Biostatistics (2006), **7**, 2, *pp.* 302–317
doi:10.1093/biostatistics/kxj008
Advance Access publication on December 2, 2005

Gradient directed regularization for sparse Gaussian concentration graphs, with applications to inference of genetic networks

HONGZHE LI*, JIANG GUI

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, 920 Blockley Hall, 423 Guardian Drive, Philadelphia, PA 19104-6021, USA
hli@cceb.upenn.edu

SUMMARY

Large-scale microarray gene expression data provide the possibility of constructing genetic networks or biological pathways. Gaussian graphical models have been suggested to provide an effective method for constructing such genetic networks. However, most of the available methods for constructing Gaussian graphs do not account for the sparsity of the networks and are computationally more demanding or infeasible, especially in the settings of high dimension and low sample size. We introduce a threshold gradient descent (TGD) regularization procedure for estimating the sparse precision matrix in the setting of Gaussian graphical models and demonstrate its application to identifying genetic networks. Such a procedure is computationally feasible and can easily incorporate prior biological knowledge about the network structure. Simulation results indicate that the proposed method yields a better estimate of the precision matrix than the procedures that fail to account for the sparsity of the graphs. We also present the results on inference of a gene network for isoprenoid biosynthesis in *Arabidopsis thaliana*. These results demonstrate that the proposed procedure can indeed identify biologically meaningful genetic networks based on microarray gene expression data.

Keywords: Empirical Bayes thresholding; Graphical models; Microarray; Threshold gradient descent.

1. INTRODUCTION

The completion of the human genome project and the development of many high-throughput genomic technologies make it possible to systematically define the organization and function of gene, protein and metabolite networks. Large-scale microarray gene expression data provide the possibility of learning gene regulation from expression profiles and constructing the gene regulatory networks and pathways or cellular networks (Ideker *et al.*, 2001; Friedman, 2004). Early research has mainly focused on using clustering analysis to identify coregulated genes (Tavazoie *et al.*, 1999). Recently, some efforts have been devoted to developing probabilistic models for modeling regulatory and cellular networks based on genome-wide high-throughout data, including both Bayesian network modeling (Friedman, 2004; Segal *et al.*, 2003)

*To whom correspondence should be addressed.

and Gaussian graphical modeling (Schafer and Strimmer, 2005; Wille *et al.*, 2004; Dobra *et al.*, 2004). The goal of such probabilistic modeling is to investigate the patterns of association in order to generate biological insights plausibly related to the underlying biological and regulatory pathways.

Graphical models use graphs to represent dependencies among stochastic variables. The graphical approach yields dependence models that are easily visualized and presented. One specific graphical model is the Gaussian graphical model, which assumes that the multivariate vector follows a multivariate normal distribution with a particular structure of the inverse of the covariance matrix, often called the precision or concentration matrix. For such Gaussian graphical models, it is usually assumed that the patterns of variation in expression for a given gene will be predicted by those of a small subset of other genes. This assumption leads to sparsity (i.e., many zeros) in the precision matrix of the multivariate distribution and reduces the problem to well-known neighborhood selection or covariance selection problems (Dempster, 1972). In such a concentration graph modeling framework, the key idea is to use partial correlation as a measure of independence of any two genes, rendering it straightforward to distinguish direct from indirect interactions. This is in contrast to the covariance graphical model where marginal correlations are used. It has been demonstrated in the literature that many biochemical and genetic networks are not fully connected (Tegner *et al.*, 2003; Jeong *et al.*, 2001; Gardner *et al.*, 2003) and many genetic interaction networks contain many genes with few interactions and a few genes with many interactions. Therefore, the genetic networks are intrinsically sparse and the corresponding precision matrix should be sparse.

In the setting when the dimension of the random variable p is relatively small as compared to the sample size n, many different procedures for model selection for the Gaussian precision graph models have been proposed (Dempster, 1972; Edwards, 2000; Drton and Perlman, 2003). The standard approach as described in Edwards (2000) is backward stepwise selection. However, as noted by Drton and Perlman (2003), the overall error rate for the stepwise procedure is not controlled. Drton and Perlman (2003) further developed a method for calculating simultaneous p-values for all pairs and partitioning these p-values into a significant (S) set, an intermediate (I) set and a non-significance (N) set. This procedure, called the SINful approach, controls the overall error rate for incorrect edge inclusion. All of these methods work well when p is small. When p is large relative to the sample size, the method of Drton and Perlman (2003) relies on the inverse of the sample covariance matrix, which could be too conservative. Moreover, the inverse of the sample covariance matrix is not unique in the case when $n < p$. In addition, none of these procedures take into account the potential sparsity of the precision matrix in the estimation step. As the number of genes increases, reliable estimates of the conditional independencies or the precision matrix require many more observations than are usually available from gene expression profiling. However, incorporating the sparse nature of the graphs can help improve the estimate of the precision matrix and therefore improve inferences of the Gaussian concentration graph structure based on such estimates, for both the cases when $p > n$ and when $p < n$.

There are several approaches in the literature to covariance selection problems in the context of microarray data analysis. Schafer and Strimmer (2005) proposed a naive approach to estimate the precision matrix by using a boosted G-inverse, then determine which off-diagonal elements are zero by a thresholding and false discovery procedure. The drawback of this approach is that the sparsity is not accounted for when estimating the precision matrix, so the procedure is expected to perform poorly. Meinshausen and Buhlmann (2006) proposed a gene-by-gene approach by using the least absolute shrinkage and selection operator (lasso) (Tibshirani, 1996) to find neighbors for each gene. Under a large set of assumptions, they showed that the neighbors can be consistently identified when the sample size goes to infinity, which is very rare for microarray gene expression data. Dobra *et al.* (2004) proposed a Bayesian approach by converting the dependency networks into compositional networks using the Cholesky decomposition. The graphs are then used to estimate the precision matrix. Since Cholesky decomposition of the precision matrix naturally imposes ordering restriction of the variables, the procedure is computationally quite intensive since it has to determine gene order in their model construction. Finally, Wille *et al.* (2004) proposed to

infer Gaussian graphs based on tri-graphs by considering all partial correlations conditioning on only one other variable. Strictly speaking, the resulting tri-graphs are not true Gaussian concentration graphs.

