CESIS Electronic Working Paper Series

Paper No. 362

Estimating Individual Mahalanobis Distance in High-Dimensional Data

Dai D.
Holgersson T.
Karlsson P.

May, 2014

The Royal Institute of technology
Centre of Excellence for Science and Innovation Studies (CESIS)
http://www.cesis.se

Estimating Individual Mahalanobis Distance in High-Dimensional Data

Dai. D. [1] Holgersson. T. [1,2] Karlsson. P.[1,2]

Abstract

This paper treats the problem of estimating individual Mahalanobis distances (MD) in cases when the dimension of the variable p is proportional to the sample size n. Asymptotic expected values are derived under the assumption $p/n \to c, 0 \leq c < 1$ for both the traditional and the leave-one-out estimators. It is shown that some estimators are asymptotically biased, but that biased corrected versions are available. Moreover, a risk function is derived for finding an optimal estimate of the inverse covariance matrix on which the MD depends. This is then used to identify the optimal estimate of the inverse covariance matrix which, unlike the standard estimator, yields efficient MD estimates over the whole range $0 \leq c < 1$.

Keywords: Increasing dimension data, Mahalanobis distance, Inverse covariance matrix, Smoothing

JEL Classification: C 55, C 38, C 46

1. Introduction

The Mahalanobis distance (MD) is a fundamental distance measure originally proposed by Mahalanobis (1936). It is used in various kinds of statistical problems

[1]Linnaeus university, SE 351 95, Växjö, Sweden
[2]Jönköping university, SE 551 11, Jönköping, Sweden

in which marginal variables within random vectors are intercorrelated. MD may be thought of as a studentized random vector, in the sense that the random vector of interest is centered and pre-multiplied by an orthogonalizing matrix, thereby transforming it to a new vector with asymptotically zero mean and uncorrelated marginals. Important cases involve the distance between a sample mean vector and a hypothesized mean value vector or that between sample mean vectors from two independent samples. In this case, the MD is used for hypothesis testing or the estimation of confidence ellipsoids (Rao, 1945) and within discriminant analysis (Fisher, 1940). But the MD is also used for measuring the distance between two individual observations of a random vector, frequently used in hierarchical cluster analysis (Friedman, Hastie and Tibshirani, 2001), and for assessing the assumption of multivariate normality (Mardia, 1974; Mardia, Kent and Bibby, 1980; Mitchell and Krzanowski, 1985; Holgersson and Shukur, 2001). Other uses involve the distance between individual observations of a random vector to its sample mean value. This distance is commonly used in the search for multivariate outliers (Wilks, 1963; Mardia, Kent and Bibby, 1980).

The above mentioned analyzes are usually fairly straightforward, and the asymptotical properties well established because of the relatively simple functional form of the MD. Recent attention, however, has been brought to applications of the MD where the sample size (n) asymptotically increase with the dimension of the random vector (p) to a constant, say $p/n \to c, 0 \leq c < 1$. This case complicates the situation considerably since the MD involves an estimator of the inverse covariance matrix, which in turn may be inconsistent or badly behaved in other senses when $c > 0$. While estimation of the inverse covariance matrix alone has been given some attention in the literature, such as in Serdobolskii (2000), Girko (1995) and Ledoit and Wolf (2004), Jonsson (1982), Efron and Morris (1976), it has been

given less attention in cases where it is used within a composite statistic, such as in the MD. An exception is Holgersson and Karlsson (2012), who proposed a family of additive ridge estimators for the MD in high-dimensional data. That method proved to outperform the standard MD estimator in a wide range of settings, but it depends on an unknown regularization coefficient that must be estimated from data, which in turn complicate its usage.

In this paper we consider a somewhat different family of estimators for the individual MD. We utilize a family of estimators of the inverse covariance matrix originally developed by Efron and Morris (1976). A risk function specifically designed for the MD is developed, and it is shown that any estimator of the inverse covariance matrix which minimizes this risk function simultaneously minimizes three different types of MDs – and that there is hence no need to treat the special cases individually. We also derive the optimal MD within the given family of estimators and compare it to the standard estimator through a small Monte Carlo simulation.

In Section 2 the MDs used in the paper are presented and defined, and the expected values of some standard MDs are derived under high-dimension asymptotics. Section 3 introduces a risk function used in the paper, while Section 4 derives the optimal estimator within this family. Finally, a brief summary is given in Section 5. Some straightforward derivations appear in the appendix.

2. Mahalanobis distances

There are several different types of MDs that arise in different applications, such as in inference of location, diagnostic testing, cluster analysis, and discriminant analysis. The distance measures usually involve either distances between sample and population mean value vectors or distances between individual observations and their expected values. This paper is concerned with the latter case. We define some specific MD measures and estimates thereof. Of particular interest are the expected values of these estimators in high-dimensional settings – that is, in cases where the dimension of data increases simultaneously with the sample size, since this has not been considered previously in the literature. The three different MDs considered in the paper are defined below:

Definition 1. *Let* $\boldsymbol{X}_i : p \times 1$ *be a random vector such that* $E[\boldsymbol{X}_i] = \boldsymbol{\mu}$ *and* $E\left[(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})'\right] = \boldsymbol{\Sigma}_{p\times p}$. *Then we make the following definitions:*

$$D_{ii} = p^{-1}(\mathbf{X}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}). \tag{1}$$

$$D_{ij} = p^{-1}(\mathbf{X}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_j - \boldsymbol{\mu}). \tag{2}$$

$$\dot{D}_{ij} = p^{-1}(\mathbf{X}_i - \mathbf{X}_j)'\boldsymbol{\Sigma}^{-1}(\mathbf{X}_i - \mathbf{X}_j). \tag{3}$$

The D_{ii} statistic measures the scaled distance between an individual observation $\mathbf{X}_i$ and its expected value $\boldsymbol{\mu}$, which is frequently used to display data, assess distributional properties, and detect influential values. The D_{ij} measures the distance between two scaled and centered observations. It is used in cluster analysis and also to calculate the Mahalanobis angle between $\mathbf{X}_i$ and $\mathbf{X}_j$ subtended at $\boldsymbol{\mu}$, defined by $cos\theta(\mathbf{X}_i, \mathbf{X}_j) = D_{ij}/\sqrt{D_{ii}D_{jj}}$. The third statistic, $\dot{D}_{ij}$, is related to D_{ij} but centers the observation $\mathbf{X}_i$ about another independent observation $\mathbf{X}_j$ and is thereby independent of an estimate of $\boldsymbol{\mu}$. Estimators of (1) – (3) may be obtained by simply replacing the unknown parameters with appropriate estimators.

If both $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are unknown and replaced by the standard estimators, we get the well-known estimators defined below:

Definition 2. *Let* $\bar{\mathbf{X}} = n^{-1}\mathbf{X}'\mathbf{1}$ *and* $\mathbf{S} = n^{-1}\sum_{i=1}^{n}\left(\mathbf{X}-\bar{\mathbf{X}}\right)\left(\mathbf{X}-\bar{\mathbf{X}}\right)'$ *. Then we make the following definitions:*

$$d_{ii} = p^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right). \tag{4}$$

$$d_{ij} = p^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_j-\bar{\mathbf{X}}\right). \tag{5}$$

$$\dot{d}_{ij} = p^{-1}\left(\mathbf{X}_i-\mathbf{X}_j\right)'\mathbf{S}^{-1}\left(\mathbf{X}_i-\mathbf{X}_j\right). \tag{6}$$

Further discussion of the statistics (4) – (6) are available in Mardia (1977) and Mardia, Kent and Bibby (1980). There are, however, other possible estimators that may be more useful. In particular, it will sometimes be desirable to circumvent the dependence between $\{\mathbf{X}_i, \mathbf{X}_j\}$ and $\left\{\mathbf{S}^{-1}, \bar{\mathbf{X}}\right\}$ in the MD. This is conveniently achieved by simply omitting $\{\mathbf{X}_i, \mathbf{X}_j\}$ from the calculations of $\left\{\mathbf{S}^{-1}, \bar{\mathbf{X}}\right\}$. Formally, it is done as follows:

