Computational Statistics and Data Analysis 75 (2014) 15–27

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

A joint convex penalty for inverse covariance matrix estimation

Ashwini Maurya *

Department of Statistics and Probability, Michigan State University, United States

ARTICLE INFO

Article history:
Received 29 July 2013
Received in revised form 8 January 2014
Accepted 22 January 2014
Available online 31 January 2014

Keywords:
Proximal gradient
Joint penalty
Convex optimization
Sparsity

ABSTRACT

The paper proposes a joint convex penalty for estimating the Gaussian inverse covariance matrix. A proximal gradient method is developed to solve the resulting optimization problem with more than one penalty constraints. The analysis shows that imposing a single constraint is not enough and the estimator can be improved by a trade-off between two convex penalties. The developed framework can be extended to solve wide arrays of constrained convex optimization problems. A simulation study is carried out to compare the performance of the proposed method to graphical lasso and the SPICE estimate of the inverse covariance matrix.

Published by Elsevier B.V.

1. Introduction

Recent surge in the use of electronic and digital technology has created vast amount of high dimensional data whose analysis demands advanced statistical tools and computational techniques. Examples are biological data of gene expression measurement, FMRI scanned images of human brain and Netflix data. In these datasets, one often have very few observations as compared to the number of variables and therefore the choice of standard statistical methods becomes inappropriate for making valid inference. Thus the concept of parsimony becomes very crucial. In many applications the problem of interest is often estimating the dependence structure of the data where the underlying probability distribution of observations is either fixed (static) or evolving over time (i.e. time varying or dynamic). In the static framework, a common assumption is that the data are independently and identically distributed (i.i.d.), whereas in dynamic setting, the distribution of data evolves over time and hence the i.i.d. assumption no longer remains valid. Here we focus on the static framework.

1.1. Background

The covariance selection was introduced by Dempster (1972) where the basic idea was to (i) introduce parsimony in parameter model fitting and (ii) exploit the powerful yet elegant theory of exponential family as a tool of practical data analysis. The computational ease along with attractive statistical features of Gaussian distribution makes it a popular choice for most of the application problems. Estimation of the inverse covariance matrix is important in a number of statistical analyses including:

* Correspondence to: 619 Red Cedar Road, Department of Statistics and Probability, Michigan State University, C508 Wells Hall, East Lansing, MI, 48824-1027, United States. Tel.: +1 517 802 7757.

E-mail addresses: mauryaas@stt.msu.edu, mauryaas@msu.edu, akmaurya07@gmail.com.

0167-9473/$ – see front matter. Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.csda.2014.01.015

- Gaussian Graphical Modeling: In Gaussian graphical modeling, a zero entry of an element in the inverse of a covariance matrix corresponds to conditional independence between the variables.
- Linear or Quadratic Discriminant Analysis: When the features are assumed to have multivariate Gaussian density, the resulting discriminant rule requires an estimate of the inverse covariance matrix.
- Principal Component Analysis (PCA): In multivariate high dimensional data it is often desirable to transform the high-dimensional feature space to a lower dimension without loosing much information. The covariance matrix method is a popular method for PCA estimates.

Several approaches have been suggested to address the estimation problem of the inverse of a covariance matrix. These approaches are either based on regularized estimation of the inverse of the covariance matrix (Banerjee et al., 2008; Friedman et al., 2007; Bickel and Levina, 2008; Rothman et al., 2008) or regularized high dimensional regression (Meinshausen and Bühlmann, 2006; Zhou et al., in press). Among earlier developments, an exact maximization of ℓ_1 penalized log likelihood using interior point methods was suggested for estimating the inverse covariance matrix (Dahl et al., 2008; Yuan and Lin, 2007; Banerjee et al., 2008). Let $X = (X_1, X_2, \ldots, X_p)^T$ be a p-variate random vector from multivariate Gaussian distribution $N_p(\mu, \Sigma)$. μ is the mean vector and Σ is the positive definite covariance matrix. Let $X^1, X^2, \ldots, X^n$ be n independent copies of X. The sample covariance matrix is given by

$$S = (1/n) \sum_{i=1}^{n} (X^i - \bar{X})(X^i - \bar{X})^T \tag{1.1}$$

where ${X^i}^T$ is the transpose of X^i and $\bar{X}$ is the mean vector of sample observations.