In this paper, we introduce a TGD regularization procedure (Friedman and Popescu, 2004) for penalized estimation of a sparse precision matrix in the setting of Gaussian graphical models and demonstrate its application to identifying genetic networks based on gene expression data. Such a regularization procedure aims to account for the sparsity of the precision matrix in the estimation stage. The procedure does not depend on the Cholesky decomposition as in Dobra *et al.* (2004) and therefore does not have to deal with the problem of ordering the variables in the Cholesky decomposition. After obtaining the estimate of the precision matrix, we propose to apply a bootstrap procedure to further identify the edges of the graph. When the sample size is larger than the dimensionality, we also introduce a procedure based on empirical Bayes thresholding (EBT) (Johnstone and Silverman, 2004) on the inverse of the sample covariance matrix. Through simulations and application to real data sets, we demonstrate that this procedure is computationally feasible for both large and small sample cases and provides biologically meaningful results.

The rest of the paper is organized as follows: we first briefly review the Gaussian concentration graphical models. We then present an EBT procedure and a TGD procedure for estimating the sparse precision matrix. Following the methods, we present simulation results and an application for inference of a gene network for isoprenoid biosynthesis pathways in *Arabidopsis thaliana*. Finally, we briefly discuss the methods and results and provide possible further extensions of the method.

2. GAUSSIAN GRAPHICAL MODELS

We assume that the gene expression data observed are randomly sampled observational or experimental data from a multivariate normal probability model. Specifically, let X be a random normal p-dimensional vector and $X_1, \ldots, X_p$ denote the p elements, where p is the number of genes. Let $V = \{1, \ldots, p\}$ be the set of nodes (genes), and $X^{(k)}$ be the vector of gene expression levels for the kth sample. We assume that

$$X \sim N_p(0, \Sigma) \tag{2.1}$$

with positive definite variance–covariance matrix $\Sigma = \{\sigma_{ij}\}$ and precision matrix $\Omega = \Sigma^{-1} = \{\omega_{ij}\}$. This model can also be summarized as a graph model. Let $G = (V, E)$ be an undirected graph with vertex set $V = \{1, \ldots, p\}$ and edge set $E = \{e_{ij}\}$, where $e_{ij} = 1$ or 0 according to whether vertices i and j, $1 \leqslant i < j \leqslant p$, are adjacent in G or not. The Gaussian graphical model consists of all p-variate normal distributions $N_p(0, \Sigma)$, where Σ is unknown but where the precision matrix satisfies the following linear restrictions:

$$e_{ij} = 0 \Rightarrow \omega_{ij} = 0.$$

This model is also called a covariance selection model (Dempster, 1972) or a Gaussian concentration graph model.

Let $[-i]$ denote the set $\{1, 2, \ldots, i-1, i+1, \ldots, p\}$. In the Gaussian graphical model, it is well-known that the partial regression coefficients of X_i on X_j in the normal linear regression $p(X_i|X_{[-i]})$ is $-\omega_{ij}/\omega_{ii}$, $j \in [-i]$, and the ijth partial correlation between the ith and the jth gene is $\rho_{ij} = -\omega_{ij}/\sqrt{\omega_{ii}\omega_{jj}}$. For a given gene g, we define the neighbor of this gene as

$$ne_g = \{j : \omega_{gj} \neq 0, j \in [-g]\},$$

which contains all the genes with a non-zero partial correlation with the gene g. From the multivariate normal distribution theory, we have the following conditional independence result,

$$X_g \perp X_{G\setminus(ne_g \cup g)}|X_{ne_g}.$$

3. EBT AND THRESHOLD GRADIENT DESCENT REGULARIZATION

We consider the estimation of the precision matrix Ω based on a sample of $i.i.d.$ observations $X^{(k)} \in R^p$, $k \in N = \{1, \ldots, n\}$, where the set N can be interpreted as indexing the samples on which we observe the variables in V and $X^{(k)}$ is the kth observation. When the sample size is larger than the number of variables, we first propose a procedure based on sample covariance matrix and EBT (Johnstone and Silverman, 2004). We then propose to develop a penalized procedure for estimating Ω using the idea of TGD (Friedman and Popescu, 2004) to take into account the sparse nature of the precision matrix for genetic networks. After obtaining the estimate of the precision matrix, we propose to use a bootstrap procedure to further select the edges of a graph.

3.1 *Estimation based on EBT when $n > p$*

When $p < n$, the maximum likelihood estimate (MLE) of the precision matrix, denoted by $\hat{\Omega}$, is simply of the inverse of the sample covariance matrix. However, such an MLE is expected to include many small values and therefore cannot be used directly to select edges of a graph. We propose to apply the EBT procedure proposed in Johnstone and Silverman (2004) on $\hat{\Omega}$ in order to select the edges of a graph and call this the MLE–EBT procedure. Specifically, starting from the MLE of Ω, and denoting its elements as $\hat{\omega}_{ij}$, we calculate the estimate of the partial correlation matrix

$$\hat{\rho}_{ij} = \frac{-\hat{\omega}_{ij}}{\sqrt{\hat{\omega}_{ii}\hat{\omega}_{jj}}}.$$

We then perform Fisher's Z-transformation on all the partial correlations and denote the Z-transformed partial correlation as z_{ij}, i.e.

$$z_{ij} = \frac{1}{2}\log\frac{1+\hat{\rho}_{ij}}{1-\hat{\rho}_{ij}}.$$

Following Johnstone and Silverman (2004), we assume the following model for z_{ij}:

$$z_{ij} = \xi_{ij} + \epsilon_{ij}, \epsilon_{ij} \sim N(0, \sigma^2),$$

where ξ_{ij} is the Z-transformation of the true partial correlation ρ_{ij}, σ^2 is the error variance, and the elements ξ_{ij} have a mixture of 0 and Laplace distribution,

$$f_{\text{prior}}(\xi) = (1-w)\delta_0(\xi) + w\,\text{Laplace}(\xi),$$

where w is the mixture probability and $\delta_0(\xi)$ is the density with mass one at zero. From this model, one can derive the posterior distribution of ξ_{ij}. Johnstone and Silverman (2004) suggested to threshold the values of z_{ij} by the posterior median of ξ_{ij} and they showed that the resulting estimate of ξ_{ij} is uniformly bounded over all signals,

$$\sup \frac{2}{p(p-1)} \sum_{ij} E|\hat{\xi}_{ij} - \xi_{ij}|^r \leqslant C_0, \quad 0 < r \leqslant 2,$$

for some constant C_0.