Definition 3. *Let* $\mathbf{S}_{(i)} = (n-1)^{-1}\sum_{k=1,k\neq i}^{n}\left(\mathbf{X}_k-\bar{\mathbf{X}}_{(i)}\right)\left(\mathbf{X}_k-\bar{\mathbf{X}}_{(i)}\right)'$, $\bar{\mathbf{X}}_{(i)} = (n-1)^{-1}\sum_{k=1,k\neq i}^{n}\mathbf{X}_k$, $\mathbf{S}_{(ij)} = (n-2)^{-1}\sum_{k=1,k\neq i,k\neq j}^{n}\left(\mathbf{X}_k-\bar{\mathbf{X}}_{(ij)}\right)\left(\mathbf{X}_k-\bar{\mathbf{X}}_{(ij)}\right)'$ *and* $\bar{\mathbf{X}}_{(ij)} = (n-2)^{-1}\sum_{k=1,k\neq i,k\neq j}^{n}\mathbf{X}_k$. *Then the following alternative estimators of (1) - (3) are defined:*

$$d_{(ii)} = p^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}_{(i)}\right)'\mathbf{S}_{(i)}^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}_{(i)}\right). \tag{7}$$

$$d_{(ij)} = p^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}_{(ij)}\right)'\mathbf{S}_{(ij)}^{-1}\left(\mathbf{X}_j-\bar{\mathbf{X}}_{(ij)}\right). \tag{8}$$

$$\dot{d}_{(ij)} = p^{-1}\left(\mathbf{X}_i-\mathbf{X}_j\right)'\mathbf{S}_{(ij)}^{-1}\left(\mathbf{X}_i-\mathbf{X}_j\right). \tag{9}$$

The estimators in (7) – (9) are frequently referred to as "leave-one-out" (Mardia, 1977) and "leave-two-out" estimators (De Maesschalck, Jouan-Rimbaud and

Massart, 2000). They have several advantages over (4) – (6). In particular, if there is a single outlier in the data set, it will not contaminate the sample mean value or covariance matrix. In addition, in case of independently identically normally distributed data, the sample mean vectors and the sample covariance matrix are independent; hence all components within the MD will be mutually independent, which in turn facilitates the derivation of their distributional properties. The estimators (7) – (9) thus provide interesting alternatives to the standard estimators (4) – (6) and are included for further investigation here.

Since the MDs have mainly been used in empirical works and very few theoretical properties are available, little is known about their behavior in high-dimensional settings – that is, in cases where the sample size n is proportional to the dimension of the random vector p. The reminder of this section is concerned with the expected values of the statistics (1) – (9) in such settings.

Proposition 1. *Let* $X_i \overset{iid}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ $i = 1, 2, ..., n$ *where* $\sup_j \{\lambda_j\}_{j=1}^p \leq r < \infty$ *and* $\inf_j \{\lambda_j\}_{j=1}^p \geq r' > 0$ $j = 1, 2, ..., p$, *where* λ_i *are the eigenvalues of* $\boldsymbol{\Sigma}_{p\times p}$. *Then the following holds:*

$$
\begin{aligned}
&(a)\ E[D_{ii}] = 1. \\
&(b)\ E[D_{ij}] = 0. \\
&(c)\ E\left[\dot{D}_{ij}\right] = 2. \\
&(d)\ E[d_{ii}] = 1. \\
&(e)\ E[d_{ij}] = -\frac{1}{n-1}. \\
&(f)\ E\left[\dot{d}_{ij}\right] = \frac{2n}{n-1}. \\
&(g)\ E\left[d_{(ii)}\right] = \frac{n}{(n-p-3)}. \\
&(h)\ E\left[d_{(ij)}\right] = \frac{1}{(n-p-4)}. \\
&(i)\ E\left[\dot{d}_{(ij)}\right] = \frac{2(n-2)}{(n-p-4)}.
\end{aligned}
$$

Proof: (a) – (c) and (g) – (i) are given in Appendix A, while (d) – (f) are given in Mardia (1977).

We are interested in estimating D_{ii}, D_{ij} and $\dot{D}_{ij}$ themselves rather than their expected values. It is, however, important to develop MD estimators that are centered about the same point as the corresponding true MD – that is, they should have the same (asymptotic) expected values. Possible biases may lead to flawed analyzes in, for instance, cluster analysis and outlier detection analysis. The differences between the expected values are listed below.

Proposition 2. *Let* $\boldsymbol{X}_i \overset{iid}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, *where* $\sup_j \{\lambda_j\}_{j=1}^p \leq r < \infty$, $\inf_j \{\lambda_j\}_{j=1}^p \geq r' > 0$, *where* λ_i *are the eigenvalues of* $\boldsymbol{\Sigma}_{p\times p}$. *Moreover, suppose we have a sample* $\{\mathbf{X_1}, ..., \mathbf{X_n}\}$ *such that* $p/n \to c$ *where* $0 \leq c < 1$, *and let* $\nu \in \mathbb{R}$ *be a non-random finite constant. Then,*

$$
\begin{aligned}
&(a) \quad E\left[\nu d_{ii} - D_{ii}\right] = \nu - 1. \\
&(b) \quad \lim_{n,p\to\infty} E\left[\nu d_{(ii)} - D_{ii}\right] = \frac{\nu}{1-c} - 1. \\
&(c) \quad \lim_{n,p\to\infty} E\left[\nu d_{ij} - D_{ij}\right] = 0. \\
&(d) \quad \lim_{n,p\to\infty} E\left[\nu d_{(ij)} - D_{ij}\right] = 0. \\
&(e) \quad \lim_{n,p\to\infty} E\left[\nu \dot{d}_{ij} - \dot{D}_{ij}\right] = 2(\nu - 1). \\
&(f) \quad \lim_{n,p\to\infty} E\left[\nu \dot{d}_{(ij)} - \dot{D}_{ij}\right] = 2\left(\frac{\nu}{1-c} - 1\right).
\end{aligned}
$$

Proof: Follows directly from Proposition 1 by inserting the expected values and taking limits.

Several interesting observations can be made based on Proposition 2. In particular, the constant ν may be set either to 1, which corresponds to estimating the MD using the biased inverse covariance matrix estimator defined by $\mathbf{S}^{-1}$, or alternatively to $\nu = (n-p-1)/n$, which yields the well-known unbiased estimator $(n-p-1)/n\mathbf{S}^{-1}$ (e.g., Mardia, Kent and Bibby (1980)). Appropriate degrees-of-freedom adjustments should be made for the leave-one-out estimator $d_{(ii)}$ and for the leave-two-out estimator $d_{(ij)}$. Note, however, that no single choice of ν yields asymptotically unbiased estimators for all six MD estimators simultaneously; it must be set individually for each estimator. For example, if $\nu = 1$, then according to Propositions 2 (a) and 2 (b), $E\left[1 \cdot d_{(ii)}\right] = E\left[D_{ii}\right]$, but $E\left[1 \cdot d_{ii}\right] \neq E\left[D_{ii}\right]$; in other words, $1 \cdot d_{ii}$ is an unbiased estimator of D_{ii}, whereas $1 \cdot d_{(ii)}$ is not. On the other hand, setting $\nu = (n-p-1)/n$ reverses the situation. Similar conclusions

can be drawn from Proposition 2 (c) – (f). Using the unbiased estimator of $\mathbf{\Sigma}^{-1}$ within the MD estimator yields biased MD estimates in some cases and unbiased ones in others. This is largely an effect of allowing for increasing-dimension asymptotics, since for a fixed p the constant $\nu = (n-p-1)/n$ will limit 1 as $n \to \infty$, and hence the choice between $\mathbf{S}^{-1}$ and $(n-p-1)/n\mathbf{S}^{-1}$ is immaterial. But when $c \not\to 0$ the choice of ν becomes crucial. To clarify things, we list the appropriate asymptotically unbiased MD estimates in Corollary 1.

