A smoothed estimator of the Mahalanobis distance in High-dimensional Data

Deliang Dai

School of Business and Economics,
Linnaeus University
deliang.dai@lnu.se

Abstract

The Mahalanobis distance (MD) is a fundamental statistic of distance measure which is frequently used as an ingredient within several statistical methods. However, it becomes instable in cases of "High-dimensional data". When the dimension p *is proportional to the sample size* n*, such that* $p/n \to c$ *where* $0 < c < 1$*.*

1. Introduction

The Mahalanobis distance is used to measure the dissimilarities between different variables, as demonstrated below.

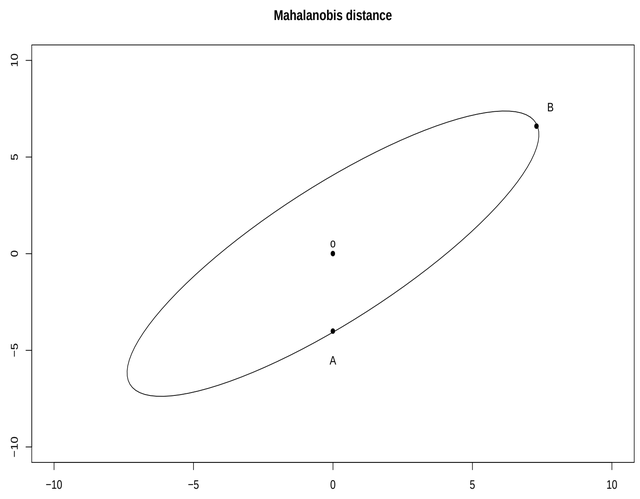

Figure 1: *The points A and B in figure have the same MDs to the original point o while the Euclidean distance is not.*

2. Basic definitions and Theories

The definition of Mahalanobis distance between an individual and its mean value is given below:

$$d_i^2 = p^{-1}\mathbf{Y}_i{}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}_i,$$

where $\mathbf{Y}_i = \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)$, $\bar{\mathbf{X}} = E\left[\mathbf{X}\right]$ and $\boldsymbol{\Sigma}_{(p\times p)} = E\left[\mathbf{Y}_i\mathbf{Y}'_i\right]$, $i = 1, ..., n$.

Since d_i^2 depends on the unknown $\boldsymbol{\Sigma}^{-\mathbf{1}}$, this is usually replaced by $\mathbf{S}^{-\mathbf{1}}$. However, the estimator $\hat{d}_i^2 = p^{-1}\mathbf{Y}_i{}'\mathbf{S}^{-1}\mathbf{Y}_i$ is not well behaved when $n \to \infty$, $p \to \infty$, $p/n \to c$ where $0 < c < 1$.

This is so because, the sample covariance matrix (**S**) on which the Mahalanobis distance depends becomes degenerate under IDA settings, which in turn produce stochastically unstable Mahalanobis distances. Therefore, some new estimators which avoid the instability problem should be developed.

For a general estimator of the covariance matrix $\hat{\boldsymbol{\Sigma}}^{-\mathbf{1}}$, the distance between the estimator and true covariance matrix is:

$$\begin{aligned}&\hat{d}_i^2 - d_i^2\\ &= p^{-1}\mathbf{Y}'_i\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{Y}_i - p^{-1}\mathbf{Y}'_i\boldsymbol{\Sigma}^{-1}\mathbf{Y}_i\\ &= p^{-1}\mathbf{Y}'_i\left(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right)\mathbf{Y}_i\end{aligned}$$

A quadratic loss function, may be defined by,

$$R^2\left(\hat{\boldsymbol{\Sigma}}^{-1}\right) = p^{-1}E\left[\mathbf{Y}'_i\left(\hat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1}\right)^2\mathbf{Y}_i\right].$$

The risk function for the average Mahalanobis distance is,

$$\begin{aligned}R^2\left(\Sigma^{-\mathbf{1}}\right) &= (np)^{-1}\sum_{i=1}^{n}\mathbf{Y}'_i\left(\Sigma^{-\mathbf{1}} - \boldsymbol{\Sigma}^{-1}\right)^2\mathbf{Y}_i\\ &= p^{-1}E\left[tr\left(\left(\Sigma^{-\mathbf{1}} - \boldsymbol{\Sigma}^{-1}\right)^2\mathbf{S}\right)\right].\end{aligned}$$

We will consider 2 approaches to find appropriate estimators for d_i^2: (a) estimators that minimize R^2 without constraints and (b) estimators that yield positive values of $\Delta = R^2(c\mathbf{S}^{-\mathbf{1}}) - R^2(\hat{\boldsymbol{\Sigma}}^{-1})$.

Some different estimators of Σ^{-1} were proposed by (James and Stein; 1961) Efron and Morris (1976)and (Haff; 1977). The improved estimator of Σ^{-1} are often of the kind aS^{-1} where $S^{-1} = \frac{1}{n}(X - \hat{X})(X - \hat{X})'$.

