Issues of Incompleteness, Outliers and Asymptotics in High-Dimensional Data

PETER S. KARLSSON

JÖNKÖPING INTERNATIONAL
BUSINESS SCHOOL
JÖNKÖPING UNIVERSITY

Jönkoping International Business School
P.O. Box 1026
SE-551 11 Jönköping
Sweden
Tel. + 46 36 15 77 00
E-mail: info@jibs.hj.se
www.jibs.se

Issues of Incompleteness, Outliers and Asymptotics in High-Dimensional Data
JIBS Dissertation Series No. 069

ISSN 1403-0470
ISBN 978-91-86345-18-1

Printed by ARK Tryckaren AB, 2011

Acknowledgments

First I would like to express my deepest gratitude to my head supervisor Associate Professor Thomas Holgersson for his guidance and support for this dissertation work over the years. I am also grateful for his patience with his PhD candidates borrowing books from his private library that has been a most appreciated resource. I would also like to thank my deputy supervisor Professor Ghazi Shukur for his support and guidance, Professor Åke E. Andersson for his valuable advice and for initiating the idea of having a closer look at the APT model, Professor Scott Hacker for his valuable advice, my colleagues at Department of Economics and Statistics (especially Kristofer Månsson and Johan Eklund), my wife Hyunjoo Kim for her support and understanding and my friends Joel, Johan and Randolf. I also wish to thank Björn Kjellander for linguistic assistance. Finally, I wish to thank my mother and father, and my sister Lizelotte for their support and encouragement.

The work on this thesis has been supported by Sparbankernas Forskningsstiftelse, for which I am most grateful.

Jönköping March 2011

Peter Karlsson

Abstract

This thesis consists of four individual essays and an introduction chapter. The essays are in the field of multivariate statistical analysis of High-dimensional data. The first essay presents the issue of estimating the inverse covariance matrix alone and when it is used within the Mahalanobis distance in High-dimensional data. Three types of ridge-shrinkage estimators of the inverse covariance matrix are suggested and evaluated through Monte Carlo simulations. The second essay deals with incomplete observations in empirical applications of the Arbitrage Pricing Theory model and the interest is to model the underlying covariance structure among the variables by a few common factors. Two possible solutions to the problem are considered and a case study using the Swedish OMX data is conducted for demonstration. In the third essay the issue of outlier detection in High-dimensional data is treated. A number of point estimators of the Mahalanobis distance are suggested and their properties are evaluated. In the fourth and last essay the relation between the second central moment of a distribution to its first raw moment is considered in an financial context. Three possible estimators are considered and it is shown that they are consistent even when the dimension increases proportionally to the number of observations.

Content

Introduction and Summary of the Thesis

ARTICLE 1

ARTICLE 2

ARTICLE 3

ARTICLE 4

Introduction and Summary of the Thesis

1 Introduction

In traditional multivariate statistical analysis contexts the dimension of the parameter (*p*) of interest is fixed so that when the sample size (*n*) increases the statistic at hand will usually converge in probability to the true parameter. This is the standard asymptotic setting in multivariate statistical analysis. In many empirical multivariate contexts, however, these traditional assumptions are violated. This includes cases where a close relation between the number of parameters and the number of observations exists. In such situations, when *p* grows with *n* so that the ratio between them limits a constant c, i.e. $p/n \to c$, $0 < c < 1$, it frequently happens that the statistic does *not* converge to its true parameter. A well-known example is the non-convergence of the sample periodigram. Estimators specially designed for high-dimensional data should therefore be developed.

Another problem occurs if the data is incomplete. In empirical statistical analyses it happens that the observed data matrix may contain empty entries. This situation arises e.g. in a number of empirical applications in Finance. Data sets of Stock returns are incomplete due to the nature of stocks leaving or entering the stock exchange and hence the empty entries are not occurring at random. This disqualifies for example the traditional approach of estimating the APT model since the covariance matrix is not estimable in a traditional sense. This calls for an alternative approach that will be further investigated in the thesis.