Corollary 1. *Let $\mathbf{X}_1, ..., \mathbf{X}_n$ be distributed as in Proposition 2. Then the following properties hold:*

$$(a) \quad E\left[d_{ii} - D_{ii}\right] = 0. \tag{10}$$

$$(b) \quad \lim_{n,p\to\infty} E\left[(1-c)\, d_{(ii)} - D_{ii}\right] = 0. \tag{11}$$

$$(c) \quad \lim_{n,p\to\infty} E\left[d_{ij} - D_{ij}\right] = 0. \tag{12}$$

$$(d) \quad \lim_{n,p\to\infty} E\left[d_{(ij)} - D_{ij}\right] = 0. \tag{13}$$

$$(e) \quad \lim_{n,p\to\infty} E\left[\dot{d}_{ij} - \dot{D}_{ij}\right] = 0. \tag{14}$$

$$(f) \quad \lim_{n,p\to\infty} E\left[(1-c)\, \dot{d}_{(ij)} - \dot{D}_{ij}\right] = 0. \tag{15}$$

Proof: Follows from Proposition 2.

We conclude from Corollary 1 that some MD estimators should involve $\mathbf{S}^{-1}$, while others should involve $((n-p-1)/n)\,\mathbf{S}^{-1}$ in order to being unbiased. For example, according to Corollary 1 (e) and (f), $\dot{d}_{ij}$ is asymptotically unbiased, while $\dot{d}_{(ij)}$ has to be pre-multiplied by $(1-c)$ to be unbiased, which is somewhat unexpected. Also note that the bias resulting from using an inappropriate estimator may be substantial if c is close to 1.

The first-order moments derived above give some insight into the high-dimensional properties of the MD estimates, but the expected values are obviously not sufficient to describe the general usefulness of an estimator. Second-order moments are important, too. In particular, there is the issue of how to develop improved MD estimators that remain well-behaved in high-dimensional asymptotics. Considering the fact that the standard MD estimator (2.4) depends on two unknown parameters which need to be estimated, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, the question arises which of the two is more important with respect to the MD estimate. Bai, Liu and Wong (2009) argue that "although the sample covariance $\mathbf{S}$ is not a good estimator of the true $\boldsymbol{\Sigma}$ when the dimension is large, the sample mean $\bar{\mathbf{X}}$ is still a good estimator of $\boldsymbol{\mu}$." One may hence expect that improved MD estimators should place focus on the (inverse) covariance matrix rather than on the mean vector. This conjecture is supported by the correlations in Proposition 3, which shows that the asymptotic correlation between the true MD and the estimated MD with $\boldsymbol{\mu}$ replaced by $\bar{\mathbf{X}}$ equals 1, whereas the correlation between the true MD and that estimated with known $\boldsymbol{\mu}$ and with $\boldsymbol{\Sigma}^{-1}$ replaced by $\mathbf{S}^{-1}$ will in turn not limit 1 unless $(p/n) \to 0$.

Proposition 3. *Let* $\mathbf{X}_1, ..., \mathbf{X}_n$ *be distributed as in Proposition 2, and let* $k = n^{-1}(n-1)$, $\mathbf{S}_{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^{n}(\mathbf{X}_i - \boldsymbol{\mu})(\mathbf{X}_i - \boldsymbol{\mu})'$ *and* $d_{i(\boldsymbol{\mu},\mathbf{S}_{\boldsymbol{\mu}})} = (\mathbf{X}_i - \boldsymbol{\mu})'\mathbf{S}_{\boldsymbol{\mu}}{}^{-1}(\mathbf{X}_i - \boldsymbol{\mu}) = \mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1}\mathbf{Z}_i$. *Then the following properties hold:*

$$(a)\ Corr\left[D_{ii}, d_{i(\boldsymbol{\mu},\mathbf{S}_{\boldsymbol{\mu}})}\right] = \frac{\left\{2p\frac{(n-p)}{n+2}\right\}}{\left\{\sqrt{2p}\sqrt{2p\frac{(n-p)}{n+2}}\right\}} = \sqrt{\frac{(n-p)}{n+2}} \underset{n,p\to\infty}{\to} \sqrt{1-c}.$$

$$(b)\ Corr\left[D_{ii}, d_{i(\bar{\mathbf{X}},\boldsymbol{\Sigma})}\right] = \frac{2p\left(\frac{n-1}{n}\right)^2}{\sqrt{2p}\sqrt{2p\left(\frac{n-1}{n}\right)^2}} = \left(\frac{n-1}{n}\right) \underset{n,p\to\infty}{\to} 1.$$

$$(c)\ Var\left[d_{i(\bar{\mathbf{X}},\mathbf{S})}\right] = \frac{2p\,(n-p-1)}{(n+1)} \underset{n,p\to\infty}{\to} 2p\,(1-c)\,.$$

Proof: First, we have that $E\left[D_i \cdot d_{i(\boldsymbol{\mu},\mathbf{S}_{\boldsymbol{\mu}})}\right] = E\left[tr\left(\mathbf{Z}_i\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1}\mathbf{Z}_i\mathbf{Z}'_i\right)\right]$, where $\mathbf{Z}_i \sim N(\mathbf{0},\mathbf{I})$ and $n\mathbf{S}_{\boldsymbol{\mu}} \sim W(n,\mathbf{I})$. Write $n\mathbf{S}_{\boldsymbol{\mu}} = n\left(\sum_{j\neq i}^{n}\mathbf{Z}_j\mathbf{Z}'_j + \mathbf{Z}_i\mathbf{Z}'_i\right) = n\left(\mathbf{S}_{\mu,(i)} + \mathbf{Z}_i\mathbf{Z}'_i\right)$. Then $\left\{\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1}\mathbf{Z}_i\right\}$ and $\{\mathbf{S}_{\boldsymbol{\mu}}\}$ are independent (Srivastava and Khatri, 1979). Hence,

$$
\begin{aligned}
& E\left[tr\left(\mathbf{Z}_i\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1}\mathbf{Z}_i\mathbf{Z}'_i\right)\right] \\
&= E\left[tr\left(\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\mathbf{Z}_i\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\mathbf{Z}_i\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\cdot\mathbf{S}_{\boldsymbol{\mu}}\right)\right] \\
&= E\left[tr\left(\left(\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\mathbf{Z}_i\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\right)^2\mathbf{S}_{\boldsymbol{\mu}}\right)\right] \\
&= tr\left(E\left[\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\mathbf{Z}_i\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1/2}\right]^2 E\left[\mathbf{S}_{\boldsymbol{\mu}}\right]\right) \\
&= nE\left[\left(\mathbf{Z}'_i\mathbf{S}_{\boldsymbol{\mu}}^{-1}\mathbf{Z}_i\right)^2\right].
\end{aligned}
$$