After the EBT, we would expect that many of the elements of the precision matrix with very small values of the partial correlations are thresholded to zero, corresponding to no edges of the Gaussian graph. This MLE–EBT approach is similar in spirit to that in Schafer and Strimmer (2005) in the settings when $p < n$.

3.2 *Regularized estimation by TGD on the off-diagonal elements*

The MLE–EBT procedure proposed above only applies when $p < n$. Even in this case, the sparse nature of the precision matrix is not accounted for in the MLE of $\mathbf{\Omega}$. In order to utilize the sparse property of the precision matrix, we propose in this section to maximize the likelihood function based on model (2.1), subject to constraint by 'sparse' precision matrix $\mathbf{\Omega}$. Let $\omega^d \equiv \{\omega_{11}, \ldots, \omega_{pp}\}$ denote the vector of the diagonal elements of the matrix $\mathbf{\Omega}$ and $\omega^o \equiv \{\omega_{ij}\}_{i \neq j}$ denote the vector of $q = p(p-1)/2$ off-diagonal elements of the $\mathbf{\Omega}$ matrix. The likelihood function can be written as

$$w(\omega^d, \omega^o) = \frac{n}{2} \log |\mathbf{\Omega}| - \frac{1}{2} \sum_{k=1}^{n} X^{(k)'} \mathbf{\Omega} X^{(k)}, \tag{3.1}$$

where $X^{(k)}$ is the kth observation. We assume that the variables are standardized. When $p < n$, the MLE of $\mathbf{\Omega}$ is simply the inverse of the sample covariance matrix, and when $n < p$, the MLE of $\mathbf{\Omega}$ is not unique.

In order to account for the sparsity of the precision matrix $\mathbf{\Omega}$, we define a loss function as the negative of the log likelihood function (3.1),

$$l(\omega^d, \omega^o) = -w(\omega^d, \omega^o).$$

Based on equation (3.1), the gradient of the loss function with respect to $\mathbf{\Omega}$ is

$$\frac{\partial l}{\partial \mathbf{\Omega}} = \frac{n}{2} \mathbf{\Omega}^{-1} - \frac{1}{2} \sum_{k=1}^{n} X^{(k)} X^{(k)'}. \tag{3.2}$$

From this we can obtain the gradient of the loss function over the off-diagonal elements ω^o. Define $g(\omega^o) = (g_1(\omega^o), \ldots, g_q(\omega^o)) = -\nabla_{\omega^o} l(\omega^o, \omega^d)$ to be the negative gradient of l with respect to ω^o. To find an optimal path from all the paths from $\mathbf{\Omega} = I$ to the MLE of $\mathbf{\Omega}$ or to a precision matrix surface formed by $\mathbf{\Omega} = S^-$ when $p > n$, we start from $\nu = 0$, $\omega^o = (0, \ldots, 0)$, and $\omega^d = (1, \ldots, 1)$ and update the elements ω^o by the following gradient descent step,

$$\hat{\omega}^o(\nu + \Delta\nu) = \hat{\omega}^o(\nu) + \Delta\nu h(\nu),$$

where $\hat{\omega}^o(\nu)$ is the ω^o value corresponding to current ν, $\Delta\nu > 0$ is an infinitesimal increment, and $h(\nu)$ is the direction in the parameter space tangent to the path evaluated at $\hat{\omega^o}(\nu)$. This tangent vector at each step represents a descent direction. In order to direct the path toward parameter points with diverse values, following Friedman and Popescu (2004), we define $h(\nu)$ as

$$h(\nu) = \{ f_j(\nu) \cdot g_j(\nu),\ j = 1, \ldots, q \},$$

where

$$f_j(\nu) = I[|g_j(\nu)| \geqslant \tau \cdot \max_{1 \leqslant k \leqslant q} |g_k(\nu)|],$$

where $I[\cdot]$ is an indicator function, and $0 \leqslant \tau \leqslant 1$ is a threshold parameter that regulates the diversity of the values of $f_j(\nu)$; larger values of τ lead to more diversity. $g(\nu)$ is the negative gradient evaluated at $\hat{\omega}^o(\nu)$ and current ω^d. Therefore, τ is the parameter which controls the degree of penalty and sparsity in the ω^o, with $\tau = 1$ giving the sparsest graphs. Instead of moving along the true gradient direction, the threshold gradient update only moves along those elements with large values of the gradient. After ω^o is updated, we update the diagonal elements of $\mathbf{\Omega}$, ω^d, by maximizing the log-likelihood function (3.1) with ω^o fixed at the current values, $\hat{\omega}^o$. This is done by using Newton–Raphson iterations.

In summary, for any threshold value $0 \leqslant \tau \leqslant 1$, the TGD regularization algorithm for the sparse Gaussian graphical model involves the following six steps:

1. Set $\omega^o(0) = 0$, $\omega^d(0) = 1$, $\nu = 0$.
2. Calculate $g(\nu) = -\partial l / \partial \omega^o$ for the current ω^o and ω^d.

3. Calculate $f_j(\nu) = I[|g_j(\nu)| \geqslant \tau \cdot \max_{1\leqslant k\leqslant q}|g_k(\nu)|]$ and $h(\nu)$.
4. Update $\omega^o(\nu + \triangle\nu) = \omega^o(\nu) + \triangle\nu \cdot h(\nu)$, $\nu = \nu + \triangle\nu$.
5. Update parameters ω^d by maximizing the log-likelihood using Newton–Raphson iterations with ω^o fixed at $\omega^o(\nu + \Delta\nu)$.
6. Repeat steps 2–5.

For a given τ, it is easy to see that the likelihood function increases as the iterations increase, and different τ correspond to different paths for Ω from I to S^-. It should be emphasized that for a given τ, the threshold gradient iterations stop before it reaches S^- and the number of gradient iterations at which to stop the algorithm can be determined by cross-validation (see Section 3.3). In this paper, we only consider the algorithm with $\tau = 1$, which corresponds to the sparsest graph for a given TGD step, and call the proposed procedure the direct TGD procedure. Such a procedure is expected to perform better for gene expression data since most biological or genetic networks are expected to be vary sparse (Barabasi and Oltvai, 2004).