An additional issue that may arise in statistical analysis is the concern of outliers. One possible way to detect outliers in $\mathbb{R}^p$ is to estimate the Mahalanobis distance (Mahalanobis, 1936) for each observation. But in a High-dimensional context (when *p* is close to *n*), the properties of traditional point estimators of the Mahalanobis distance may be poor and new estimators with improved properties are needed.

This introduction chapter includes a discussion of the issues mentioned so far and a summary of the papers in the thesis.

High-Dimensional Data and Increasing Dimension Asymtotics

In empirical analysis conducted on a given data set with fixed values of *n* and *p* one might wonder why asymptotic properties are relevant. The answer may be that for example in univariate settings we want to ensure that a sample of size $n = 50$ will on average give a more accurate analysis than a sample of size $n = 15$. Hence, as we collect more information (*n* increase) the analysis will be more accurate. But in multivariate contexts, this is not generally valid if $p/n = c$. Put differently, an analysis is not necessarily improved if *p* is increased from, say, 20 to 50. Hence, new asymptotic theory is needed for situations where the dimension increases along with the sample size. This kind of asymptotics is referred to as Increasing Dimension Asymptotics (IDA).

IDA was first initiated by Andreij Kolmogorov in the 1970s when he conducted pioneering research of statistical problems where the number of variables and the number of observations were allowed to grow simultaneously such that their quotient limits a constant (*c*) (Serdobolskii, 2008). Later on a number of important contributions were made to IDA where problems of estimating the inverse covariance matrix, expected value vectors and in discriminant analysis were investigated in IDA theory (Girko 1975, 1990, 1995; Serdobolskii 1985, 2000, 2008). In particular, the IDA theory estimators of the inverse covariance matrix are sometimes based on resolvent-type estimators where the central idea is to ensure that the eigenvalues of the estimated covariance matrix are not getting too small, hence avoiding that the estimator getting stochastically instable.

In many IDA contexts the limiting p/n ratio is restricted to the interval $0 < c < 1$, but it is questionable if this is a reasonable approach. A good IDA methodology should work well even in classical asymptotic situations, i.e. when $c = 0$, which means that the p/n ratio should be extended to the interval $0 \leq c < 1$ so that c is allowed to limit zero. The consequence of this is that for example in the first article (where three estimators of the inverse covariance matrix are investigated), it was found that some estimators behaves well over $0 \leq c < 1$, whereas others may behave poorly when c is in the neighborhood of $c = 0$ and that the standard estimator has optimal properties only for c close or equal to zero.

On the other hand, while some estimators may have optimal asymptotic properties, it is of great interest to know how they behave in a situation where *p* and *n* are fixed. An estimator that may be preferred due to its large sample properties may not be useful in small samples. This issue also prevails for estimators that have optimal IDA properties, since if an estimator is optimal when $p \to \infty$ and $p/n \to c$ in the interval $0 \leq c < 1$, then this does not imply that the estimator is optimal for a given pair of *p* and *n*. While the IDA properties of estimators of the inverse covariance matrix are known, their finite sample properties have only been investigated for some highly restrictive cases previously. Hence, their finite sample properties are investigated in this thesis through Monte Carlo simulations.

Outliers

The definition of an outlier varies among researchers in the literature where some are “a value which is dubious in the eyes of the analyst” (Dixon, 1950), “an outlying observation, or outlier, is one that appear to deviate markedly from other members of the sample in which it occurs” (Grubbs, 1969), and Beckman and Cook (1983) conclude that the term outlier seems “to indicate any observation that does not come from the target population”. The question then is how outliers can be identified and there are a number of proposed methods in the literature. Among these, the most widely used method for identification of multivariate outliers is the Mahalanobis distance (MD) and several outlier detection methods in line with this have been suggested (Comrey, 1985; Rasmussen, 1988; Wilks, 1963; Mardia, 1977). The MD is defined by:

$$d_i^2 = \left(\mathbf{x}_i - \boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu}\right),$$

where $\boldsymbol{\mu}$ is the population mean and $\boldsymbol{\Sigma}$ is the population variance. The traditional Mahalanobis distance estimator (MDE) is achieved by plugging in $\hat{\boldsymbol{\mu}} = \bar{\mathbf{x}}$ and $\hat{\boldsymbol{\Sigma}}^{-1} = \mathbf{S}^{-1}$ in d_i^2.