But $n\mathbf{S}_{\mu} = \left(\sum_{j\neq i}^{n}\mathbf{Z}_j\mathbf{Z}'_j + \mathbf{Z}_i\mathbf{Z}'_i\right) \sim \left(\mathbf{W}_{(i)} + \mathbf{W}_i\right)$, where $\mathbf{W}_{(i)} \sim Wishart(n-1,\mathbf{I})$, $\mathbf{W}_i \sim Wishart(1,\mathbf{I})$, and the two terms are independent. We may then use the identity $1 - \mathbf{Z}'_i\mathbf{S}^{-1}\mathbf{Z}_i = \dfrac{\left|\mathbf{W}_{(i)}\right|}{\left|\mathbf{W}_{(i)} + \mathbf{W}_i\right|} = \dfrac{\left|\mathbf{S}_{(i)}\right|}{|\mathbf{S}|}$, and it follows from Rao (2009) that $\mathbf{Z}'_i n\mathbf{S}^{-1}\mathbf{Z}_i \sim B\left(\dfrac{p}{2}, \dfrac{(n-1)-p+1}{2}\right)$, where $B(\alpha,\beta)$ is a Type I Beta distribution (Johnson, Kotz and Balakrishnan, 1995), and hence $E[\mathbf{Z}'_i\mathbf{S}^{-1}\mathbf{Z}_i]^2 = p(p+2)\left(\dfrac{n}{n+2}\right)$, and the expected value and variance are obtained similarly from the Beta distribution, which proves (a). Next, we note that

$$
\begin{aligned}
& E\left[D_i \cdot d_{i(\bar{\mathbf{X}},\boldsymbol{\Sigma})}\right] \\
&= E\left[\mathbf{Z}'_i\mathbf{Z}_i\left(\mathbf{Z}_i - \bar{\mathbf{Z}}\right)'\left(\mathbf{Z}_i - \bar{\mathbf{Z}}\right)\right] \\
&= E\left[\mathbf{Z}'_1\mathbf{Z}_1\mathbf{Z}'_1\mathbf{Z}_1 - 2\bar{\mathbf{Z}}'\mathbf{Z}_1\mathbf{Z}'_1\mathbf{Z}_1 + \mathbf{Z}'_1\mathbf{Z}_1\bar{\mathbf{Z}}'\bar{\mathbf{Z}}\right]
\end{aligned}
$$

$$
\begin{aligned}
&= E\left[\chi^2_{(p)}\right]^2 - 2n^{-1}E\left[\mathbf{Z'}_1\mathbf{Z}_1\mathbf{Z'}_1\mathbf{Z}_1 + \left(\sum\nolimits_{i\neq 1}^{n}\mathbf{Z}_i\right)'\mathbf{Z}_1\mathbf{Z'}_1\mathbf{Z}_1\right] \\
&+ n^{-2}E\left[\mathbf{Z'}_1\mathbf{Z}_1\mathbf{Z'}_1\mathbf{Z}_1 + (n-1)\,E(\mathbf{Z'}_1\mathbf{Z}_1\mathbf{Z'}_i\mathbf{Z}_i)_{i\neq 1}\right] \\
&= \left(p^2+2p\right)\left(1-2n^{-1}\right) + n^{-2}\left(p^2+2p\right) + n^{-2}\,(n-1)\,p^2 \\
&= 2pk^2 + p^2k.
\end{aligned}
$$

Next, $E\left[d_{i(\bar{\mathbf{X}},\boldsymbol{\Sigma})}\right]$ and $Var\left[d_{i(\bar{\mathbf{X}},\boldsymbol{\Sigma})}\right]$ may be obtained from the property $d_{i(\bar{\mathbf{X}},\boldsymbol{\Sigma})} = \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right) \sim \left(\mathbf{Z}_i - \bar{\mathbf{Z}}\right)'\left(\mathbf{Z}_i - \bar{\mathbf{Z}}\right)$, where $\left(\mathbf{Z}_i - \bar{\mathbf{Z}}_{(i)}\right) \sim N\left(\mathbf{0}, k\mathbf{I}\right)$, where k is defined above, and $E\left[D_i\right]$ and $Var\left[D_i\right]$ are obtained from $D_i \sim \chi^2_{(p)}$, which proves (b). Finally, the moments of d_i are given in Mardia (1977), which yields (c).

Several important conclusions can be made from Proposition 3. First, in comparison with (a), where the mean value but not the covariance matrix is known, and with (b), which reverses the situation, the correlation between the true and the estimated MD does not limit 1 in (a), but it does limit 1 in (b). This verifies the conjecture that the covariance matrix is a concern in estimating the MD in high-dimensional data, whereas the mean vector is not. Hence improved estimators should involve alternative estimators of the (inverse) covariance matrix rather than the sample mean vector. Second, the variance of the standard estimator d_i does not limit the variance of the true MD, $Var\left[D_i\right] = 2p$, and hence the range of d_i will typically not correspond to that of D_i , meaning that they have fundamentally different limiting distributions. In the next section we develop an appropriate risk function for the MD that may be used to develop estimators with certain optimality properties. This includes estimators of a wider class than those considered above.

3. A new class of MD estimators

In this section we consider a wider family of estimators, of which those considered in Section 2 are special cases. Of particular interest is a family of estimators proposed by Efron and Morris (1976), defined by

$$\hat{\mathbf{\Sigma}}_{a,b}^{-1} = a\mathbf{S}^{-1} + b\big(p^{-1}tr\mathbf{S}\big)^{-1}\mathbf{I}, \tag{16}$$

where a and b are non-random constants (the original estimator by Efron and Morris (1976) is defined slightly differently, but the format (16) is more convenient for our purposes). This estimator is interesting from several points of view. A typical eigenvalue of $\hat{\mathbf{\Sigma}}_{a,b}^{-1}$ is given by $al_j^{-1} + b\big(\bar{l}\big)^{-1}$, where l_j is an eigenvalue of $\mathbf{S}$ and $\bar{l}$ its arithmetic mean. Hence, if $a+b=1$, the eigenvalues of $\hat{\mathbf{\Sigma}}_{a,b}^{-1}$ are smoothed towards $\bar{l}^{-1}$. In this sense, the estimator is a regularized estimator and may hence be expected to perform better than the standard estimator $\mathbf{S}^{-1}$ in high-dimensional settings. It may also be shown that $p^{-1}tr\left(\hat{\mathbf{\Sigma}}_{a,b}^{-1}\right)$ is algebraically smaller than the "overestimated" $p^{-1}tr\left(\mathbf{S}^{-1}\right)$ and hence stochastically closer to $p^{-1}tr\left(\mathbf{\Sigma}^{-1}\right)$ (proof omitted). Moreover, unlike other non-linearly regularized estimators, such as resolvent estimators (Serdobolskii, 2007) whose expected values are unknown, $\hat{\mathbf{\Sigma}}_{a,b}^{-1}$ has a closed, simple form which in turn facilitates its use within MD estimates. Using the estimator (16) within the MD, we define the following normalized estimators of D_{ii}, D_{ij} and $\dot{D}_{ij}$:

Definition 4. *Let* $\bar{\mathbf{X}} = n^{-1}\mathbf{X}'\mathbf{1}$, $\mathbf{S} = n^{-1}\sum_{i=1}^{n}\left(\mathbf{X}-\bar{\mathbf{X}}\right)\left(\mathbf{X}-\bar{\mathbf{X}}\right)'$ *and* $\hat{\mathbf{\Sigma}}_{a,b}^{-1} = a\mathbf{S}^{-1} + b(p^{-1}tr\mathbf{S})^{-1}\mathbf{I}$, *where* a *and* b *are non-random constants. Then we make the following definitions:*

$$e_{ii} = p^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\hat{\mathbf{\Sigma}}_{a,b}^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right). \tag{17}$$

$$e_{ij} = p^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\hat{\mathbf{\Sigma}}_{a,b}^{-1}\left(\mathbf{X}_j - \bar{\mathbf{X}}\right). \tag{18}$$

$$\dot{e}_{ij} = p^{-1}\left(\mathbf{X}_i - \mathbf{X}_j\right)'\hat{\mathbf{\Sigma}}_{a,b}^{-1}\left(\mathbf{X}_i - \mathbf{X}_j\right). \tag{19}$$

The expected values of these estimators are given in Proposition 4.

