different squared error loss functions (describing the risk of the estimators) by means of Monte Carlo simulations. Moreover, it is demonstrated that the risk function proposed by Serdobolskii (2000) for finding an optimal value of the ridge coefficient within the resolvent-type estimator may be modified to comprise the other two estimators as well, and hence that all three estimators are feasible. There is also an issue surrounding the properties of the three estimators when used within a composite statistic, since the inverse covariance matrix alone is only of partial interest. Therefore the simulations also includes the Mahalanobis distance, which is a function of the inverse covariance matrix. Furthermore, adaptive risk functions for identifying the optimal ridge coefficients with respect to the Mahalanobis distance residual sum of square are developed, and their consistencies are proved. In other words, feasible estimators are developed for estimating the Mahalanobis distance in IDA contexts. These, in turn, should be of great interest as a tool for outlier detection in IDA contexts.
2 Estimation of covariance matrices

A frequently used estimator of \( \text{Cov}(X) = \Sigma_{(p \times p)} \) is given by

\[
S_{p \times p} = n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})',
\]

(2.1)

where \( \{X_i\}_{i=1}^{n} \) is a set of \( n \) independent observations on a \( p \)-dimensional random vector \( X_{(p \times 1)} \) (sometimes with a small-sample adjustment of the divisor). It is well known that \( S \) converges in probability to \( \Sigma \) under fairly mild moment restrictions when \( p \) is fixed and finite. Alternative estimators have also been proposed. Important references include Kubokawa and Srivastava (2003), Daniels and Kass (2001), Shäfer and Strimmer (2005). In this paper no further attention will be paid to the problem of estimating \( \Sigma \), instead concentration will be on the estimation of its inverse \( \Sigma^{-1} \). An obvious estimator is given by the inverse of the standard estimator, \( S^{-1} \). This estimator, however, is not well behaved if \( p \) increases with the same rate as \( n \). For example, setting

\[
R_p(\hat{\Sigma}^{-1}) = E[p^{-1}\text{tr}(\hat{\Sigma}^{-1} - \Sigma^{-1})^2],
\]

where \( E[\cdot] \) is the expectation operator, \( \text{tr}(\cdot) \) is the trace operator of a matrix and letting \( \hat{\Sigma}^{-1} = S^{-1} \), the limit in (2.2) is

\[
R(S^{-1}) = \lim_{p \to \infty} R_p(S^{-1}).
\]

(2.3)

The function (2.3) limits infinity if \( p/n \to q^{-1} > 0 \) and so the standard estimator is degenerate in that case. Therefore much attention has been paid to developing better estimators of \( \Sigma^{-1} \). One family of estimators consists of scalar transformations of the standard estimator, say \( \hat{\Sigma}^{-1} = \alpha S^{-1} \). Such estimators are frequently referred to as Stein-type estimators or shrinkage estimators as \( \alpha \) is usually smaller than one. While there are estimators within this family which have asymptotically bounded risk functions even for \( q \to 1 \), they are often of little value in cases when \( \Sigma^{-1} \) is to be used
in a wider context, e.g. when the statistic of main interest is a composite function $g(\Sigma^{-1})$. This is so because scalar transformations will usually result in mappings of the type $g(\alpha \Sigma) = \alpha g(\Sigma^{-1})$, or even $g(\alpha \Sigma^{-1}) = g(\Sigma^{-1})$ as in e.g. Zellner's SUR models (Zellner, 1962). For this reason such estimators seem to exercise their main potential in cases when the sole purpose is to estimate $\Sigma^{-1}$ alone, and therefore other estimators that have a wider field of usages will be considered instead. Of particular interest is an estimator defined by

$$\hat{\Sigma}^{-1} = \hat{\Sigma}(t) = (I + tS)^{-1}, \ 0 \leq t. \quad (2.4)$$

The normalized trace of an estimator of the inverse covariance matrix in the form of the estimator (2.4), in the literature called the reolvent*, is denoted by $\hat{h}(t) = p^{-1}trace(\hat{\Sigma}^{-1})$. A key property of (2.4) is that its trace limits a fixed constant as $p$ increases whereas $p^{-1}trace(S^{-1})$ does not, as discussed above. Moreover, (2.4) also plays an important role as an estimate of $\Sigma^{-1}$, and its theoretical IDA properties have been investigated by Girko (1990) and Serdobolskii (1985, 2000, 2008). The key difference between $S^{-1}$ and $\hat{\Sigma}^{-1}$ may be examined by their eigenvalues: let $\lambda_i$ be the $i$:th ordered eigenvalue of $S$. Then the $i$:th eigenvalue of $S^{-1}$ is $\lambda_i^{-1}$ while the $i$:th eigenvalue of $\hat{\Sigma}^{-1}$ is $(1+t\lambda_i)^{-1}$. Hence the eigenvalue of $\hat{\Sigma}^{-1}$, say $\hat{\lambda}_i^{-1}$, is always bounded above zero. The idea of ensuring that the smallest eigenvalue of an inverse quadratic form does not become too small, and hence produce estimators with large sampling variance, is not unique to resolvent-type estimators. In the early 1970s, Hoerl and Kennard (1970a, 1970b) proposed a family of ridge estimators in a context of regression analysis, though similar estimators were also proposed before that (e.g. Foster, 1961). Since estimators of regression parameters involve the inversion of a quadratic form, say $(XX)^{-1}$, these authors proposed the alternative statistic $(tI+XX)^{-1}$ for some $t > 0$, thus ensuring that the smallest eigenvalues do not come too close to zero.
In terms of estimating inverse covariance matrixes, a similar estimator may be expressed by

