Computational Statistics and Data Analysis 93 (2016) 404–420

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust estimation of precision matrices under cellwise contamination

G. Tarr*, S. Müller, N.C. Weber

School of Mathematics and Statistics, University of Sydney, NSW 2006, Australia

ARTICLE INFO

Article history:
Received 1 April 2014
Received in revised form 8 February 2015
Accepted 8 February 2015
Available online 23 February 2015

Keywords:
Precision matrix
Covariance matrix
Robust estimation
Data mining

ABSTRACT

There is a great need for robust techniques in data mining and machine learning contexts where many standard techniques such as principal component analysis and linear discriminant analysis are inherently susceptible to outliers. Furthermore, standard robust procedures assume that less than half the observation rows of a data matrix are contaminated, which may not be a realistic assumption when the number of observed features is large. The problem of estimating covariance and precision matrices under cellwise contamination is investigated. The use of a robust pairwise covariance matrix as an input to various regularisation routines, such as the graphical lasso, QUIC and CLIME is considered. A method that transforms a symmetric matrix of pairwise covariances to the nearest covariance matrix is used to ensure the input covariance matrix is positive semidefinite. The result is a potentially sparse precision matrix that is resilient to moderate levels of cellwise contamination. Since this procedure is not based on subsampling it scales well as the number of variables increases.

1. Introduction

Often the aim of data mining and statistics is to extract information about the relationships between the variables and identify any features or structure in the data. The covariance matrix, $\mathbf{\Sigma} = \text{var}(\mathbf{y})$, where $\mathbf{y} \sim \mathbf{F}$, the distribution of the true data generating process, and its inverse, the precision matrix $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$ are fundamental components of many statistical procedures, such as principal component analysis (PCA) and linear discriminant analysis. However, it is well known that the classical covariance matrix is inherently non-robust to outliers and suffers from distortion in its eigenstructure in high dimensions (Johnstone, 2001). This paper combines pairwise covariance matrix estimation with recent regularisation routines currently used in bioinformatics and machine learning to produce an estimated precision matrix that is robust to moderate levels of cellwise contamination.

The need for robust statistics in data mining and associated fields is well known, see Barnett and Lewis (1994) for a general overview. In particular, it is desirable for learning algorithms to be stable with respect to noisy features and unusual fluctuations in the inputs. For example Li (2004) considers robust incremental PCA applied to multi-view face modelling and Mavroeidis and Marchiori (2014) consider the stability of sparse PCA in the context of feature selection in microarray gene expression data. Other situations where robust techniques are important include speech recognition and neural networks, see Gales and van Dalen (2007) and Bieroza et al. (2011), respectively. Outlier detection in high dimensions has also been addressed by Filzmoser et al. (2008).

* Corresponding author.
E-mail address: garth.tarr@sydney.edu.au (G. Tarr).

http://dx.doi.org/10.1016/j.csda.2015.02.005
0167-9473/

In the statistics literature, robust estimation of covariance matrices has received much attention in the past, notably the minimum volume ellipsoid and minimum covariance determinant (MCD) estimators, projection type estimators and M-estimators, see Hubert et al. (2008) for a survey. Furthermore, research into covariance matrix estimation and its applications is ongoing, see for example Filzmoser et al. (2014) who use the MCD estimator to construct robust Mahalanobis distances to identify local multivariate outliers; Hubert et al. (2014) who study the shape bias of a range of existing robust covariance matrix estimators; or Cator and Lopuhaä (2010, 2012) who consider asymptotic expansions and establish asymptotic normality for general MCD estimators.

An alternative approach is to estimate the covariance matrix in a component-wise manner based on a robust estimator of scale as outlined by Ma and Genton (2001). It is well known that the resulting symmetric matrix is not guaranteed to be positive definite (PD). Methods to ensure the resulting estimator is PD have previously been explored by Rousseeuw and Molenberghs (1993) with notable updates in the robustness literature by Maronna and Zamar (2002) and quite separately in the finance literature by Higham (2002). Alqallaf et al. (2002) also proposed a pairwise approach to covariance matrix estimation by means of first Winsorising the data. The resulting Quadrant Covariance estimate does not necessarily require a transformation to ensure the result is positive definite.