3.3 *Model selection by cross-validation and bootstrap*

As the iterations continue, more and more non-zero elements are selected in the precision matrix and the corresponding undirected graphs grow larger. The final model should provide the best balance between coverage (correctly identified connections/total true connections) and false positives (incorrectly identified connections/total identified connections) (Gardner *et al.*, 2003). We propose to use K-fold cross-validation for choosing the number of TGD iterations, ν, where for each ν, the K-fold cross-validated log-likelihood criterion is defined as

$$\mathrm{CV}(\nu) = \frac{1}{K}\sum_{k=1}^{K}\left(-n_k \log|\Omega_{-k}| + \sum_{i\in V_k} X^{(i)}\Omega X^{(i)}\right),$$

where n_k is the size of the kth validation set V_k and Ω_{-k} is the TGD estimate of the precision matrix based on sample $V\backslash V_k$ evaluated at $\hat{\Omega}(\nu)$. Alternatively, we can use the Bayesian information criterion (BIC) criteria for selecting ν, where the degrees of freedom can be defined as the number of non-zero entries of the off-diagonal elements of the precision matrix. This is similar in spirit to the lasso in linear regression where the degrees of freedom is defined as the number of non-zero coefficients (Zou *et al.*, 2004).

Since the number of the off-diagonal elements in the precision matrix is often quite large compared to the sample size, there is often considerable uncertainty in the edges chosen. As a final step in the procedure, we propose to use the bootstrap method to determine the statistical accuracy and the importance of each of the edges identified by the TGD procedure. In bootstrapping, B bootstrap data sets, $X^{*1}, \ldots, X^{*B}$, are sampled with replacement from the original data set such that each bootstrap sample contains n observations. We then apply the TGD procedure to each bootstrap data set and examine which edges are in the final models. One can then choose only the edges with high probability of being non-zero in the precision matrix over the bootstrap samples.

4. Simulations

We performed simulations to investigate how well the proposed threshold gradient procedure estimates the precision matrix and to compare the new estimate with the inverse of the sample covariance matrix, i.e. the maximum likelihood estimator of the precision matrix. We also evaluated the performance of the proposed procedure in the case when $p > n$.

4.1 *Estimation when $p < n$*

We consider Gaussian precision graphs with 40 nodes and the following four precision matrices (Ω) with different degrees of sparsity:

1. Very sparse precision matrix (Ω_1): the numbers of true neighbors or edges for each gene range from one to four.
2. Sparse precision matrix (Ω_2): the numbers of true neighbors or edges for each gene range from five to nine.
3. Less sparse precision matrix (Ω_3): the numbers of true neighbors or edges for each gene range from 8 to 14.
4. Dense precision matrix (Ω_4): the numbers of true neighbors or edges for each gene range from 14 to 35.

Specifically, we generate 40 points randomly on a $[0, 1] \times [0, 1]$ space and then calculate all the pairwise distances between the points. For each point (corresponding to one gene), define the k neighbors as those with k smallest distances to this gene. By choosing different numbers of k, we can obtain graphs for models 1–4 with different degrees of sparsity. For the pairs with edges, the corresponding elements in the precision matrix are first generated from uniform distribution between 0.5 and 1 or between -1 and -0.5. For each row, the diagonal element is defined as a factor of the sum of the absolute values of the elements of the given row. Finally, each row is divided by the corresponding diagonal element so that the final precision matrix has diagonal elements of 1 and is positive definite. The factors chosen are 2, 1, 0.8, and 0.5 for models 1–4 to ensure that the precision matrices are positive definite and the final partial correlations are in similar ranges for all four models. See Figure 1 for a heat map plot of the partial correlation used for simulations for the four different models. It is clear that the precision matrix gets denser from model 1 to model 4. The actual ranges of the partial correlations range between -0.2 and 0.2 for most values (see plots in the left panel of Figure 2).

For each of the precision matrices Ω, we simulated 120 $i.i.d.$ $\mathrm{N}(0, \Omega^{-1})$ 40-dimensional vectors, i.e. sample sizes of 120 and dimensionality of 40. For each simulated data set, we estimated the precision matrix using the proposed TGD, TGD–EBT, MLE, and MLE–EBT procedures. For the TGD procedure, 10-fold cross-validation was used for choosing the TGD iteration step. We repeate this procedure 100 times and computed the estimates of the following two loss functions:

$$l(\Omega, \hat{\Omega}) = \operatorname{tr} \Omega^{-1}\hat{\Omega} - \log|\Omega^{-1}\hat{\Omega}| - n,$$

$$l_2(\Omega, \hat{\Omega}) = \operatorname{tr}(\Omega^{-1}\hat{\Omega} - I)^2,$$

where $\hat{\Omega}$ is an estimate of Ω. The first loss is called entropy loss and the second loss is called quadratic loss (Lin and Perlman, 1985).

Figure 2 presents the box plots of the two loss functions for the four different estimation procedures based on 100 simulation runs. The estimators we considered include the inverse of the sample covariance matrix (i.e. the MLE), the EBT estimator using the MLE of the precision matrix (MLE–EBT), the proposed TGD estimate, and the TGD estimate with EBT. Overall, we observed that the TGD estimate with or without further EBT outperformed the MLE and MLE–EBT procedures in all cases for both loss criteria and the improvements are substantial over MLE. In addition, the more sparse the precision matrix is, the greater gain in risk reduction we have by using the proposed TGD-based estimates. Also, although further EBT on MLEs, indeed, reduces the loss greatly, further EBT on the TGD estimates actually results in increase in loss. One reason for such increases in loss is that the off-diagonal elements of the estimated precision matrix are already very sparse and the assumptions of the EBT procedure of Johnstone and Silverman (2004) may not hold. Based on this observation, the EBT will not be applied to TGD estimates of the precision matrix in the following analyses.