$$\tilde{\Sigma}^{-1} = (I + S)^{-1}, \quad 0 \leq t.$$  \hspace{1cm} (2.5)

While this kind of estimator has been extensively used in regression contexts (see e.g. Hoerl and Kennard 1970a, 1970b; Obenchain 1975, 1978; Goldstein and Smith 1974; Kibria 2003 and Gruber 1998) its potential as an estimator of the inverse covariance matrix has been given less attention in the past, especially in IDA contexts. The estimators (2.4) and (2.5) seem very alike but are actually quite different. The $i$:th eigenvalue of $\tilde{\Sigma}^{-1}$, say $\tilde{\lambda}^{-1}$, is given by $(t + \lambda)^{-1}$ and it is readily seen that there is no functional relationship between $\tilde{\lambda}^{-1}$ and $\tilde{\lambda}^{-1}$. The ridge coefficient $t$ operates additively in $\tilde{\lambda}^{-1}$ whereas it operates multiplicatively in $\hat{\lambda}^{-1}$. A compromise between the two estimators may be obtained by

$$\tilde{\Sigma}^{-1} = (tI + (1-t)S)^{-1}, \quad 0 \leq t \leq 1.$$  \hspace{1cm} (2.6)

This estimator has not been considered previously and may be thought of as a compromise between the two others since the ridge coefficient $t$ operates in both a multiplicative and an additive way. The estimator (2.4) will hence forth be referred to as the multiplicative estimator, (2.5) will be referred to as the additive estimator while the estimator (2.6) will be referred to as the mixture estimator. When it comes to the properties of the above estimators, (2.4) has been fairly well documented analytically (Serdobolskii, 1985, 2000, 2008; Girko, 1990). The properties of the estimators (2.5) and (2.6) are not known however, and are not easy to obtain analytically, at least not within an IDA context, and are therefore subject of investigation. Estimators of inverse covariance matrixes are frequently described with respect to the loss function

$$L_t(\hat{\Sigma}^{-1}, \Sigma^{-1}) = p^{-1}tr\left((\hat{\Sigma}^{-1} - \Sigma^{-1})^2\right).$$  \hspace{1cm} (2.7)
An alternative loss function is available from

\[ L_2\left( \hat{\Sigma}^{-1}, \Sigma^{-1} \right) = p^{-1} tr \left( \hat{\Sigma}^{-1} \Sigma - I \right)^2. \]  \hspace{1cm} (2.8)

Other loss functions have also been considered in the literature (e.g. Stein, 1956) but the paper will be restricted to (2.7) and (2.8). It should be noted that these loss functions are usually presented without the divisor \( p^{-1} \) when the dimension \( p \) is taken to be fixed, but as the argument within the trace operator is monotonically increasing in \( p \), it is convenient to work with normalized measures. Some asymptotic properties of the estimator (2.4) with respect to (2.7) are given in Serdobolskii (1985, 2000, 2008) but little is known about the properties of (2.5) or (2.6) with respect to any of these measures, particularly for increasing dimension asymptotics. The next section is concerned with this matter.

* Functions of the kind \((A - kI)^{-1}\) are usually referred to as ‘resolvents’ in the functional analysis literature (e.g. Kolmogorov and Fomin, 1957) while the same term has been used for the function \((kA + I)^{-1}\) in for instance Girko (1990) and Serdobolskii (1985, 2000).

3 Risk functions

Ridge estimators are usually connected with the problem of choosing an optimal value of the ridge coefficient \( t \) in this case. For a double coefficient estimator of the type

\[ \alpha \hat{\Sigma}^{-1} = \alpha (1 + \tau S)^{-1}, \quad \alpha \geq 0, \quad \tau \geq 0, \]  \hspace{1cm} (3.1)

the limit quadratic risk defined in (2.3) is the limit of

\[ R_p \left( \hat{\Sigma}^{-1} \right) = p^{-1} tr \left( \Sigma^{-2} \right) - 2 p^{-1} E \left[ tr \left( \Sigma^{-1} \hat{\Sigma}^{-1} \right) \right] + p^{-1} E \left[ tr \left( \Sigma^{-2} \right) \right]. \]  \hspace{1cm} (3.2)