In practice, it is often the precision matrix, the inverse of the covariance matrix, that is primarily of interest. This is the case, for example, in Gaussian graphical model selection. As such, this paper is primarily concerned with robustly estimating the precision matrix. While there is an obvious link between covariance matrices and precision matrices, it is not obvious that a good (robust) estimator for one results in a good estimator for the other. For a recent example where robust covariance matrices have been used to estimate the precision matrix see Gottard and Pacillo (2010). We will employ robust pairwise covariance matrices as a starting point for various regularisation techniques to facilitate the estimation of robust, potentially sparse, precision matrices.

Classical robust estimators assume that contamination occurs within a restricted subset of the observation vectors, however, in recent years there has been interest in developing robust estimators that perform well under cellwise contamination. The cellwise contamination model was initially explored in a data mining context by Alqallaf et al. (2002) and later defined comprehensively by Alqallaf et al. (2009). This form of contamination is prevalent in large, automatically generated data sets, found in data mining and bioinformatics, where there is often little quality control over the inputs. Cellwise contamination is common in the context of missing data, however, it represents a philosophical divergence from the traditional approach to robustness. Recent examples where the problem of cellwise contamination have arisen include, Farcomeni (2014), Van Aelst et al. (2012) and Agostinelli et al. (2014).

We perform a detailed simulation study to assess the performance of a variety of precision matrix estimators in the presence of cellwise contamination over a number of scenarios and levels of p while keeping the sample size fixed. Our results are distilled from a comprehensive range of performance indices. We outline these indices and consider their applicability to the various scenarios in the supplementary material accompanying this article (see Appendix A). We show that the pairwise nature of the covariance estimates enables the resulting precision matrix to have a higher level of robustness than when using standard robust covariance matrix estimation procedures in the presence of cellwise contamination. This is a novel result and a significant first step towards dealing with cellwise contamination in this context.

The remainder of this paper is structured as follows. Section 2 outlines the cellwise contamination model and highlights why standard robust techniques fail in this setting. Sections 3 and 4 outline the theory for existing pairwise covariance matrix estimation techniques and regularisation routines and we propose a new procedure which combines robust pairwise covariance matrix estimation with regularisation. Sections 5 and 6 present the results of an extensive simulation study and Section 7 summarises the important findings.

2. Cellwise contamination

Consider a data set $\mathbf{X} \in \mathbb{R}^{n\times p}$ consisting of n observations on p variables. Classically, even the most robust procedures are designed such that they only work when at most half of the rows in $\mathbf{X}$ have contamination present.

Alqallaf et al. (2009) formally outline the cellwise contamination model as an extension of the standard Tukey–Huber contamination model which was first introduced in the univariate location–scale setup (Tukey, 1962; Huber, 1964). Consider the data generating process for the n rows in $\mathbf{X}$, $\mathbf{x}_i = (\mathbf{I} - \mathbf{B}_i)\mathbf{y}_i + \mathbf{B}_i\mathbf{z}_i$, where $\mathbf{y}_i \sim \mathbf{F}$, the distribution of well-behaved data, $\mathbf{z}_i \sim \mathbf{G}$, some outlier generating distribution and $\mathbf{B}_i = \text{diag}(B_1, \ldots, B_p)$ is a diagonal matrix, where $B_1, \ldots, B_p$ are Bernoulli random variables, $B_j \sim \mathcal{B}(1, \varepsilon_j)$. When $\mathbf{y}$, $\mathbf{B}$ and $\mathbf{z}$ are independent we have a situation that is similar to the missing completely at random model, where the missingness does not depend on the values of $\mathbf{y}$, see, for example, Little and Rubin (2002).

The structure of $\mathbf{B}_i$ determines the contamination model. If $B_1, \ldots, B_p$ are fully dependent, then $\mathbf{B}_i = U_i\mathbf{I}$, where $U_i \sim \mathcal{B}(1, \varepsilon)$, and we recover the fully dependent contamination model, the standard model on which classical robust procedures are based. In this setting, the probability that an observation is uncontaminated, $1-\varepsilon$, is independent of the dimensionality. Furthermore, the proportion of contaminated observations is preserved under affine equivariant transformations.

In contrast, if $B_1, \ldots, B_p$ are mutually independent we have the fully independent contamination model, where each element of $\mathbf{x}_i$ is drawn from $\mathbf{F}$ or $\mathbf{G}$ independently of the other $p-1$ elements in $\mathbf{x}_i$. That is, contaminating observations occur independently at the univariate level. In this setting, it may be unreasonable to assume that less than half the rows have contamination. Furthermore, if p is large and there is only one outlier in an observation vector, then down-weighting the entire observation may be wasteful.