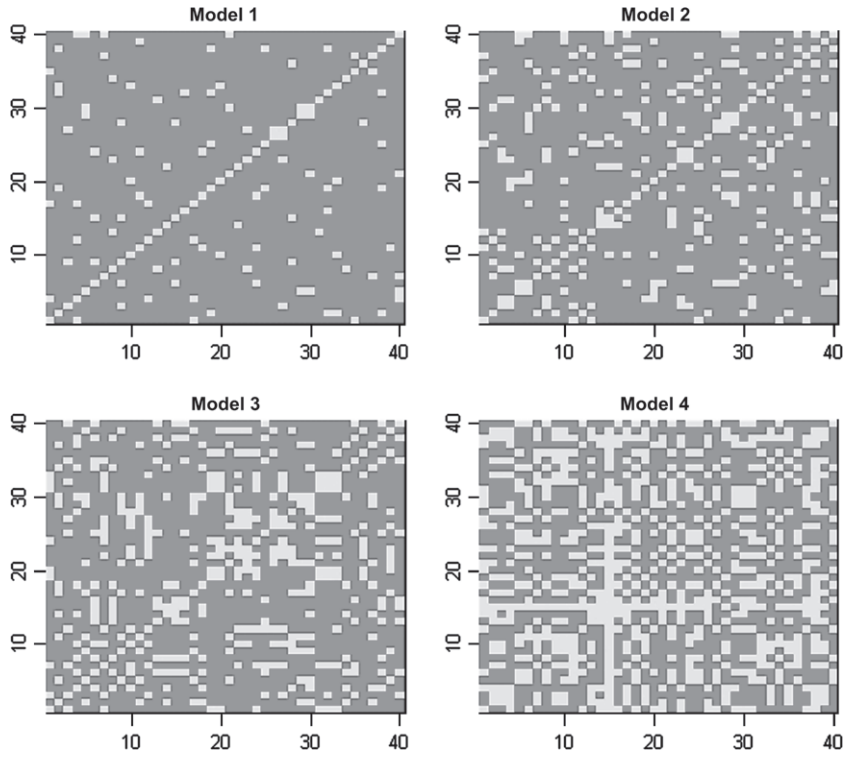

Fig. 1. Heat map of the true partial correlation matrix for models 1–4 used in the simulations, where model 1 corresponds to a sparse graph and model 4 corresponds to a dense graph. For each plot, the white dots indicate the non-zero partial correlation between two genes and the two axes index the genes.

To further demonstrate that the TGD procedure gives sensible estimates of the partial correlations, we plot in the left panel of Figure 3 the estimated partial correlations versus the true partial correlations for all the true edges (i.e. those pairs with non-zero true partial correlations) and the estimated partial correlations for those pairs with zero partial correlations (plots on the right panel). Clearly, the estimates of the partial correlations from TGD correlate quite well with the true values for gene pairs with edges. In comparison, the estimates are zeros for most of the conditionally independent pairs, especially for the case when the graph is very sparse (model 1). In addition, as expected for estimates based on any regularized procedure, we note that the estimates of the partial correlations from TGD are in general smaller than the true values. In other words, the TGD procedure shrinks the estimates of partial correlations toward zero.

4.2 *Estimation when $p > n$*

Finally, we demonstrate that the proposed procedure is computationally feasible and provide sensible results even when $p > n$. We simulated sparse graphs with $n = 100$ and $p = 200$. A similar procedure was used for generating the precision matrix, and the resulting 286 non-zero off-diagonal elements range from -0.56 to 0.48 with most values between -0.2 and 0.2.

In order to assess how the results change as the TGD iterations go, we plot in Figure 4 the sensitivity, specificity, and false-negative and false-positive (or false discovery) rates as a function of the TGD interaction step, for a total of 10 000 steps ($\Delta\nu$ is taken to be 10^{-4}), where the sensitivity is defined as

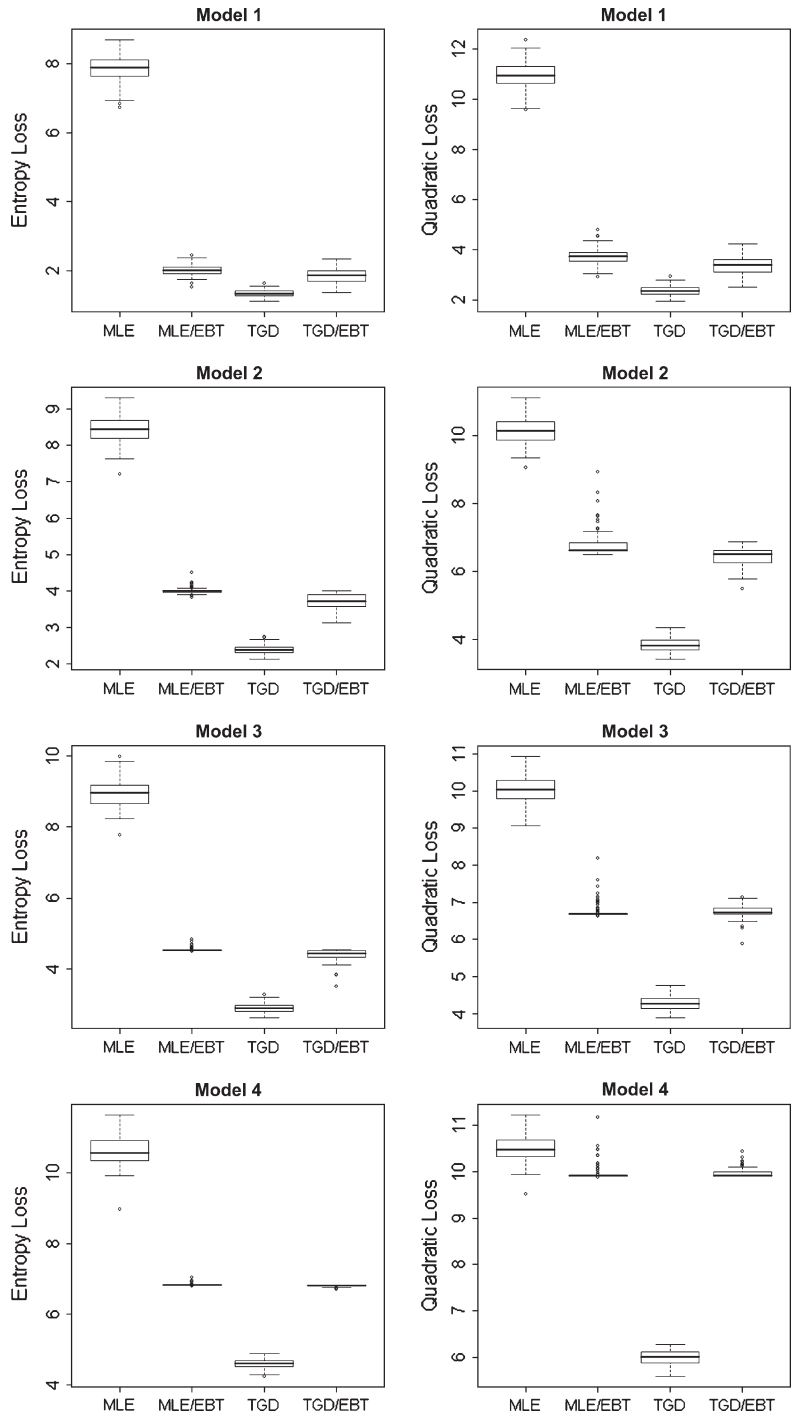

Fig. 2. Box plot of the loss functions based on 100 replications, where the left panel represents the entropy loss and the right panel represents the quadratic loss.