If the type of convergence is not explicitly mentioned, then it is assumed that the convergence is in mean square. In order to establish consistent estimators of the limit risk functions of (2.5) and (2.6) we will use the following result:
It is known that if \( X \sim N(0, \Sigma) \), \( F_p(u) = p^{-1} \sum_{i=1}^{p} \text{ind}(\lambda_i \leq u) \rightarrow F(u) \) where \( \lambda_i \) are eigenvalues of \( \Sigma \) such that \( \lambda_i \in [c_1, c_2] \) where \( c_1 > 0 \) and \( c_2 < \sqrt{3} \| \Sigma \| \), \( \| \| \) denotes the spectral norm of a matrix, \( \text{ind}(\lambda_i \leq u) \) denotes an indicator function of the \( i \)th eigenvalue, \( F(u) \in \mathbb{R} \) is the limit spectral distribution function and \( p/n \rightarrow (0 < y < 1) \), then a statistic approximating (3.2) is available from:

\[
\hat{R}_p\left(\hat{\Sigma}^{-1}\right) = \hat{\Lambda}_{-2} - 2\alpha \left( p^{-1} - n^{-1} \right) \text{tr}\left(S^{-1}\hat{\Sigma}^{-1}\right) + 2\alpha p^{-1}n^{-1} \left( \text{tr}\left(S^{-1}\right) \right)^2 + \alpha^2 p^{-1} \text{tr}\hat{\Sigma}^{-2},
\]

(3.3)

where, \( \hat{\Sigma}^{-1} = (I + \tau S)^{-1} \), \( \hat{\Lambda}_{-1} = \left( p^{-1} - n^{-1} \right) \text{tr}\left(S^{-1}\right) \), \( \hat{\Lambda}_{-2} = \left( 1 - pn^{-1} \right)^2 p^{-1} \text{tr}\left(S^{-2}\right) + \left( 1 - pn^{-1} \right) p^{-1}n^{-1} \left( \text{tr}\left(S^{-1}\right) \right)^2 \). The result is due to (Serdobolski, 2000).

For a fixed value of \( \alpha \) the minimum of \( \hat{R}_p\left(\hat{\Sigma}^{-1}\right) \) can be found numerically or graphically. An operational estimator of (2.4) is hence available by first setting \( \alpha = 1 \) and defining \( \hat{t} = \tau_{\text{opt.}} = \min_{\tau} \hat{R}_p \) and then defining an operational estimator by

\[
\hat{\Sigma}^{-1} = (I + \hat{\tau} S)^{-1}.
\]

(3.4)

The risk function (3.3) is actually rich enough to also provide estimates of the ridge coefficients of the other two estimators as well. To see this, note that (2.5) and (2.6) may be expressed as follows:

\[
\hat{\Sigma}^{-1} = (I + S)^{-1} = \alpha (I + \tau S)^{-1}, \quad \alpha = t^{-1}, \quad \tau = t^{-1},
\]

(3.5)

and

\[
\hat{\Sigma}^{-1} = (I + (1-t)S)^{-1} = \alpha (I + \tau S)^{-1}, \quad \alpha = t^{-1}, \quad \tau = t^{-1}(t-1).
\]

(3.6)

In other words, all three ridge estimators considered in the paper may be written in the form (3.1) and optimal values of the respective ridge coefficient \( t \) may be obtained...
from (3.3) numerically. Letting \( \tilde{t} \) and \( \bar{t} \) denote the values optimising \( \hat{R}_p(\hat{\Sigma}^{-1}) \) with respect to (3.5) and (3.6) respectively, the following estimators are obtained

\[
\hat{\Sigma}^{-1} = (\hat{\Pi} + S)^{-1}, \tag{3.7}
\]

and

\[
\bar{\Sigma}^{-1} = (\bar{\Pi} + (1-\bar{t})S)^{-1}. \tag{3.8}
\]

Hence three alternative estimators for the inverse covariance matrix in IDA settings are available. The finite sample properties of these estimators need to be compared with respect to different values of \( n \) and \( p \), and also to different distributional properties of \( \mathbf{X} \). Moreover, the inverse covariance matrix is often used as a component within another statistic and optimality of \( \hat{\Sigma}^{-1} \), for example, is not necessarily preserved when used as an ingredient within a composite statistic. Some examples involving the inverse covariance matrixes are discriminant analysis (Pavlenko and von Rosen, 2001), Mardia’s measures of skewness and kurtosis (Mardia, 1970), Zellner’s SUR models (1962) and the Mahalanobis distance (Mahalanobis, 1936). The next section will therefore treat the latter with respect to its IDA properties.