Proposition 4. *Let* $\mathbf{X}_i \overset{iid}{\sim} N\left(\boldsymbol{\mu}_{p\times 1}, \boldsymbol{\Sigma}_{p\times p}\right)$ $i = 1, 2, ..., n$, *where* $\sup_j \{\lambda_j\}_{j=1}^p \leq r < \infty$, $\inf_j \{\lambda_j\}_{j=1}^p \geq r' > 0$ $j = 1, 2, ..., p$, *where* λ_i *are the eigenvalues of* $\boldsymbol{\Sigma}_{p\times p}$ *and* a *and* b *are fixed non-random constants, and* $\hat{\boldsymbol{\Sigma}}_{a,b}^{-1} = a\mathbf{S}^{-1} + b(p^{-1}tr\mathbf{S})^{-1}\mathbf{I}$, *where* $\mathbf{S} = n^{-1}\sum_{i=1}^n \left(\mathbf{X} - \bar{\mathbf{X}}\right)\left(\mathbf{X} - \bar{\mathbf{X}}\right)'$. *Then the following holds:*

$$
\begin{aligned}
&(a) \quad E\left[e_{ii}\right] = a + b. \\
&(b) \quad E\left[e_{ij}\right] = -\frac{a+b}{n-1}. \\
&(c) \quad E\left[\dot{e}_{ij}\right] = -\frac{2n\left(a+b\right)}{n-1}.
\end{aligned}
$$

Proof:

(a) Since the variables $\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\left\{\mathbf{I}(p^{-1}tr\left(\mathbf{S}\right))^{-1}\right\}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)$ are identically distributed and $\sum_{i=1}^n \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\left\{\mathbf{I}(p^{-1}tr\left(\mathbf{S}\right))^{-1}\right\}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right) = (p^{-1}tr\left(\mathbf{S}\right))^{-1}tr\left(n\mathbf{S}\right) = np$, it follows that

$$
\begin{aligned}
&E\left[e_{ii}\right] \\
&= E\left[p^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'a\mathbf{S}^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\right] + p^{-1}E\left[\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\left\{b\mathbf{I}\left(p^{-1}tr\left(\mathbf{S}\right)\right)^{-1}\right\}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\right] \\
&= a + b,
\end{aligned}
$$

where the first term is given by Proposition 1(d). ■

(b) The terms $\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)$ are identically distributed, and hence

$$\begin{aligned}
E\left[e_{ij}\right] &= E\left[p^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right] \\
&= p^{-1}(n-1)^{-1} \sum\nolimits_{i \neq j}^{n} E\left[\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right] \\
&= p^{-1}(n-1)^{-1} E\left[\left(\sum\nolimits_{i \neq j}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)'\right) \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right] \\
&= p^{-1}(n-1)^{-1} E\left[\left(\sum\nolimits_{i=1}^{n} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' - \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)'\right) \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right] \\
&= p^{-1}(n-1)^{-1} E\left[-\left(\mathbf{X}_j - \bar{\mathbf{X}}\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right] = -(n-1)^{-1} E\left[e_{ii}\right] \\
&= -\frac{a+b}{n-1}. \quad \blacksquare
\end{aligned}$$

(c) It holds that

$$\begin{aligned}
E\left[\dot{e}_{ij}\right] &= E\left[p^{-1}(\mathbf{X}_i - \mathbf{X}_j)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} (\mathbf{X}_i - \mathbf{X}_j)\right] \\
&= E\left[p^{-1}\left(\left(\mathbf{X}_i - \bar{\mathbf{X}}\right) - \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\left(\mathbf{X}_i - \bar{\mathbf{X}}\right) - \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right)\right] \\
&= E\left[p^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)\right] - 2E\left[p^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right] \\
&+ E\left[p^{-1}\left(\mathbf{X}_j - \bar{\mathbf{X}}\right)' \hat{\mathbf{\Sigma}}_{a,b}^{-1} \left(\mathbf{X}_j - \bar{\mathbf{X}}\right)\right].
\end{aligned}$$

From Proposition 3.2 (a) and (b) it then follows that

$$E\left[\dot{e}_{ij}\right] = (a+b) - 2\left(-\frac{a+b}{n-1}\right) + (a+b) = \frac{2n\,(a+b)}{n-1}. \quad \blacksquare$$

The expected values of Proposition 4 are interesting because they show that under the constraint $(a+b) = 1$, the expectations of the MD in (17) – (19) coincide with those of the traditional MD estimates in Proposition 1 (d) - (f).

The first-order moment properties are, however, not sufficient to describe the general usefulness of an estimator; second-order moments are also important.

Hence, in order to derive optimal values of a and b, an appropriate risk function to be optimized is required. Moreover, such a function must be derived with respect to the MD since an estimator optimal for estimating $\mathbf{\Sigma}^{-1}$ alone need not necessarily produce an optimal estimator of the MD. In order to develop an appropriate risk function, we start by considering the common MD estimate d_{ii} and note that the expected value of the distance between this MD estimate and the population counterpart D_{ii} may be written as

$$\begin{aligned}
&E\left(d_{ii}-D_{ii}\right)\\
&=n^{-1}\sum\nolimits_{i=1}^{n}E\left(d_{ii}-D_{ii}\right)\\
&=n^{-1}\sum\nolimits_{i=1}^{n}E\left\{\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\mathbf{S}^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)-\right.\\
&\quad\left.\left(\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)-\left(\mu-\bar{\mathbf{X}}\right)\right)'\mathbf{\Sigma}^{-1}\left(\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)-\left(\mu-\bar{\mathbf{X}}\right)\right)\right\}\\
&=n^{-1}\sum\nolimits_{i=1}^{n}E\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\left(\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\right)\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)+E\left(\bar{\mathbf{X}}-\mu\right)'\mathbf{\Sigma}^{-1}\left(\bar{\mathbf{X}}-\mu\right).
\end{aligned}$$

The second term of this expression is positive w.p.1; hence, squaring $\left(\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\right)$ yields a strictly positive measure. The risk function normalized by p^{-1} hence becomes

$$\begin{aligned}
R\left(\mathbf{S}^{-1}\right) &= p^{-1}n^{-1}E\left[\sum\nolimits_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\left(\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\right)^2\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\right]+\\
&\quad p^{-1}E\left[\left(\mu-\bar{\mathbf{X}}\right)'\mathbf{\Sigma}^{-1}\left(\mu-\bar{\mathbf{X}}\right)\right]\\
&\quad = p^{-1}n^{-1}E\left[tr\left\{\left(\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\right)^2\sum\nolimits_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\right\}\right]+\\
&\quad p^{-1}trE\left[\mathbf{\Sigma}^{-1}\left(\mu-\bar{\mathbf{X}}\right)\left(\mu-\bar{\mathbf{X}}\right)'\right]\\
&\quad = p^{-1}tr\left\{E\left[\left(\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\right)^2\mathbf{S}\right]\right\}+n^{-1}.
\end{aligned}$$

For an arbitrary estimator of the inverse covariance matrix, say $\hat{\mathbf{\Sigma}}^{-1}$, and ignoring the constant n^{-1} term, we arrive at the following risk function:

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

The risk measure (16) has been used earlier in Holgersson and Karlsson (2012) and also coincides with a risk measure previously developed by Efron and Morris

(1976), though the latter authors derived it from an empirical Bayes perspective and for the purpose of estimating $\boldsymbol{\Sigma}^{-1}$ alone. Moreover, it may be shown that any $\hat{\boldsymbol{\Sigma}}^{-1}$ which minimizes (16) will minimize not only the difference $(d_{ii} - D_{ii})$ but also $(d_{ij} - D_{ij})$ and $\left(\dot{d}_{ij} - \dot{D}_{ij}\right)$ (see Appendix B); hence there is no need to develop separate covariance matrix estimates for d_{ii}, d_{ij} and $\dot{d}_{ij}$ individually. Once an estimator of $\boldsymbol{\Sigma}^{-1}$ that minimizes $R\left(\hat{\boldsymbol{\Sigma}}^{-1}\right)$ is identified, it is simultaneously optimal for d_{ii}, d_{ij} and $\dot{d}_{ij}$. When it comes to developing an explicit estimator $\hat{\boldsymbol{\Sigma}}^{-1}_{a,b}$, we will start with the constrained estimator $\hat{\boldsymbol{\Sigma}}^{-1}_{a,b=0}$ and derive the optimal value of the scalar a within $a\mathbf{S}^{-1}$ through the risk function (17).