This family, however, is not useful when $\frac{p}{n} \to c \neq 0$ because,

- when $a = 1$, it degenerates to the standard estimator $\Sigma^{-1} = S^{-1}$.
- when a is the unbiasing constant, the $\dfrac{n}{n-p-1}$ Mahalanobis distance limits $\dfrac{1}{1-c} \to \infty$ as $c \to 1$.
- Serdobolskii (2000) proposed an optimal estimator of the scalar α,

$$\alpha^{opt} = 1 - y - y\frac{(1-y)^{-2}\Lambda_{-1}^2}{(1-y)^{-2}\Lambda_{-2} + y(1-y)^{-3}\Lambda_{-1}^2}$$

  Therefore, as $y \to 0$, the $\alpha^{opt} \to 1$, then MD of sample inverse covariance matrix doesn't change. On the other hand, as $y \to 1$, the $\alpha^{opt} \to 0$, the MD of sample inverse covariance matrix becomes useless. Therefore, it is necessary to focus on other estimators.

An alternative is available through regularized estimators,

Serdobolskii (2000) and Girko (1995) proposed one resolvent type of estimator. The basic definition is given below,

$$\hat{\boldsymbol{\Sigma}}^{-1} = (I + tS)^{-1}$$

(Holgersson and Karlsson; 2012) also have some improvement basing on another form of resolvent estimator as below,

$$\hat{\boldsymbol{\Sigma}}^{-1} = (tI + S)^{-1}$$

These estimators reduce the variance dramatically.(Friedman, Hastie and Tibshirani; 2001) The "biased-variance trade-off" gives a reasonable compromise between the estimation and error. It is pointed out (Serdobolskii; 2007) that, by choosing the coefficient value of t empirically, we can get an estimator of Σ^{-1} which remains stable in IDA settings..

Serdobolskii (2007) proposed an asymptotically unimprovable estimator $\hat{\Sigma}^{-1} = \hat{\Gamma}_{n\varepsilon}^{opt}$ based on the eigenvalues as below:

$$\hat{\Gamma}_{n\varepsilon}^{opt}(\lambda_i) = \left(1 - \frac{n}{N}\right)\frac{\lambda_i}{\lambda_i^2 + \varepsilon^2} + \frac{2}{N}\sum_{j=1}^{n}\frac{\lambda_i - \lambda_j}{\left(\lambda_i - \lambda_j\right)^2 + \varepsilon^2}$$

where $\varepsilon^2 = \bar{\lambda}(p/n)^2$.

3. Simulation results

Some simulations based on the optimized estimator $\hat{\Sigma}^{-1} = \hat{\Gamma}_{n\varepsilon}^{opt}$ are given in the figures below.

The risk ratio equals $R^2(\hat{\Sigma}^{-1})/R^2(S^{-1})$. Therefore, the value near to 1 indicates that the smoothed is as good as the regular estimator while a value less then one indicate that it has a lower risk and hence is better.

The results of simulations are given in the figures below,

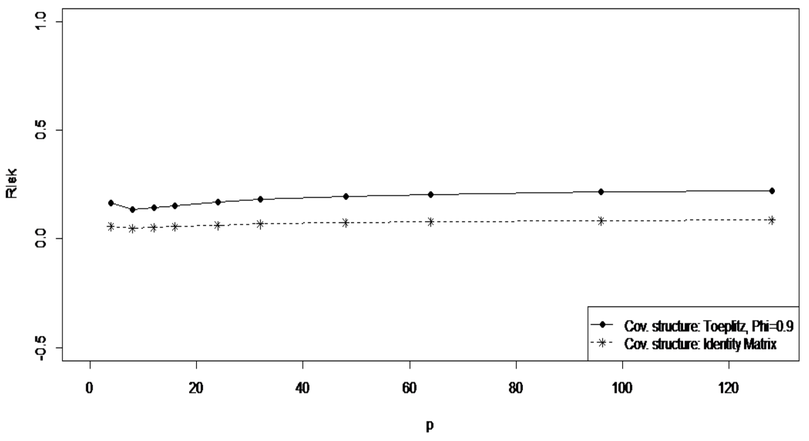

Figure 2: *Simulation with increasing p dimensions*

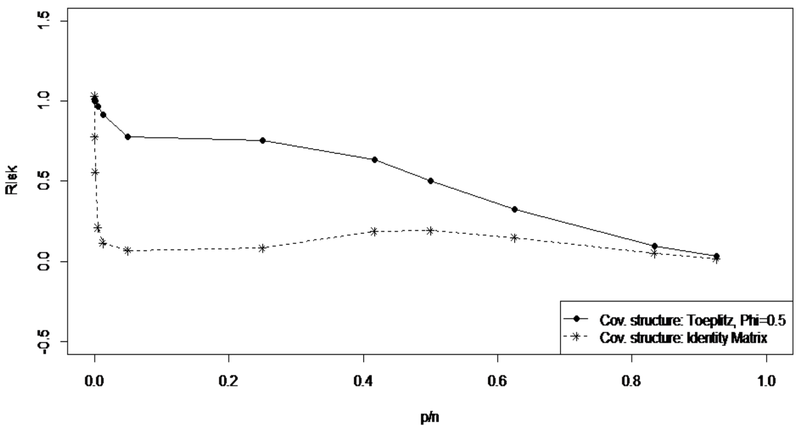

Figure 3: *Simulation with increasing p/n*

References

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

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.

Haff, L. (1977). Minimax estimators for a multinormal precision matrix, *Journal of Multivariate Analysis* **7**(3): 374–385.

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

James, W. and Stein, C. (1961). Estimation with quadratic loss, *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Vol. 1, pp. 1–379.

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

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