4 The inverse covariance matrix within the Mahalanobis distance

The Mahalanobis distance for a single observation is defined by

\[
d_i^2 = p^{-1} Y_i' \Sigma^{-1} Y_i, \tag{4.1}
\]

where \( Y_i = (X_i - \mu_X) \) and the mean value vector \( \mu_X \) is assumed to be known. This measure is important in diagnostics, particularly in the search for outliers in \( \mathbb{R}^p \). An estimator of (4.1) may be obtained by simply replacing the inverse covariance matrix with an appropriate estimate, say \( \hat{\Sigma}^{-1} \):

\[
\hat{d}_i^2 = p^{-1} Y_i' \hat{\Sigma}^{-1} Y_i. \tag{4.2}
\]
This estimator has been considered by, for instance, Mardia (1977) and Wilks (1963) who also proposed methods for explicit outlier tests, but the investigation will here be restricted to the estimator (4.2) and its properties with respect to point estimation in IDA contexts. Clearly, the stochastic properties of (4.2) will depend on the estimator $\hat{\Sigma}^{-1}$ and hence, if $\hat{\Sigma}^{-1} = \hat{\Sigma}^{-1}(t)$, a tool for deciding an appropriate value of the ridge coefficient $t$ is needed. The loss function considered in Section 3 is not necessarily useful in this context. A more specialized choice, which may be thought of as the expected value of the residual sum of squares between (4.1) and (4.2), is available by

$$R^2(\hat{d}_i^2) = p^{-1}E\left[ Y_i'(\hat{\Sigma}^{-1} - \Sigma^{-1})^2 Y_i \right].$$  \hspace{1cm} (4.3)

To achieve an estimator of (4.3) it is convenient to expand it as follows:

$$pR^2(\hat{d}_i^2) = n^{-1}E\left[ \sum_{i=1}^n Y_i'(\hat{\Sigma}^{-1} - \Sigma^{-1})^2 Y_i \right] = n^{-1}E\left[ tr\left( \sum_{i=1}^n (\hat{\Sigma}^{-1} - \Sigma^{-1})^2 Y_i Y_i' \right) \right] =$$

$$E\left[ tr\left( (\hat{\Sigma}^{-1} - \Sigma^{-1})^2 S \right) \right] = tr\left( E\left[ \Sigma^{-1} S \right] \right) - 2tr\left( E\left[ \Sigma^{-1} S \Sigma^{-1} S \right] \right) + tr\left( E\left[ \Sigma^{-2} S \right] \right),$$ \hspace{1cm} (4.4)

where $S = n^{-1}\sum_{i=1}^n Y_i Y_i'$. In order to derive a consistent estimate of (4.4) some simplifying lemmas, under the assumption that $S$ is of full rank, are listed below:

**Lemma 1**: $\hat{\Sigma}^{-1} = I - t S \hat{\Sigma}^{-1}$ where $\hat{\Sigma}^{-1} = (I + t S)^{-1}$.

**Proof**: Post-multiply both sides by $(I + t S)$.

**Lemma 2**: $\hat{\Sigma}^{-1} S = t^{-1} (I - \hat{\Sigma}^{-1})$.

**Proof**: By Lemma 1, $t^{-1} (I - \hat{\Sigma}^{-1}) = S \hat{\Sigma}^{-1}$, and since $t^{-1} (I - \hat{\Sigma}^{-1})$ is symmetric, $S \hat{\Sigma}^{-1}$ is also symmetric and the Lemma follows.

**Lemma 3**: $\hat{\Sigma}^{-1} = t^{-1} (I - \hat{\Sigma}^{-1}) S^{-1}$.

**Proof**: Follows from Lemma 2.
Theorem 1: Let $Y_i$ be distributed as the variable of (3.3). Then, if $p/n \to y$, where $0 < y < 1$, the following convergence holds:

$$\hat{R}_2^d(t) = t^{-1}\left\{ \hat{h}(t) - p^{-1}\text{tr}\left(\hat{\Sigma}^{-2}\right) \right\} - 2\left\{ \hat{h}(t) \hat{s}(t) \right\} + \hat{\Lambda}_1 \to R_2^d(t),$$

where $\hat{h}(t) = p^{-1}\text{tr}\left((1+tS)^{-1}\right)$, $\hat{s}(t) = 1 - y + \hat{y}(t)$ and $\hat{\Lambda}_1 = (p^{-1} - n^{-1})\text{tr}\left(S^{-1}\right)$.

Proof: By Lemma 3 it follows that

$$E\left[ p^{-1}\text{tr}\left(\hat{\Sigma}^{-1}\left(\hat{\Sigma}^{-1}\right)S\right) \right] = E\left[ p^{-1}\text{tr}\left(t^{-1}\hat{\Sigma}^{-1}\right) \right] - E\left[ p^{-1}\text{tr}\left(t^{-1}\hat{\Sigma}^{-2}\right) \right].$$

In Serdobolskii (2000, 2008) it is shown that $p^{-1}\text{tr}\left(\hat{\Sigma}^{-1}\right) \to E\left[ p^{-1}\text{tr}\left(\Sigma^{-1}\right) \right]$ and $p^{-1}\text{tr}\left(\hat{\Sigma}^{-2}\right) \to E\left[ p^{-1}\text{tr}\left(t^{-1}\Sigma^{-2}\right) \right]$ which proves convergence of the first term of $\hat{R}_2^d(t)$.