Let $\hat{W}$ be the estimate of inverse of the covariance matrix Σ. Banerjee et al. (2008) have considered the following determinant maximization (MAXDET) (Vandenberghe and Boyd, 2004) problem:

$$\hat{W} = \arg\max_{X \succ 0} \{\log(\det(X)) - \mathrm{tr}(SX) - \lambda \|X\|_1\} \tag{1.2}$$

where $\mathrm{tr}(S)$ is the trace of the matrix S, $\|X\|_1$ is the ℓ_1 norm and defined as the sum of absolute values of elements of matrix X and λ is the regularization parameter which controls the sparsity structure of the estimated inverse covariance matrix. Another approach to solve the above optimization problem is based on high dimensional regression (Yuan, 2009; Meinshausen and Bühlmann, 2006; Wainwright et al., 2006). For Gaussian graphical models, the main motivation behind the regression approach is the explicit relation between the elements of the inverse covariance matrix and coefficients of the predictor variables (Yuan, 2009). Meinshausen and Bühlmann (2006) follow this approach and estimate the neighborhood structure of the variables by fitting ℓ_1 regularized regression to each of the variables on the remaining set of variables as predictors. They also have established the consistency of their estimates under certain assumptions of sparsity and stability.

Friedman et al. (2007) introduced graphical lasso which has better computational power compared to earlier methods. In graphical lasso, the algorithm obtains the lasso estimate of each row/column of the covariance matrix given the sample covariance matrix. It uses the block co-ordinate descent algorithm for optimization of the objective function.

Rothman et al. (2008) has proposed the sparse permutation invariance covariance estimate (SPICE). This method uses Cholesky decomposition of the inverse covariance matrix and a quadratic approximation of the likelihood function to simplify the problem as finding the minimum of each univariate parameter in the objective function in closed form. The objective function is invariant and consequently the estimator remains permutation invariant. The method uses the cyclical co-ordinate descent algorithm to do the optimization.

In another approach, Sheena and Gupta (2003) have proposed a constrained maximum likelihood estimator with restrictions on the lower or upper bound of the eigenvalues. This method focuses on only one of the two ends of the eigenspectrum and thus the resulting estimator does not correct for the overestimation of the large eigenvalues and underestimation of the small eigenvalues simultaneously. Consequently their approach does not address the distortion of the entire eigenspectrum especially in small sample sizes. Won et al. (2012) consider a maximum likelihood estimation of the covariance matrix with condition number constraint. The condition number of a matrix is defined as the ratio of largest to smallest eigenvalue of the matrix. However this approach itself requires an estimation of condition number.

To control the distortion of eigenspectrum of the covariance matrix, we consider a joint penalty of sum of singular values (trace norm) in addition to the ℓ_1 norm. By minimizing the joint penalty function of ℓ_1 and the trace norm, the resulting estimated inverse covariance matrix is sparse as well as singular values of the corresponding covariance matrix are more centered than the observed sample covariance matrix. Rolfs et al. (2012) consider the estimation of the inverse covariance matrix which can be seen as a particular case of the proposed approach by setting-off the trace norm penalty. A single penalty of the ℓ_1 norm is appropriate when the underlying true inverse covariance matrix is sparse. However it does not control the distortion in eigenspectrum of the inverse covariance matrix. Controlling the eigenspectrum is an intuitive way to get a well conditioned estimate of the inverse covariance matrix.

1.2. Contribution

We propose a joint convex penalty of ℓ_1 and the *trace norm* to the inverse covariance matrix estimation. The estimator thus obtained is simultaneously sparse and gives better performance than graphical lasso for small sample size in terms of