There are a number of alternative specifications of the Mahalanobis distance in outlier detection analysis. For example Mardia (1977) suggests “leave-one-out” estimators. The idea behind the leave-one-out estimators is to ensure that an outlier should not contaminate the estimation of the inverse covariance matrix and thereby avoiding being detected. Then there are extensions of this estimator. For example in some MD estimators the observation is also removed from the mean vector. This can also be done for several observations, i.e. the so called “leave-few-out” estimators (Maesschalck el al, 2000).

In this thesis the investigation of MDE is focused on point estimators of the MD in High-dimensional data. While estimating the Mahalanobis distance is straightforward in the traditional asymptotical setting (*p* fixed and $n \to \infty$) it becomes less obvious how to proceed in an IDA context (when $p/n \to c$ as $p \to \infty$ where $0 \leq c < 1$). This due to the fact that MDE is a composite estimator, which includes an estimator of the inverse covariance matrix (since $\boldsymbol{\Sigma}^{-1}$ is generally unknown) and in IDA settings the inverse sample covariance matrix may be degenerate with respect to some risk function. On the other hand, ridge-shrinkage estimators of the inverse covariance matrix have bounded risk functions. The potential of using ridge-shrinkage estimators within a number of specifications of MDE are investigated in article 1 and article 3. Moreover, such estimators require appropriate choices of the ridge coefficient. Methods for doing this are explored in this thesis.

Incompleteness

The Increasing Dimension Asymptotics is a typical setting in financial contexts where relations are investigated between all stocks in the cross section over a time period. This is because the number of stocks listed on a stock exchange at any time point during the time period will increase as we increase the time period, since new firms will enter the stock market. Simultaneously a number of stocks may exit the stock market due to a company may go bankrupt or become acquired by another firm, causing those stocks to have empty entries in the end. Hence there is a situation where an observed data matrix is incomplete. In general, a typical return matrix will have the pattern shown in the figure below:

$$\mathbf{r}_{4\times 5} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & \cdot \\ r_{21} & r_{22} & r_{23} & r_{24} & r_{25} \\ r_{31} & \cdot & r_{33} & r_{34} & r_{35} \\ r_{41} & \cdot & r_{42} & r_{44} & r_{45} \end{bmatrix}.$$

The above matrix shows an incomplete return matrix consisting of 5 assets measured over a period of 4 time units, where the second asset exited at $t = 3$ and the fifth asset entered at $t = 2$. Hence, as we increase the number of observations (T) then simultaneously the number of stocks with an incomplete set of observations increases. This is referred to as *the incompleteness problem of the APT model* in article 2. Depending on the nature of these empty entries, the method for dealing with them shifts. The term incomplete is used in the thesis to pinpoint that the empty entries in the data matrix are not missing values. Hence, it follows that for this kind of data matrixes with incomplete entries, methods such as the EM Algorithm are not applicable and alternative methods are required. One solution for situations with incomplete data as described above is to use a model-independent approach where one may extract principal components from an incomplete dataset without estimating the covariance matrix (Cristoffersson, 1970; Ruhe, 1974; Wiberg, 1976). This method will only use the observed data to find principal components that in turn can be used in further statistical analyses. This method has the strength of not assuming an underlying distribution and is discussed in detail in article 2.

This thesis consists of four papers, which deal with the above-presented issues. The first paper investigates and develops estimators of the inverse covariance matrix in an IDA setting. These estimators are then applied within estimators of the Mahalanobis Distance in situations when the population mean is known. The issue of Incompleteness of the APT model is investigated in the second paper. A solution to the problem is suggested and a case study is performed for the purpose of demonstration. The third paper treats the issue of outliers in a high-dimensional context, where a number of estimators of the Mahalanobis distance are suggested and evaluated. In the fourth paper three estimators for estimating Mean-Standard Deviation Ratios of Financial Data are developed in high-dimensional contexts.