By Lemma 2 it follows that

$$E\left[ p^{-1}\text{tr}\left(\Sigma^{-1}\left(\Sigma^{-1}\right)S\right) \right] = E\left[ p^{-1}\text{tr}\left(t^{-1}\Sigma^{-1}\right) \right] - E\left[ p^{-1}\text{tr}\left(t^{-1}\Sigma^{-1}\Sigma^{-1}\right) \right].$$

Moreover, in Serdobolskii (2008) it is shown that $\hat{\Lambda}_1 \to E\left[ p^{-1}\text{tr}\left(\Sigma^{-1}\right) \right]$ and

$$\left(\hat{\Lambda}_1 - \hat{h}(-t)\hat{s}(-t)\right) \to E\left[ p^{-1}\text{tr}\left(\Sigma^{-1}\Sigma^{-1}\right) \right],$$

and $E[\Sigma] = \Sigma + o\left(p^0\right)K$, where $K$ is a symmetric positively semidefinite matrix consisting of residual terms with bounded norm so that $E\left[n^{-1}\left(\Sigma^{-2}\right)\right] = p^{-1}\Sigma^{-2}\Sigma^{-1} + o\left(p^0\right)K$. Hence $\hat{h}(-t)\hat{s}(-t)$ converges to the second term of $\hat{R}_2^d(t)$, $\lim_{n \to \infty} E\left[n^{-1}\text{tr}\left(\Sigma^{-2}\Sigma^{-1}\right) \right] = \Lambda_{-1}$, and since $\hat{\Lambda}_1 \to \Lambda_{-1}$ the proof is complete.

Corollary 1: From the assumptions and definitions of Theorem 1,

$$\hat{R}_2^d(\tau, \alpha) = \tau^{-1}\alpha^2\left\{ p^{-1}\text{tr}\left(\hat{\Sigma}^{-1}\right) + p^{-1}\text{tr}\left(\hat{\Sigma}^{-2}\right) \right\} - 2\alpha\hat{h}(-\tau)\hat{s}(-\tau) + \hat{\Lambda}_1 \to R_2^d(\tau, \alpha),$$

where $\alpha \in \mathbb{R}$ is a non-random constant and $\hat{h}(t, \alpha) = p^{-1}\alpha\text{tr}\left((1+tS)^{-1}\right)$, $\hat{s}(t, \alpha) = 1 - y + \hat{y}(t, \alpha)$, $\hat{\Lambda}_1 = (p^{-1} - n^{-1})\text{tr}\left(S^{-1}\right)$ and $\tau$ is a function of $t$ determined by the inverse covariance-estimator (3.5) and (3.6).
\textbf{Proof:} Follows from the proof of (3.3).

The risk function (4.4) provides a tool for deciding an appropriate value of the ridge coefficient \( t \) for estimating the Mahalanobis distance between \( \mathbf{X}_i \) and \( \mu \). Moreover, the estimator \((\mathbf{I}+t\mathbf{S})^{-1}\) within (4.2) may be replaced by the double coefficient estimator \( \alpha(\mathbf{I}+r\mathbf{S})^{-1} \) which in turn encompasses all three estimators (2.4) - (2.6).

Hence two alternatives to (4.2) are given by

\[
\tilde{d}_i = p^{-1}\mathbf{Y}_i'\tilde{\Sigma}^{-1}\mathbf{Y}_i,
\]

(4.5)

\[
\overline{d}_i = p^{-1}\mathbf{Y}_i'\bar{\Sigma}^{-1}\mathbf{Y}_i.
\]

(4.6)

Through Corollary 1 a tool for deciding appropriate values of the ridge coefficient for any of (4.2), (4.5) and (4.6) is available. Adhering to the notation of (3.4) – (3.6) the following feasible estimators are obtained:

\[
\hat{d}_i = p^{-1}\mathbf{Y}_i'\hat{\Sigma}^{-1}\mathbf{Y}_i.
\]

(4.7)

\[
\hat{\hat{d}}_i = p^{-1}\mathbf{Y}_i'\hat{\hat{\Sigma}}^{-1}\mathbf{Y}_i.
\]

(4.8)

\[
\bar{d}_i = p^{-1}\mathbf{Y}_i'\bar{\Sigma}^{-1}\mathbf{Y}_i.
\]

(4.9)