Proposition 5. *Let* $\mathbf{X}_1, ..., \mathbf{X}_n$ *be distributed as in Proposition 4, and let* $p/n \to c$, *where* $0 \leq c < 1$ $\mathbf{S} = n^{-1}\sum_{i=1}^{n}\left(\mathbf{X} - \bar{\mathbf{X}}\right)\left(\mathbf{X} - \bar{\mathbf{X}}\right)'$. *Then* $\lim_{n,p\to\infty} R\left(\hat{\boldsymbol{\Sigma}}^{-1}_{a,b=0}\right)$ *is minimized when* $a = 1 - c$.

Proof: The risk function is given by

$$\begin{aligned} R\left(\hat{\boldsymbol{\Sigma}}^{-1}_{a,b=0}\right) &= p^{-1}tr\left\{E\left[\left(a\mathbf{S}^{-1} - \boldsymbol{\Sigma}^{-1}\right)^2\mathbf{S}\right]\right\} \\ &= p^{-1}tr\left\{E\left[a^2\mathbf{S}^{-1} - 2a\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-2}\mathbf{S}\right]\right\} \\ &= p^{-1}tr\left\{a^2\left(\frac{n}{n-p-1}\right)\boldsymbol{\Sigma}^{-1} - 2a\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}^{-1}\right\} \\ &= p^{-1}\left(a^2\left(\frac{n}{n-p-1}\right) - 2a + 1\right)tr\left\{\boldsymbol{\Sigma}^{-1}\right\}. \end{aligned}$$

Taking the derivative of $R\left(a\mathbf{S}^{-1}\right)$ w.r.t. a, equating at zero and solving yields $a_{opt} = \dfrac{n-p-1}{n}$ and so $\lim_{n,p\to\infty} a_{opt} = 1 - c$. ■

In view of Corollary 1 (a) and (c), this is an interesting finding because the value of a yielding asymptotic unbiasedness for ad_{ii} and ad_{ij} is $a = 1$. Hence, the leave-one-out estimator $d_{(ii)}$ defined in (7) is the only unbiased estimator which

minimizes the risk function and is from this point of view the only reasonable estimator of D_{ii} to use in high-dimensional settings. Below the constraint in Proposition 5 is reversed, and the optimal value of b is derived when a is constrained to equal zero.

Proposition 6. *Let* $\mathbf{X}_1, ..., \mathbf{X}_n$ *be distributed as in Proposition 4, and let* $p/n \to c$, *where* $0 \leq c < 1$ *and* $\mathbf{S} = n^{-1} \sum_{i=1}^{n} \left(\mathbf{X} - \bar{\mathbf{X}}\right)\left(\mathbf{X} - \bar{\mathbf{X}}\right)'$. *Then* $\lim_{n,p\to\infty} R\left(\hat{\mathbf{\Sigma}}_{0,b}^{-1}\right)$ *is minimized when* b=1.

Proof: Using the identity $\frac{1}{\Lambda_{-1}} E\left[\frac{(np-2)}{ntr(\mathbf{S})}\right] = \frac{1}{\Lambda_{-1}} E\left[\frac{tr(\mathbf{\Sigma}^{-1}\mathbf{S})}{tr(\mathbf{S})}\right] = \varphi$ (Efron and Morris, 1976), where $\Lambda_{-1} = p^{-1}tr(\mathbf{\Sigma}^{-1})$, we get

$$
\begin{aligned}
R\left(\hat{\mathbf{\Sigma}}_{0,b}^{-1}\right) &= p^{-1}tr\left\{E\left[\left(\frac{b}{p^{-1}tr(\mathbf{S})}\mathbf{I} - \mathbf{\Sigma}^{-1}\right)^2 \mathbf{S}\right]\right\} \\
&= p^{-1}tr\left\{E\left[\frac{b^2}{p^{-2}tr^2(\mathbf{S})}\mathbf{S} - \frac{2b}{p^{-1}tr(\mathbf{S})}\mathbf{\Sigma}^{-1}\mathbf{S} + \mathbf{\Sigma}^{-2}\mathbf{S}\right]\right\} \\
&= \Lambda_{-1}\left\{E\left[\frac{b^2}{p^{-1}(pn-2)n^{-1}} \cdot \frac{(pn-2)}{\Lambda_{-1}ntr(\mathbf{S})} - 2b \cdot \frac{tr(\mathbf{\Sigma}^{-1}\mathbf{S})}{\Lambda_{-1}tr(\mathbf{S})} + \frac{p^{-1}tr(\mathbf{\Sigma}^{-1})}{\Lambda_{-1}}\right]\right\} \\
&= \Lambda_{-1}\left(\frac{b^2}{p^{-1}(pn-2)n^{-1}} \cdot E[\varphi] - 2b \cdot E[\varphi] + 1\right).
\end{aligned}
$$

Taking the derivative of $R\left(\hat{\mathbf{\Sigma}}_{0,b}^{-1}\right)$ w.r.t. b, equating at zero, and solving for b yields $2\Lambda_{-1}\left(\frac{b_{opt}}{p^{-1}(pn-2)n^{-1}} - 1\right) E[\varphi] = 0 \Rightarrow b_{opt} = 1 - 2n^{-1}p^{-1}$. As $n, p \to \infty$ we find that $b_{opt} \to 1$ ■.

It is also possible to find the optimal values of a and b without constraining one of them to equal zero. However, as shown in Efron and Morris (1976), the resulting optimal value of a, or equivalently of $1 - a$, will depend on unknown parameters. On the other hand, imposing the restriction that the MD estimator should be unbiased, equivalent to the constraint $a + b = 1$, an asymptotically optimal and

unbiased estimator is obtained by $\hat{\mathbf{\Sigma}}^{-1}_{(1-c),c}$. Propositions 5 and 6 then follow as special cases when $c = 1$ or $c = 0$. Unbiased and asymptotically optimal versions of the estimators (17) - (19) are hence available by substituting $a = 1 - c$ and $b = c$. Unlike the bias-adjusted estimators in Corollary 1, the new MD estimators defined by $e_{ii} = p^{-1}\big(\mathbf{X}_i - \bar{\mathbf{X}}\big)'\hat{\mathbf{\Sigma}}^{-1}_{1-c,c}\big(\mathbf{X}_i - \bar{\mathbf{X}}\big)$, $e_{ij} = p^{-1}\big(\mathbf{X}_i - \bar{\mathbf{X}}\big)'\hat{\mathbf{\Sigma}}^{-1}_{1-c,c}\big(\mathbf{X}_j - \bar{\mathbf{X}}\big)$ and $\dot{e}_{ij} = p^{-1}(\mathbf{X}_i - \mathbf{X}_j)'\hat{\mathbf{\Sigma}}^{-1}_{1-c,c}(\mathbf{X}_i - \mathbf{X}_j)$ are not only unbiased but also optimal with respect to the risk function (16), even as $p/n \to c$ when $0 \leq c < 1$. Moreover, these proposed estimators are simple to conduct and depend on no estimated parameters other than the sample mean $\bar{\mathbf{X}}$ and the covariance matrix $\mathbf{S}$, and since they coincide with the standard estimators in classical asymptotics – that is, when $c = 0$ – they should be useful in a wide range of applications. Moments and optimal values of a and b for the leave-one-out estimators could be derived in a similar way, but the process would be tedious and is hence omitted here.