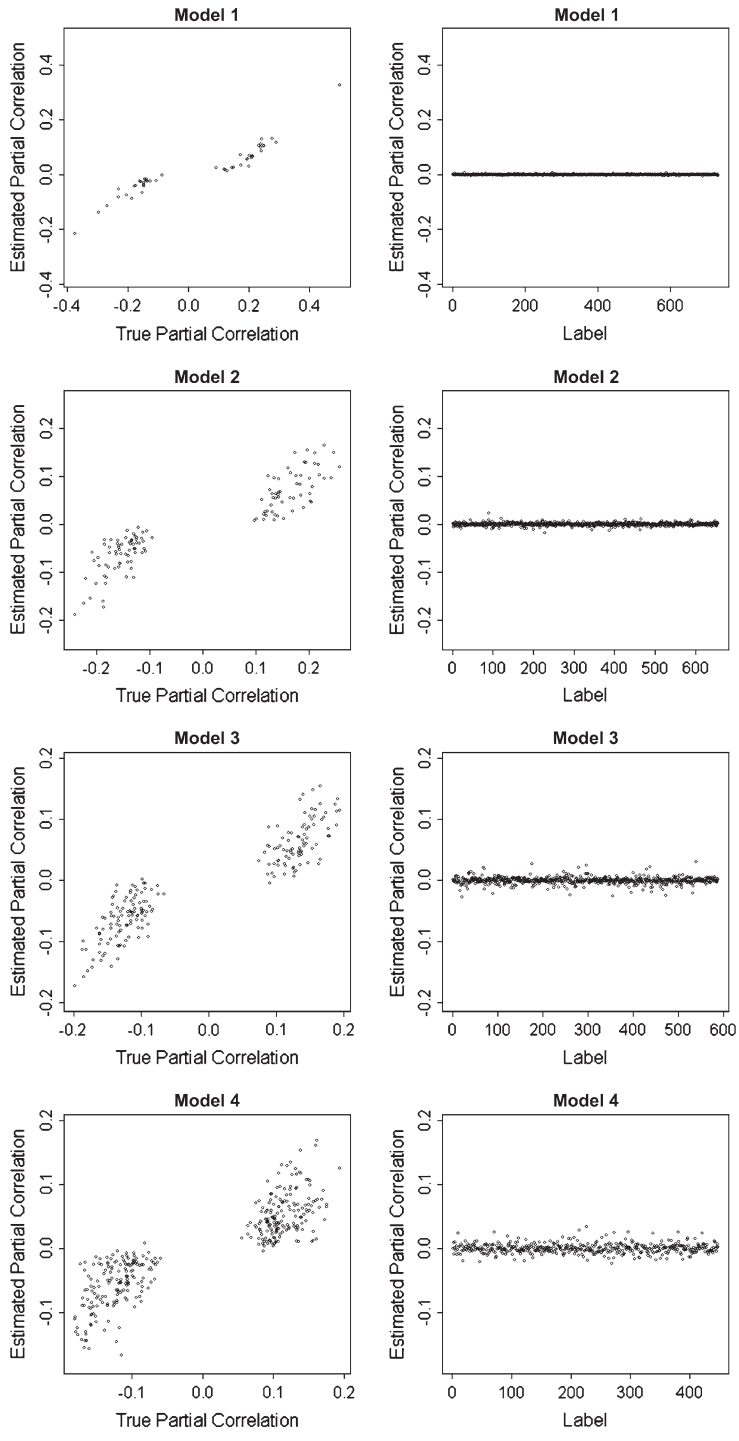

Fig. 3. Scatter plots of true against the estimated partial correlations for models 1–4 over 100 replications. The left panel represents the true edges and the right panel represents those with zero partial correlations.