Thus, using any of the estimators (4.7) – (4.9) its efficiency may be evaluated with respect to (4.3) by Monte Carlo simulations. Further details are supplied in the next section.
This section investigates some properties of the estimators of the inverse covariance matrix \( \Sigma^{-1} \) and the Mahalanobis distance \( p^{-1}Y_i^\prime \Sigma^{-1}Y_i \) considered in the paper. In order to do this, various simulations will be conducted while varying relevant parameters. The properties of any estimator of the inverse covariance matrix will depend upon the true covariance matrix \( \Sigma \). Two types of covariance matrixes are presented below which are simple to generate for any value of the dimension \( p \):

a. \( \Sigma_{a,(p \times p)}(\phi) = \text{toeplitz}(\theta^0, \theta^1, \ldots, \theta^{(p-1)}), |\phi| < 1 \).

b. \( \Sigma_{b,(p \times p)} = P_{p \times p} \Lambda_{p \times p} P_{p \times p}^\prime \) where \( P \) is the eigenvector matrix of \( \Sigma \) and \( \Lambda \) is a diagonal matrix consisting of the eigenvalues of \( \Sigma \). \( P \) is obtained by drawing a set of numbers \( \{ P_{ij} \}_{i,j} \) from a uniform distribution on the interval \( (1,10) \) using the ISBN number of (Rao, 1973) as random seed, and then applying the singular value decomposition to make the columns orthonormal. The eigenvalues \( \{ \Lambda_1, \ldots, \Lambda_p \} \) were drawn from a \( \chi^2 \) distribution. \( P \Lambda P^\prime = \Sigma_b \) is then held fixed in all replicates.

The above covariance matrixes are both interesting in their own way. The multiplicative estimator (2.4) limits the identity matrix as the ridge coefficient \( t \) limits zero which is equivalent to the special case \( \Sigma_a(0) \), while it limits a singular matrix when \( \phi \) limits unity. Moreover, the covariance matrix in (a) has its eigenvalues bounded on a finite segment, i.e. \( 0 < \lambda_{\min} \leq \lambda_{\max} < \infty \). Since most of the methods treated in the paper assume the eigenvalues to be bounded, it is of interest to investigate covariance matrixes that violate this assumption, though in a modest sense. This is achieved by the covariance matrix in (b) since the eigenvalues obtained by a single drawing of numbers of the chi-square distribution have the property that \( \lim_{p \to \infty} \lambda_{\max} = \infty \). The two types of covariance matrixes therefore represent rather different aspects of the true covariance matrix. Apart from the parametric structure of the covariance matrix, it is also important to consider transformations of the type \( \Sigma \mapsto \beta \Sigma \) for some scalar parameter \( \beta \). Hence both the functional form of \( \Sigma \) and scalar transformations thereof will be considered. Another quantity of relevance is the
The multiplicative estimator has its greatest potential relative to the standard estimator $S^{-1}$ for values of $q$ just above the value 1, say $1 < q < 5$. Moreover, the risk function of equation (3.3) can only be considered to be a good approximation to the true risk function for large values of $p$. Therefore the simulations will also include a range of values of $p$. Carefully selected combinations of the above parameterisations should allow for investigations of the main properties of the considered estimators. The parametric setting used in the experiment in Tables 1 and 2 are listed below:

**Table 1 Parameters used in simulation A (inverse covariance matrix)**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional form of covariance matrix</td>
<td>$\Sigma_a, \Sigma_b$</td>
<td>Equations (a, b)</td>
</tr>
<tr>
<td>Scalar transformation</td>
<td>$\beta$</td>
<td>1, 50</td>
</tr>
<tr>
<td>Dimension of covariance matrix</td>
<td>$p$</td>
<td>5, 10, 25, 50, 100, 150, 200</td>
</tr>
<tr>
<td>Sample quotient</td>
<td>$q = n / p$</td>
<td>1.1, 2, ..., 8</td>
</tr>
<tr>
<td>Distribution of $\varepsilon$</td>
<td>$\varepsilon_{ij} \sim N(0,1)$</td>
<td>Normal distribution</td>
</tr>
</tbody>
</table>

**Table 2 Parameters used in simulation B (Mahalanobis distance)**

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional form of covariance matrix</td>
<td>$\Sigma_a$</td>
<td>Equation (a)</td>
</tr>
<tr>
<td>Scalar transformation</td>
<td>$\beta$</td>
<td>5, 500</td>
</tr>
<tr>
<td>Correlation parameter</td>
<td>$\phi$</td>
<td>0, 0.9</td>
</tr>
<tr>
<td>Dimension of covariance matrix</td>
<td>$p$</td>
<td>5, 10, 25, 50</td>
</tr>
<tr>
<td>Sample quotient</td>
<td>$q = n / p$</td>
<td>1.1, 2, ..., 5.5</td>
</tr>
<tr>
<td>Distribution of $\varepsilon$</td>
<td>$\varepsilon_{ij} \sim N(0,1)$</td>
<td>Normal distribution</td>
</tr>
<tr>
<td></td>
<td>$\varepsilon_{ij} \sim 2{\text{Exp}(2)-(1/2)}$</td>
<td>Exponential distribution</td>
</tr>
</tbody>
</table>