4. Summary

In this paper the expected values of a number of individual Mahalanobis distances are derived in the case when the dimension p of the random vector increases proportionally to the sample size n. It is shown that some types of standard estimators remain unbiased in this case, while others are asymptotically biased, a finding which is somewhat unexpected. Moreover, a new family of MD estimates is proposed that utilizes an estimate of the inverse covariance matrix, which is an ingredient within the MD, previously proposed in the literature. Since this new family of estimators depends on two unknown constants a and b, their optimal values need to be derived. The paper therefore derives a risk function specifically designed for the MD. This risk function in turn coincides with a previously proposed risk function which conveniently links the properties of any inverse co-

variance matrix to individual Mahalanobis distance, in the sense that any inverse covariance matrix which is optimal with respect to minimizing the risk function will simultaneously produce optimal MD estimates. Moreover, using this risk function in conjunction with a first-order moment restriction facilitates derivation of closed-form optimal values for a and b. These optimal values are non-adaptive and non-random; hence, an operational, unbiased, and asymptotically efficient MD is available. It is argued that the proposed new family of MD estimators should be superior to the standard estimators in a wide range of settings involving low-dimensional as well as high-dimensional data.

Appendices

A.

Proof of Proposition 1 *(g):* Under the assumption $\mathbf{X}_i \sim iidN(\mu, \mathbf{\Sigma})$, we have that $(n-1)\mathbf{S}_{(i)} \sim W(n-2, \mathbf{\Sigma})$ so that $\mathbf{S}_{(i)}^{-1} \sim (n-1)^{-1} W^{-1}(n-2, \mathbf{\Sigma})$, where $W^{-1}(,)$ denotes the inverse Wishart distribution (see Siskind (1972), Mardia, Kent and Bibby (1980), von Rosen (1988)), and we get that $E\left[\mathbf{S}_{(i)}^{-1}\right] = \frac{(n-1)}{(n-p-3)}\mathbf{\Sigma}^{-1}$. Moreover, since $\left\{\mathbf{X}_i, \mathbf{S}_{(i)}, \bar{\mathbf{X}}_{(i)}\right\}$ are all mutually independent, we get

$$
\begin{aligned}
E\left[d_{(ii)}\right] &= p^{-1} E\left[\left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right)' \mathbf{S}_{(i)}^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right)\right] \\
&= p^{-1} E\left[tr\left(\mathbf{S}_{(i)}^{-1}\left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right)'\right)\right] \\
&= p^{-1} tr\left(E\left[\mathbf{S}_{(i)}^{-1}\right] E\left[\left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right)\left(\mathbf{X}_i - \bar{\mathbf{X}}_{(i)}\right)'\right]\right) \\
&= p^{-1} \frac{(n-1)}{(n-p-3)} tr E\left[\mathbf{\Sigma}^{-1}\left(\left(\mathbf{X}_i - \mu\right)\left(\mathbf{X}_i - \mu\right)' + \left(\bar{\mathbf{X}}_{(i)} - \mu\right)\left(\bar{\mathbf{X}}_{(i)} - \mu\right)'\right)\right] \\
&= p^{-1} \frac{(n-1)}{(n-p-3)} tr\left(\left[\mathbf{\Sigma}^{-1}\left(\mathbf{\Sigma} + (n-1)^{-1}\mathbf{\Sigma}\right)\right]\right) \\
&= \frac{n}{(n-p-3)}. \quad \blacksquare
\end{aligned}
$$

Proof of Proposition 1 *(h):* Noting that $E\left[\mathbf{S}_{(ij)}^{-1}\right] = \frac{(n-2)}{(n-p-4)}\mathbf{\Sigma}^{-1}$ and that $\left\{\mathbf{X}_i, \mathbf{X}_j, \mathbf{S}_{(ij)}, \bar{\mathbf{X}}_{(ij)}\right\}$ are all mutually independent, we get

$$
\begin{aligned}
E\left[d_{(ij)}\right] &= p^{-1}E\left[\left(\mathbf{X}_i-\bar{\mathbf{X}}_{(ij)}\right)'\mathbf{S}_{(ij)}^{-1}\left(\mathbf{X}_j-\bar{\mathbf{X}}_{(ij)}\right)\right] \\
&= p^{-1}E\left[tr\left(\mathbf{S}_{(i)}^{-1}\left(\mathbf{X}_i-\bar{\mathbf{X}}_{(i)}\right)\left(\mathbf{X}_i-\bar{\mathbf{X}}_{(i)}\right)'\right)\right] \\
&= p^{-1}\frac{(n-2)}{(n-p-4)}tr\left(E\left[\mathbf{\Sigma}^{-1}\left(\left(\bar{\mathbf{X}}_{(ij)}-\mu\right)\left(\bar{\mathbf{X}}_{(ij)}-\mu\right)'\right)\right]\right) \\
&= p^{-1}\frac{(n-2)}{(n-p-4)}tr\left[\mathbf{\Sigma}^{-1}\left((n-2)^{-1}\mathbf{\Sigma}\right)\right] \\
&= \frac{1}{(n-p-4)}. \quad \blacksquare
\end{aligned}
$$

Proof of Proposition 1 *(i):* Noting that $E\left[\mathbf{S}_{(ij)}^{-1}\right] = \frac{(n-2)}{(n-p-4)}\mathbf{\Sigma}^{-1}$ and that $\left\{\mathbf{X}_i, \mathbf{X}_j, \mathbf{S}_{(ij)}\right\}$ are mutually independent, we get

$$
\begin{aligned}
E\left[\dot{d}_{(ij)}\right] &= p^{-1}E\left[\left(\mathbf{X}_i-\mathbf{X}_j\right)'\mathbf{S}_{(ij)}^{-1}\left(\mathbf{X}_i-\mathbf{X}_j\right)\right] \\
&= p^{-1}\frac{(n-2)}{(n-p-4)}tr\left(E\left[\mathbf{\Sigma}^{-1}\left(\left(\mathbf{X}_i-\mu\right)\left(\mathbf{X}_i-\mu\right)'+\left(\mathbf{X}_j-\mu\right)\left(\mathbf{X}_j-\mu\right)'\right)\right]\right) \\
&= \frac{2\left(n-2\right)}{(n-p-4)}. \quad \blacksquare
\end{aligned}
$$

B.

Derivation of the risk function for the Mahalanobis distance d_{ij} :

Noting that there are *n-1* identically distributed d_{ij} terms for each i, where $i, j = 1, ..., n$, the average distance $(d_{ij} - D_{ij})$ equals the average of $(d_{i1} - D_{i1})$, say, which in turn may be written as

$$
\begin{aligned}
&(n-1)^{-1}\sum_{i\neq 1}^{n}\left(d_{i1}-D_{i1}\right)\\
&=(n-1)^{-1}\sum_{i\neq 1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)\left(\mathbf{X}_1-\bar{\mathbf{X}}\right)-\left(\mu-\bar{\mathbf{X}}\right)'\boldsymbol{\Sigma}^{-1}\left(\mu-\bar{\mathbf{X}}\right).
\end{aligned}
$$

By squaring $(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1})$, normalizing with p^{-1}, and taking expectation, we get

$$
\begin{aligned}
&p^{-1}(n-1)^{-1}E\left[\sum_{i\neq 1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)^2\left(\mathbf{X}_1-\bar{\mathbf{X}}\right)\right]+n^{-1}\\
&=p^{-1}(n-1)^{-1}E\left[\sum_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)^2\left(\mathbf{X}_1-\bar{\mathbf{X}}\right)\right.\\
&\left.-\left(\mathbf{X}_1-\bar{\mathbf{X}}\right)'\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)^2\left(\mathbf{X}_1-\bar{\mathbf{X}}\right)\right]+n^{-1}\\
&=p^{-1}(n-1)^{-1}E\left[0-\left(\mathbf{X}_1-\bar{\mathbf{X}}\right)'\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)^2\left(\mathbf{X}_1-\bar{\mathbf{X}}\right)\right]+n^{-1}\\
&=-p^{-1}(n-1)^{-1}E\left[tr\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)^2\sum_{i=1}^{n}\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)\left(\mathbf{X}_i-\bar{\mathbf{X}}\right)'\right]+n^{-1}\\
&=-p^{-1}n(n-1)^{-1}E\left[tr\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)^2\mathbf{S}\right]+n^{-1}.\quad\blacksquare
\end{aligned}
$$