2 Summary of the papers

Article 1 An Investigation and Development of Three Estimators of Inverse Covariance Matrices with Applications to the Mahalanobis Distance

The first paper in the thesis 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. The latter setting is referred to as IDA in the literature where the ratio between dimension and the number of observations limits a constant, in contrast to the classical setting where the dimension is fixed and the number of observations limits infinity. In this type of setting, three different ridge-type estimators are investigated of which two are proposed in the paper and one has been considered previously in the literature. The estimators are characterized by their different functional form in which the ridge coefficient operates multiplicatively, additively or as a mixture of these two. While the properties of the multiplicative estimator are partly known from previous research, the properties of the two new estimators are not known and these are not easy to obtain analytically, at least not within an IDA context. In addition little is known about these estimators' behavior in finite, given combinations of *n* and *p*. For these reasons a Monte Carlo Simulation is performed where the proposed IDA estimator's properties are evaluated relative to the traditional estimator. Furthermore, in a practical situation it is needed to decide an appropriate value of the ridge coefficient. By rewriting the three investigated estimators into a general form the authors were able to derive a readily available estimator of the IDA risk function (Serdobolskii, 2000, 2008). This gives a tool for finding the optimal value of the ridge coefficient within the proposed estimators.

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.

Article 2 The Incompleteness Problem of the APT Model

In the second paper, two serious problems arising in empirical analysis of the APT model are discussed. It is argued that the traditional solution to the incompleteness problem by simply excluding stocks without a complete set of observations from the analysis leads to (i) an asymptotically empty set containing no observations at all and (ii) will lead to selection bias in that only the largest companies will remain for any fixed time period.

The first fact follows almost trivially, while the bias part is demonstrated empirically. Also, as soon as a stock has been delisted from the stock exchange, there will not be any further observations. This causes an observed data matrix to have empty entries. These are referred to as incomplete entries (as opposed to "missing values") to emphasize the fact that they are not supposed to be there, i.e. the empty entries in the data matrix are not missing values. Hence, imputation methods such as EM Algorithm are not applicable and alternative methods should be considered instead. More specifically, it is proposed that in most practical situations, one should either estimate the covariance matrix (the main ingredient of the APT model) by a parametric model containing only a few parameters, or use some optimization algorithm for the data at hand. In particular, it is demonstrated that band matrices may be consistently estimated with respect to a certain norm, regardless of whether the data matrix is incomplete and of increasing dimension. It is also argued that non-parametric methods may be less risky, when compared to parametric approaches and that appropriate techniques for this purpose have already been developed – such as Wiberg's method (Wiberg, 1976) – though they do not seem to have been used in this context previously. We also conduct a case study on the Stockholm OMX data using the APT model. The empirical specification of the APT model in this paper is based on the two-step technique initially proposed by Roll and Ross (1980). Four factors are extracted for the data at hand and we find that a high proportion of the cross-sectional regressions are significant and hence the extracted risk factors are priced. It is empirically shown that the method of only including assets with a long history will lead to a selection bias, and that the non-parametric optimization technique for extracting principal components works surprisingly well and should be implemented in all similar pricing models.

Article 3 Point estimators of the Mahalanobis distance in High-Dimensional Data

This paper treats the problem of estimating the Mahalanobis distance when the dimension of the matrix is comparable to the sample size. Earlier research suggests that the traditional inverse sample covariance matrix is not very useful in such high-dimensional cases, and alternative estimators are required.

The Mahalanobis distance is, e.g., used for identifying potential outliers which can be done graphically by plotting the Mahalanobis distance for a data set at hand in a box plot. In this paper two ridge-shrinkage estimators of the Mahalanobis distance between a single observation and the mean value vector are proposed. Leave-one-out estimators as well as the traditional estimator for known or unknown mean value are considered. These include resolvent-type and ridge-type estimators for estimating the inverse covariance matrix within the composite distance estimator. Both estimators depend on a ridge-shrinkage coefficient that has to be determined from data. Hence the paper also proposes a consistent estimator of a risk function that may be applied to find this coefficient. In other words, the estimators are adaptive. The risk function may be used to either identify the best (optimal) estimator or to identify a set of estimators superior to the standard estimator (good estimators). Analytical properties along with a small Monte Carlo simulation were used to investigate the main properties of the estimators, with respect to the true parameter and also to the traditional estimator. The main findings are that both the resolvent-type estimator and the ridge-type estimator perform very well relative to the traditional estimator. This difference becomes accentuated when the dimension is high relative to the sample size.