The variables used to estimate the inverse covariance matrixes are defined by $X' = \beta (L_{(p \times p)} + \varepsilon_{(p \times n)})$ where $LL' = \Sigma$. The matrix $\Sigma$ is held constant in all replications while a new $\varepsilon$ is drawn in each replicate. For each combination of parameters 10 000 replicates were used.
A. Estimators of inverse covariance matrixes

In order to make the simulations easier to interpret the loss function of each ridge estimator has been divided with the corresponding loss for the standard estimator, e.g. \( \frac{L_1(\hat{\Sigma}^{-1})}{L_1(S^{-1})} \). The ridge coefficient of the mixture estimator ranges between zero and one, but is in the below graphs scaled to fit in the same graphs as the other estimators. Figures 1-2 present the risk of the three estimators (2.4), (2.5) and (2.6) when \( p = 150 \) and \( q \) increase vertically (between rows of graphs). In Figure 1 all three ridge estimators dominate \( S^{-1} \) for all values of the ridge coefficient \( t \) when \( q = 1.1 \) (the quotients being uniformly below 1). For \( q = 5 \), however, it is seen that the ridge estimators are inferior to the standard estimators for nearly all values of the ridge coefficient. In particular, the multiplicative estimator and the mixture estimator may perform drastically worse than \( S^{-1} \) for poor choices of \( t \). In comparison of Figure 1 and 2 it is seen that the higher value of \( \beta \) leads to higher relative risk for the multiplicative and mixture estimator respectively while the risk of the additive estimator actually decreases by this change. The difference between the three estimators generally becomes more accentuated for high values of \( \beta \). It is also interesting to compare the properties of the estimators with respect to the two loss functions (left vs. right column) in that good properties with respect to \( L_1 \) does not necessarily hold for \( L_2 \). For example, the difference between Figures 1e and 1f is notable. Moving on to the graphs within Figure 3, the effect of \( q \) is compared horizontally and the value of \( \beta \) vertically. Again it is seen that the ridge estimators possess their greatest potential for low values of \( q \) and that the difference between the three estimators becomes accentuated when \( \beta \) increases. In Figure 4 the performance of the estimators are investigated for a covariance matrix of type \( \Sigma_b \) as defined on p. 13. The ridge estimators may excel at estimating the inverse covariance matrix as seen from this figure, but, at the same time, they may also perform considerably worse than the standard estimator for inappropriately chosen values of \( t \). Hence it is of great importance to examine the properties of the estimators when the ridge coefficient is determined from data. Figure 5 therefore concerns the adaptive two-stage estimators, i.e. when the value of \( t \) has been obtained from (3.3). From these graphs it is evident that the multiplicative and the additive ridge estimators perform very well for values
of $q$ below 6, though high values of $\phi$ (= high covariances) deteriorate the relative performance. The mixture estimator is uniformly inferior to the other two, at least for the setting considered here. It should also be stressed that the estimator $S^{-1}$ is (for normal distribution and fixed value of $p$) well known to asymptotically limit the maximum likelihood estimator of $\Sigma^{-1}$ (Anderson, 2003). Hence it will be asymptotically efficient and outperform other estimators as the quotient $q$ increases, i.e., an estimator should optimally limit $S^{-1}$ when the quotient between $n$ and $p$ grows. The mixture and the additive estimator coincide with $S^{-1}$ for a specific value of the ridge coefficient ($t = 0$) which indeed is seen to happen asymptotically as $q$ grows (Figures 5c-5d). From this point of view, these two estimators should be considered just as useful as the multiplicative estimator even though the latter is frequently superior in the settings considered here.
Figure 1. $L_1$ and $L_2$ risk for $\beta = 1$, $p = 150$, $\phi = 0.9$ and $q = \{1.1, 2, 5\}$
Figure 2. $L_1$ and $L_2$ risk for $\beta = 50$, $p = 150$, $\phi = 0.9$ and $q = \{1.1, 2, 5\}$

Fig 2a. $L_1$ risk for $q = 1.1$

Fig 2b. $L_2$ risk for $q = 1.1$

Fig 2c. $L_1$ risk for $q = 2$

Fig 2d. $L_2$ risk for $q = 2$

Fig 2e. $L_1$ risk for $q = 5$

Fig 2f. $L_2$ risk for $q = 5$
Figure 3. $L_t$ risk for $\beta = \{1, 50\}$, $p = 150$, $\phi = 0$ and $q = \{1.1, 2\}$