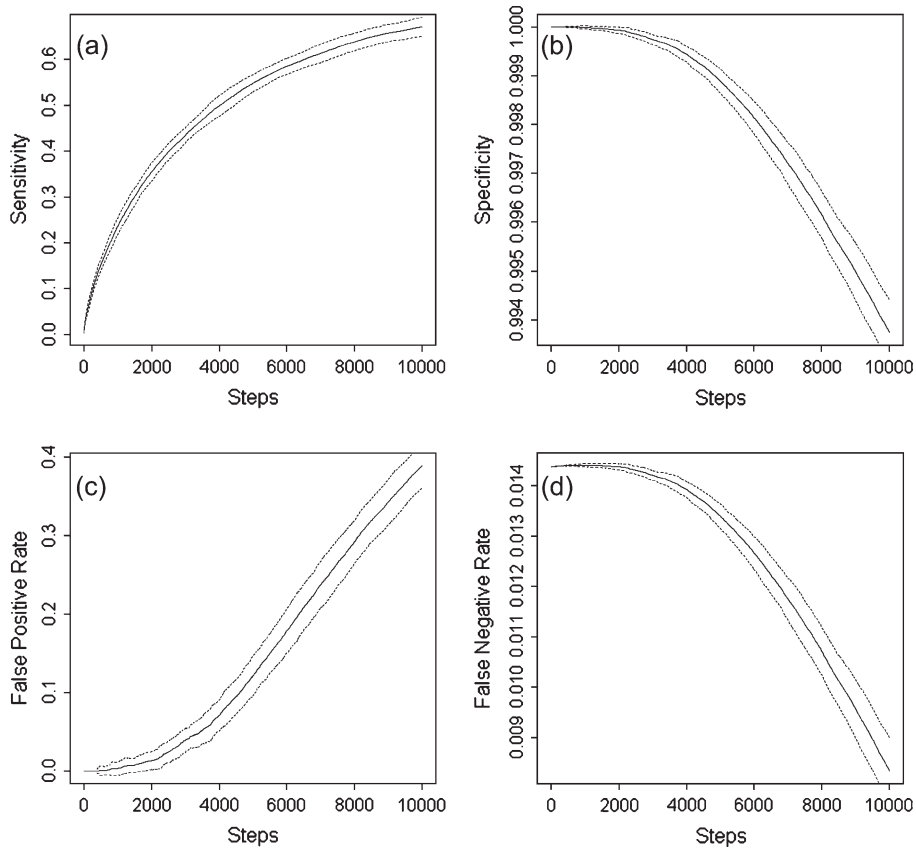

Fig. 4. Results based on simulation for Gaussian graphs with $p = 200$ and sample size of $n = 100$. For each plot, the x-axis is the TGD step, and the y-axis is sensitivity (a), specificity (b), false discovery rate (c), and false-negative rate. The dashed lines are ± 1 SE based on 50 replications.

the proportion of the identified edges as being the true edges and the false discovery rate is defined as the proportion of wrong edges among those identified by the TGD procedure. Specificity and false-negative rates are similarly defined. First, as expected for very sparse graphs, we observe that the TGD procedure results in very high specificity and very low false-negative rate, and both rates decrease as iterations go. On the other hand, as the TGD iterations go, both the sensitivity and false discovery rate increase. For example, for a false discovery rate of 20%, the sensitivity is about 60%, and for a discovery rate of 30%, the sensitivity increases to about 65%. Based on BIC criteria, treating the number of non-zero off-diagonal elements as the number of effective parameters, the algorithm stops at about the 8000th TGD step, which corresponds to a sensitivity of about 65%. This example demonstrates that the TGD procedure behaves well even when $p > n$.

5. APPLICATIONS TO ISOPRENOID PATHWAYS IN *A. Thaliana*

The isoprenoid biosynthetic pathway provides intermediates of many natural products including steroles, chlorophylls, carotenoids, plastoquinone, and abscisic acid. It is now known that plants contain two pathways for the synthesis of the structural precursors of isoprenoids: the mevalonate (MVA) pathway, located

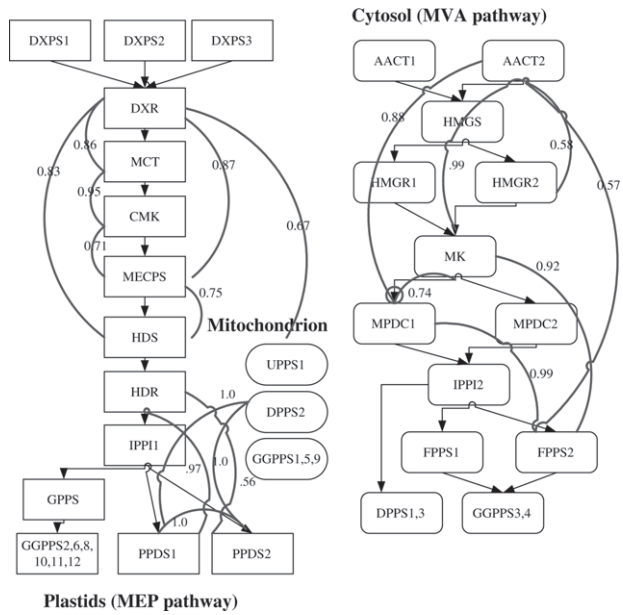

Fig. 5. Pathways identified by the TGD method for the 40 genes in the isoprenoid pathways, where the solid arrows are the true pathways and the curved undirected lines are the estimated edges with bootstrap probability of greater than 0.5 for the TGD method. For this plot, the left panel represents a subgraph of the gene module in the MEP pathway and the right panel represents a subgraph of the gene module in the MVA pathway. The numbers on the estimated edges are the bootstrap probabilities.

in the cytosol/endoplasmic reticulum, and the recently discovered methylerythritol 4-phosphate (MEP) pathway, located in the plastids. The pathway in plastids, which is MVA independent, occurs and is responsible for the subsequent biosynthesis of plastidial terpenoids such as carotenoids and the side chains of chlorophyll and plastoquinone (Wille *et al.*, 2004). It is therefore important to understand the organization and regulation of this complex metabolic pathway, with the long-term goal of using the generated knowledge to undertake metabolic engineering strategies oriented to increase the production of isoprenoids with pharmaceutical and food applications, and also to the design and development of new antibiotics.

In order to better understand the pathway and gain insights into the cross-link between the two pathways at the transcriptional level, Wille *et al.* (2004) reported a data set including the gene expression patterns monitored under various experimental conditions using 118 GeneChip microarrays. For the construction of the genetic network, they focused on 40 genes, 16 of which were assigned to the cytosolic MVA pathway, 19 assigned to the plastidal MEP pathway, and five genes encoding proteins located in the mitochondria. See the solid lines of Figure 5 for the MVA and the MEP pathways and the genes involved.

5.1 *Results from the TGD procedure*

In order to demonstrate whether the proposed TGD method can identify the known isoprenoid pathways of these 40 genes based on the 118 gene expression measurements, we first estimated the precision matrix by the threshold gradient methods. Using 10-fold cross-validation, the TGD procedure resulted in 20 non-zero off-diagonal elements. We next used a bootstrap with the TGD procedure to estimate the confidence of the edges. With bootstrap probability of 0.50 or higher, we identified 19 pairs of genes