Derivation of the risk function for the Mahalanobis distance $\dot{d}_{ij}$: : Allowing for permutations, there are $m=n\,(n-1)$ terms of $\dot{d}_{ij}-\dot{D}_{ij}$. Defining $\Delta:=(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1})$ and normalizing with p^{-1}, we have

$$
\begin{aligned}
&p^{-1}E\left[\dot{d}_{ij}-\dot{D}_{ij}\right]=\\
&p^{-1}m^{-1}E\left[\sum_{i=1}^{n}\sum_{j=1,i\neq j}^{n}\left(\left(\mathbf{X}_i-\mathbf{X}_j\right)'\left(\mathbf{S}^{-1}-\boldsymbol{\Sigma}^{-1}\right)^2\left(\mathbf{X}_i-\mathbf{X}_j\right)\right.\right.\\
&\left.\left.-p^{-1}\left(\mu-\bar{\mathbf{X}}\right)'\boldsymbol{\Sigma}^{-1}\left(\mu-\bar{\mathbf{X}}\right)\right)\right]\\
&=p^{-1}m^{-1}E\left[\left\{\sum_{i=1,i\neq 1}^{n}\left(\mathbf{X}_i-\mathbf{X}_1\right)'\Delta^2\left(\mathbf{X}_i-\mathbf{X}_1\right)\right\}\right.\\
&\left.+...+\left\{\sum_{i=1,i\neq n}^{n}\left(\mathbf{X}_i-\mathbf{X}_n\right)'\Delta^2\left(\mathbf{X}_i-\mathbf{X}_n\right)\right\}\right]-n^{-1}.
\end{aligned}
$$

An arbitrary {} term may be expanded as,

$$\begin{aligned}
&\sum_{i=1,i\neq a}^{n}\left(\mathbf{X}_i-\mathbf{X}_a\right)'\Delta^2\left(\mathbf{X}_i-\mathbf{X}_a\right)\\
&=\sum_{i=1}^{n}\left(\mathbf{X}_i-\mathbf{X}_a\right)'\Delta^2\left(\mathbf{X}_i-\mathbf{X}_a\right)-\left(\mathbf{X}_a-\mathbf{X}_a\right)'\left(\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\right)^2\left(\mathbf{X}_a-\mathbf{X}_a\right)\\
&=\sum_{i=1}^{n}\left(\mathbf{X}_i{}'\Delta^2\mathbf{X}_i-2\mathbf{X}'_i\Delta^2\mathbf{X}_1+\mathbf{X}_1{}'\Delta^2\mathbf{X}_1\right).
\end{aligned}$$

Substituting this into the above and omitting the last n^{-1} term, we get

$$\begin{aligned}
&p^{-1}E\left[\dot{d}_{ij}-\dot{D}_{ij}\right]\\
&=p^{-1}m^{-1}E\left[\left\{\sum_{i=1}^{n}\mathbf{X}_i{}'\Delta^2\mathbf{X}_{\ i}-2n\bar{\mathbf{X}}'\Delta^2\mathbf{X}_1+n\mathbf{X}_1{}'\Delta^2\mathbf{X}_1\right\}+\right.\\
&\left....+\left\{\sum_{i=1}^{n}\mathbf{X}_i{}'\Delta^2\mathbf{X}_{\ i}-2n\bar{\mathbf{X}}'\Delta^2\mathbf{X}_n+n\mathbf{X}_n{}'\Delta^2\mathbf{X}_n\right\}\right]\\
&=p^{-1}m^{-1}E\left[n\sum_{i=1}^{n}\mathbf{X}_i{}'\Delta^2\mathbf{X}_i-2n\bar{\mathbf{X}}'\Delta^2n\bar{\mathbf{X}}+n\sum_{i=1}^{n}\mathbf{X}_i{}'\Delta^2\mathbf{X}_i\right]\\
&=p^{-1}2m^{-1}nEtr\left[\Delta^2\left(\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i{}'-n\bar{\mathbf{X}}\bar{\mathbf{X}}'\right)\right]\\
&=p^{-1}2m^{-1}nE\left[tr\left(\Delta^2\mathbf{S}\right)\right]\\
&=2p^{-1}(n-1)^{-1}E\left[tr\left(\Delta^2\mathbf{S}\right)\right]\\
&\rightarrow 2n^{-1}p^{-1}E\left[tr\left(\left(\mathbf{S}^{-1}-\mathbf{\Sigma}^{-1}\right)^2\mathbf{S}\right)\right]\quad\blacksquare
\end{aligned}$$

Bai, Z., Liu, H. and Wong, W.-K. (2009). Enhancement of the applicability of markowitz's portfolio optimization by utilizing random matrix theory, *Mathematical Finance* **19**(4): 639–667.

De Maesschalck, R., Jouan-Rimbaud, D. and Massart, D. L. (2000). The mahalanobis distance, *Chemometrics and Intelligent Laboratory Systems* **50**(1): 1–18.

Efron, B. and Morris, C. (1976). Multivariate empirical bayes and estimation of covariance matrices, *The Annals of Statistics* pp. 22–32.

Fisher, R. A. (1940). The precision of discriminant functions, *Annals of Human Genetics* **10**(1): 422–429.

Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning*, Vol. 1, Springer Series in Statistics.

Girko, V. L. (1995). *Statistical analysis of observations of increasing dimension*, Vol. 28, Springer.

Holgersson, H. and Shukur, G. (2001). Some aspects of non-normality tests in systems of regression equations, *Communications in Statistics-Simulation and Computation* **30**(2): 291–310.

Holgersson, T. and Karlsson, P. (2012). Three estimators of the mahalanobis distance in high-dimensional data, *Journal of Applied Statistics* .

Johnson, N., Kotz, S. and Balakrishnan, N. (1995). *Continuous univariate distributions*, number çňň 2 åIJů in *Wiley series in probability and mathematical statistics: Applied probability and statistics*, Wiley & Sons.
URL: *http://books.google.se/books?id=0QzvAAAAMAAJ*

Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix, *Journal of Multivariate Analysis* **12**(1): 1–38.

Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices, *Journal of multivariate analysis* **88**(2): 365–411.

Mahalanobis, P. (1936). On the generalized distance in statistics, *Proceedings of the National Institute of Sciences of India*, Vol. 2, New Delhi, pp. 49–55.

Mardia, K. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies, *Sankhyā: The Indian Journal of Statistics, Series B* pp. 115–128.

Mardia, K. (1977). Mahalanobis distances and angles, *Multivariate analysis IV* pp. 495–511.

Mardia, K., Kent, J. and Bibby, J. (1980). Multivariate analysis.

Mitchell, A. and Krzanowski, W. (1985). The mahalanobis distance and elliptic distributions, *Biometrika* **72**(2): 464–467.

Rao, C. R. (1945). Familial correlations or the multivariate generalisations of the intraclass correlations, *Current Science* **14**(3): P66–67.

Rao, C. R. (2009). *Linear statistical inference and its applications*, Vol. 22, John Wiley & Sons.

Serdobolskii, V. (2000). *Multivariate statistical analysis: A high-dimensional approach*, Vol. 41, Springer.

Serdobolskii, V. I. (2007). *Multiparametric statistics*, Elsevier.

Siskind, V. (1972). Second moments of inverse elements wishart matrix, *Biometrika* pp. 690–691.

Srivastava, S. and Khatri, C. (1979). *An introduction to multivariate statistics*, North-Holland/New York.
URL: *http://books.google.se/books?id=swbvAAAAMAAJ*

von Rosen, D. (1988). Moments for the inverted wishart distribution, *Scandinavian Journal of Statistics* pp. 97–109.

Wilks, S. (1963). Multivariate statistical outliers, *Sankhyā: The Indian Journal of Statistics, Series A* pp. 407–426.