Fig 3a. Risk for $\beta = 1, q = 1.1$

Fig 3b. Risk for $\beta = 1, q = 2$

Fig 3c. Risk for $\beta = 50, q = 1.1$

Fig 3d. Risk for $\beta = 50, q = 2$
Figure 4. $L_i$ risk for $\Sigma_b, \beta = \{1, 50\}, q = \{1.1, 2\}$

Fig 4a. Risk for $\beta = 1, q = 1.1$

Fig 4b. Risk for $\beta = 1, q = 2$

Fig 4c. Risk for $\beta = 50, q = 1.1$

Fig 4d. Risk for $\beta = 50, q = 2$
B. Estimators of the Mahalanobis distance

The estimators of the Mahalanobis distance are investigated in Figures 6 and 7. For low values of the scalar parameter $\beta$ (Figure 6a and 6b) the ridge estimators perform very well while for values equal to 500 the standard estimator will dominate for a wide range of the ridge coefficient. Moreover, it is perhaps a bit surprising that the behavior of the estimators is drastically affected by the distribution; if the distribution is changed from a standard normal distribution to a standardized exponential distribution when $q$ is low (Figure 6a and 6b) the relative performance of the ridge estimators improves markedly. This suggests that the standard estimator is more affected by asymmetric distributions than the ridge estimators. Moreover, from Figure 7 it is seen that all three adaptive estimators work very well, at least for values of $q$ just above unity. But when $q$ increases the multiplicative estimator asymptotically
deteriorates while the other two estimators limit the standard estimator. This, again, underlines the use of ridge estimators that coincide with the standard estimator for some value of the ridge coefficient.

Figure 6. \( R^2 \) risk for Normal and Exponential distribution, \( \beta = \{5, 500\}, q = \{1.1, 2\} \)

<table>
<thead>
<tr>
<th></th>
<th>Multiplicative</th>
<th>Mixed</th>
<th>Additive</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fig 6a. Risk for ~ ( N, \beta = 5, q = 1.1 )</td>
<td><img src="graph1.png" alt="Graph" /></td>
<td><img src="graph2.png" alt="Graph" /></td>
<td><img src="graph3.png" alt="Graph" /></td>
</tr>
<tr>
<td>Fig 6b. Risk for ~ ( \text{Exp}, \beta = 5, q = 1.1 )</td>
<td><img src="graph4.png" alt="Graph" /></td>
<td><img src="graph5.png" alt="Graph" /></td>
<td><img src="graph6.png" alt="Graph" /></td>
</tr>
<tr>
<td>Fig 6c. Risk for ~ ( N, \beta = 500, q = 2 )</td>
<td><img src="graph7.png" alt="Graph" /></td>
<td><img src="graph8.png" alt="Graph" /></td>
<td><img src="graph9.png" alt="Graph" /></td>
</tr>
<tr>
<td>Fig 6d. Risk for ~ ( \text{Exp}, \beta = 500, q = 2 )</td>
<td><img src="graph10.png" alt="Graph" /></td>
<td><img src="graph11.png" alt="Graph" /></td>
<td><img src="graph12.png" alt="Graph" /></td>
</tr>
</tbody>
</table>
6 Conclusive Summary

This paper concerns the functional form of the estimator of the inverse covariance matrix. The main attention lies in settings where the dimension of the covariance matrix is comparable to the sample size, or even growing asymptotically. Specifically, three different ridge-type estimators have been investigated of which two are proposed by the authors and one has been considered previously in the literature. The estimators are characterized by their different functional form in that the ridge coefficient operates multiplicatively, additively or as a mixture of these two. Risk functions for choosing optimal values of the ridge coefficients are derived in the paper and their corresponding two-stage (adaptive) estimators are examined through Monte Carlo simulations, in which the properties of the three estimators are demonstrated to be very unlike. The mixture estimator has some tractable properties but is not optimal.
for any of the parameter settings investigated in the paper. The additive ridge estimator on the other hand is shown to perform surprisingly well over a wide range of parameter settings which is interesting since it has not been considered previously in the literature. In particular, while the multiplicative estimator outperforms the additive only when the quotient between the sample size and the dimension is low, the additive estimator, as well as the mixture estimator, have the advantage of coinciding with the standard maximum likelihood estimator for a certain value of the ridge coefficient. The simulations reveal that the proposed adaptive estimators indeed coincides with the standard estimator precisely when it is optimal, i.e. when the sample size is large relative to the dimension. This is not achieved by the previously proposed multiplicative estimator due to its different functional form. Hence, even though the multiplicative estimator has superior properties for low values of the sample quotient the careful practitioner may for this reason prefer the additive estimator.

Moreover, three adaptive estimators of the Mahalanobis distance in high dimensional contexts are also proposed in the paper and appropriate risk functions have been derived for choosing the optimal ridge coefficient. Monte Carlo simulations show that the resulting estimators expectedly outperform the traditional estimator for a wide range of parameter settings. Hence a tool for outlier detection in high dimensional data is developed which should be useful in diagnostic analysis.
References