Article 4 Estimating Mean-Standard Deviation Ratios of Financial Data

In the final paper the relation between the second central moment of a distribution to its first raw moment is considered. The linkage between signal (in terms of centrality) and noise (spread) has long been considered in statistics. The most commonly used expression is probably that of the coefficient of variation suggested by Pearson (1895), defined by the ratio between the standard deviation to the mean value, and its reciprocal value has been frequently used in fields such as imaging (Lopes et al, 1993) and finance (Osteryoung, 1977) etc.

In this paper the later relation is of interest, where according to the theory of finance, an investor is compensated by increased expected returns for taking higher risks (i.e. variance of the investments returns). This is sometimes represented by the ratio of the mean value to the standard deviation (Sharpe, 1966). In most cases these applications have involved the problem of making inference of the relation between the mean and standard deviation of a single variable. In this context, however, there is a systematic relation between the standard deviations and mean values of a large number of heterogeneous variables. This relation is investigated by a simple model linking the mean and the standard deviation (risk) of excess returns. The model allows for a common relation in form of a scalar parameter (β) and allows each asset's relation between mean and the standard deviation (risk) of excess returns to deviate from the common relation. Three estimators are proposed and it is shown that they are consistent even when the dimension increases proportionally to the number of observations. An empirical study is conducted on the stocks of the Stockholm stock exchange during June 1995 - June 2010. The population was divided into three segments depending on the market capitalization value. It is seen that the three segments β are heterogeneous. Moreover, it is also argued that two out of the three estimators are sensitive to extreme observations, whereas the third is robust.

References

Christofferson, A. (1970). The one component model with incomplete data. Ph.D. Thesis, Institute of Statistics. Uppsala University, Uppsala.

Comrey, A. L. (1985). A method for removing outliers to improve factor analytical results, *Multivariate Behavioral Research*, 20, 273-281.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977**).** Maximum Likelihood from Incomplete Data via the EM Algorithm**.** *Journal of the Royal Statistical Society, Series B (Methodological)*, 39(1), 1-38.

Girko, V.L. (1975). *Random Matrices*. Kiev. Vyshcha shkola.

Girko, V.L. (1990). *Theory of Random Determinants*. Kluwer Academic Publisher.

Girko, V.L. (1995). *Statistical Analysis of Observations of Increasing Dimension.* Kluwer Academic Publisher.

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

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India* 2 (1): 49–55.

Mardia, K. V. (1977). (Krishnaiah, ed.) *Mahalanobis Distances and Angles, Multivariate Analysis IV*, North-Holland.

Rasmussen, J. L. (1988). Evaluating outlier identification tests: Mahalanobis D Squared and Comrey D. *Multivariate Behavioral Research*, 23(2), 189-202.

Roll, R. and Ross, S. A. (1980). An Empirical Investigation of the Arbitrage Pricing Theory. *Journal of Finance*, 35(5), 1073-1103.

Ruhe, A. (1974). Numerical computation of principal components when several observations are missing. Technical report. UMINF-48-74, Dept. Information Processing, Umeå University, Umeå.

Serdobolskii, V. I. (2000). *Multivariate Statistical Analysis: A High-dimensional Approach.* Springer-Verlag.

Serdobolskii, V. I. (1985). “The resolvent and the spectral functions of sample covariance matrices of increasing dimension”. *Russian Mathematical Surveys* 40: 232–233.

Serdobolskii, V. I. (2008). *Multiparametric Statistics*. Elsevier.

Wiberg, T. (1976). Computation of principal components when data are missing. Proc. Second Symp. Computational Statistics, 229–236.

Wilks, S. S. (1963). Multivariate Statistical Outliers. *Sankhya A*, 25, 407-426.