Computational Statistics and Data Analysis 56 (2012) 2626–2636

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Fitting very large sparse Gaussian graphical models

Harri Kiiveri [a,*], Frank de Hoog [b]

[a] CSIRO Mathematics Informatics and Statistics, The Leeuwin Centre, 65 Brockway Rd, Floreat, 6014, Western Australia, Australia
[b] CSIRO Mathematics Informatics and Statistics, GPO Box 664, Canberra, ACT, Australia

ARTICLE INFO

Article history:
Received 5 April 2011
Received in revised form 6 February 2012
Accepted 7 February 2012
Available online 18 February 2012

Keywords:
Covariance selection
Gene networks
Graphical models
High dimensional
Large scale optimisation
Limited memory quasi-Newton

ABSTRACT

In this paper we consider some methods for the maximum likelihood estimation of sparse Gaussian graphical (covariance selection) models when the number of variables is very large (tens of thousands or more). We present a procedure for determining the pattern of zeros in the model and we discuss the use of limited memory quasi-Newton algorithms and truncated Newton algorithms to fit the model by maximum likelihood. We present efficient ways of computing the gradients and likelihood function values for such models suitable for a desktop computer. For the truncated Newton method we also present an efficient way of computing the action of the Hessian matrix on an arbitrary vector which does not require the computation and storage of the Hessian matrix. The methods are illustrated and compared on simulated data and applied to a real microarray data set.

The limited memory quasi-Newton method is recommended for practical use.

1. Introduction and motivation

High throughput biotechnology measurement platforms are now routinely generating data sets with numbers of variables varying from thousands to millions. To perform simple multivariate analyses of these data sets requires consideration of the correlation structure of these large sets of variables, i.e. estimation of large covariance matrices. For example, traditional microarray analysis (Smyth, 2004) makes the unrealistic assumption that gene expression measurements are independent. The simplest way to relax this assumption is to specify a covariance matrix for all the measurements. This will typically be of the size 20 000 by 20 000 or more for microarray data. Clearly there are issues in computing and storing such a large covariance matrix. One possible solution to this problem is to make this large matrix, or its inverse, sparse. In this paper we will consider sparse inverse covariance matrices (Dempster, 1972) which can be used to model some of these data sets (for an application to microarray data sets see Kiiveri, 2011). Such matrices define Gaussian graphical models (sometimes referred to as covariance selection models), see for example Speed and Kiiveri (1986) and Lauritzen (1996), which have conditional independence interpretations. In particular the nonzero entries in the ith row of the inverse covariance matrix define the regression coefficients of the ith variable on its neighbouring variables (genes), the other regression coefficients being zero. Writing Y_i for the ith variable, Y_{-i} for the vector of the remaining variables, μ_i for the mean of variable Y_i and σ^{ij} for the i,jth element of the inverse covariance matrix Σ^{-1}, we have

$$\begin{aligned} E\{Y_i \mid Y_{-i}\} &= \mu_i - \sum_{j \neq i} \left(\frac{\sigma^{ij}}{\sigma^{ii}}\right)(Y_j - \mu_j) \\ \operatorname{var}\{Y_i \mid Y_{-i}\} &= (\sigma^{ii})^{-1} \end{aligned} \tag{1}$$

where E denotes conditional expectation and var conditional variance. Hence the regression coefficient of variable i on j is $\beta_{ij} = -(\sigma^{ij}/\sigma^{ii})$.

* Corresponding author. Tel.: +61 8 93336315; fax: +61 8 93336121.
E-mail addresses: harri.kiiveri@csiro.au (H. Kiiveri), frank.dehoog@csiro.au (F. de Hoog).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2012.02.007

Fig. 1. Example zero pattern (* denotes nonzero) in inverse covariance matrix and corresponding graphical representation.

The interpretation of zero elements of the inverse being equivalent to regression coefficients being zero makes this class of models attractive for analysing microarray data as it provides useful information about the (linear) interrelationships between genes.

An undirected graph can be associated with any pattern of zeros in the inverse covariance matrix by the relation: there is an edge between vertices i and j if and only if $\sigma^{ij} \neq 0$, where vertex i denotes variable Y_i. An example of this is given in Fig. 1.

The cliques (maximal sets of vertices which are all adjacent) of the graph are {1, 2, 5}, {2, 3, 5}, {3, 4, 5}, and {1, 4, 5}. For this example, the regression of variable 1 on the rest of the variables has nonzero regression coefficients for variables 2, 4 and 5. The associated undirected graph of a model gives a compact visual representation of the connections between variables. Graphical representations of the model are also useful in describing properties of the maximum likelihood estimate and we will make use of this below.

This paper is somewhat unusual in that it contains a collection of existing material which is scattered throughout the statistical and numerical analysis literature and as such is not novel. (However, we would point out that details in Section 5.2 would appear to be new material.) What is novel is the creative use of this material to make it possible to do something new, that is to be able to fit very large sparse Gaussian graphical models, which to the best of our knowledge has not been possible to date.

The rest of the paper is structured as follows. In Section 2 we outline a strategy for the determination of the zero pattern in high dimensional covariance selection models. Given a pattern of zeros, Section 3 presents the likelihood function to be maximised as well as expressions for its first two derivatives. Section 4 discusses existing algorithms for fitting inverse covariance matrices and gives a unified background for the quasi-Newton algorithm and the truncated Newton method. Section 5 explains how to organise the calculations so as to make the limited memory quasi-Newton and truncated Newton algorithms feasible for fitting very large inverse covariance matrices. An illustrative example is given in Section 6 along with a comparison with an existing method. Finally we finish with a discussion and conclusions in Section 7.

2. Determining the pattern of zeros

Determining the pattern of zeros, i.e. model choice, is an important issue and we present a practical strategy for doing that in this section. Later, as we focus on the computation of maximum likelihood estimates, we will assume that a pattern of zeros has been specified and an estimate of Σ^{-1} is required.

There are two common methods for determining the pattern of zeros for a given a data set. The first method involves computing p individual regressions of each variable on the remaining variables. This is intuitively reasonable given our earlier discussion about the interpretation of elements of the inverse covariance matrix in terms of regression coefficients. The regression method could incorporate a sparsity penalty, see Meinshausen and Bühlmann (2006), which also contains a theoretical justification for the method, or simply be some form of consistent stepwise variable selection, either using a forward stepwise variable selection method as in Zhang (2009), or a combination of forward and backward selection with a modified Bayesian information criterion (BIC), An et al. (2008). An attractive feature of these methods is the ability to easily distribute the problem over a number of processors.

A second class of methods is maximum likelihood estimation with L1 (more generally sparsity) constraints on the elements of sigma inverse. These methods accomplish simultaneous model selection and fitting, see for example Banerjee et al. (2008) and Friedman et al. (2008a). Unfortunately these methods currently do not scale up well to higher dimensions. An order of magnitude increase in number of variables makes a big difference in computing times and storage requirements. See Kiiveri (2011) for more discussion on this point. Note that if we use these methods as a pattern selector, we still may wish to compute maximum likelihood estimates of parameters for a selected pattern of zeros.

2.1. A practical strategy

A practical strategy for determining the zero pattern when p is very large, motivated by Meinshausen and Bühlmann (2006), Chen and Chen (2008) and Eq. (1) is as follows.

(i) For each variable, considering the remaining variables as potential predictors, produce a (nested) sequence of models by forward stepwise regression. The R package lars implementing the methods of Efron et al. (2004) can be used to do this.