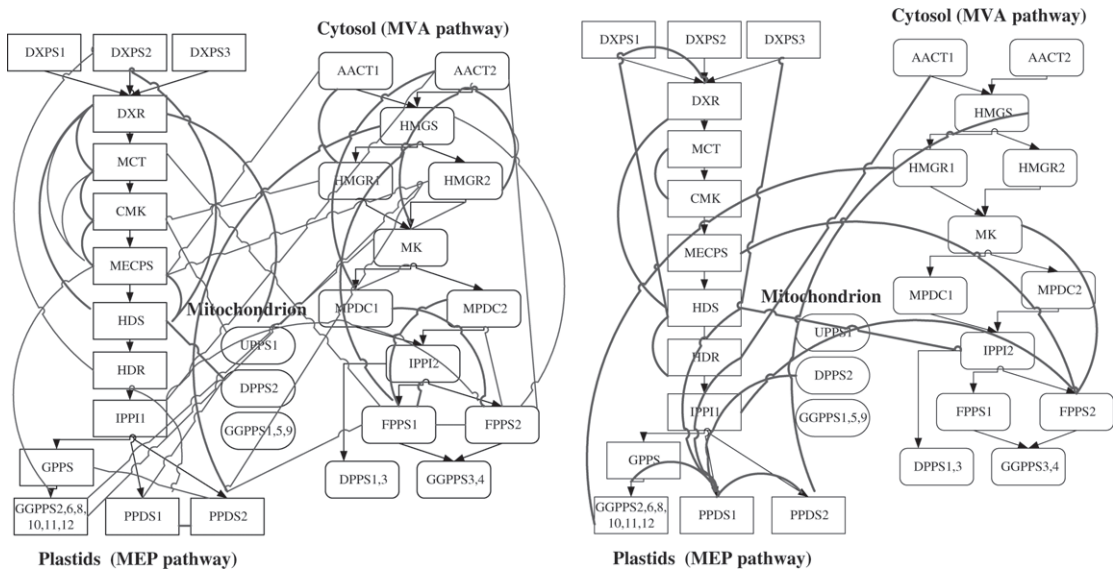

Fig. 6. Pathways identified by the tri-graph method by Wille *et al.* (2004) (left plot) and the SINful approach with cutoff p-value of 0.50 (right plot) for the 40 genes in the isoprenoid pathways, where the solid arrows are the true pathways and the curved undirected lines are the estimated edges. For each plot, the left pane includes a subgraph of the gene module in the MEP pathway and the right panel includes a subgraph of the gene module in the MVA pathway.

which are connected with high confidence, of which 12 pairs have a bootstrap probability of 0.80 or higher. These 19 pairs are plotted on the true network in Figure 5. We find a module with strongly interconnected genes in each of the two pathways. For the MEP pathway, 1-deoxyxylulose-5-phosphate synthase (DXPS), 1-deoxyxylulose-5-phosphate-reductoisomerase (DXR), 2-*C*-methylerythritol-4-phosphate cytidyltransferase (MCT), 4-(cytidine-5′-diphospho)-2-*C*-methylerythritol kinase (CMK), and 2-*C*-methylerythritol-2,4-cyclodiphosphate synthase (MECPS) are connected as the true pathway. Similarly, the genes in the MVA pathways, acetyl-CoA/acetyl-CoA *C*-acetyltransferase (AACT), 3-hydroxy-3-methylglutaryl-CoA reductase (HMGR), mevalonate kinase (MK), mevalonate 5-diphosphate decarboxylase (MPDC), and Farnesyl diphosphate synthase (FPPS) are closely connected. In addition, there are also several genes in the MEP pathway which are linked to proteins in the mitochondria.

It is interesting to note that although both the TGD method and the tri-graph method of Wille *et al.* (2004) identified two closely connected genetic modules (see left plot of Figure 6), the method based on tri-graph seems to include many more edges for each module and more cross-links between the two pathways. While there is some evidence of cross-links between the two pathways, one should not expect that the two pathways are so closely linked since genes of the two pathways belong to two different cell compartments. One possible explanation of such a difference is that the tri-graph conditions only on one other gene at each calculation and therefore cannot capture multigene effects when considering the partial correlations for a given pair of genes.

5.2 *Comparison with other methods*

As a comparison, we applied the SINful procedure using the inverse of the sample covariance matrix and the MLE–EBT procedure to the same data set. If we used p-value less than 0.10 for the SINful procedure, we only identify 11 edges, all from the MEP pathway, and none of the edges in the MVA pathways were

identified by either of the two methods. Even if the p-value is set to 0.50, many false edges between MEP and MVA pathways are identified and even the tightly connected DXR–HDS [1-hydroxy-2-methyl-2-(*E*)-butenyl-4-diphosphate synthase] module cannot be identified (see right plot of Figure 6). Similarly, the MLE–EBT procedure also only identified a few edges and failed to identify the DXR–HDS module (not shown).

6. Discussion

We have proposed a TGD-based regularization procedure for performing penalized estimation for the Gaussian graph models in order to account for the sparsity of the precision matrix. Such a procedure is computationally feasible even when the number of variables is greater than the sample size. We have demonstrated the method by simulations and application to identify the *A. thaliana* isoprenoid pathways based on 40 genes and 118 experimental conditions. The results indicate that empirically defined associations based on the sparse Gaussian graphs, indeed, link to functional activity in isoprenoid metabolic pathways and many key biological interactions in the isoprenoid metabolic pathways are captured by graphs constructed by our method. However, biologically speaking, it is important to keep in mind that the Gaussian concentration graphs built based on our proposed method should properly be considered as coexpression or coregulation networks and not as genetic regulatory networks *per se*.

We have demonstrated the importance of accounting for sparsity of the precision matrix in the estimation stage, even in cases when the sample covariance matrix is invertible and the sample precision matrix is unique. As clearly demonstrated by our simulations, by accounting for the sparsity in the estimation stage, the estimate of the precision matrix is closer to the true matrix than the naive method of inverting the sample covariance matrix. This is in contrast to the procedure proposed by Schafer and Strimmer (2005) when such sparsity is not accounted for in the estimation stage. Our simulation also indicates that the TGD procedure has no computational difficulty in high-dimensional settings when $p > n$ for p in the order of hundreds. When p is very large, the major computational burden is on updating the diagonal elements when the off-diagonal elements are known and are sparse during the TGD iterations. In this paper, we simply used Newton–Raphson iteration for updating the diagonal elements. For $p = 200$ and $n = 100$, it took about 70 min to finish 10 000 TGD iterations on a desktop personal computer using R (3.2 GHz and 1.0 G RAM). More efficient computation may be developed to fully utilize the sparsity of the off-diagonal elements. Alternatively, one may only perform one-step Newton–Raphson updates during each of the TGD steps. This deserves further investigation.

For sparse graphs, it is expected that the TGD procedure should give high specificity and low false-negative rates. For the settings when $p > n$, one crucial step of the proposed TGD algorithm is to decide when to stop the TGD iterations. In this paper, we used cross-validation based on the likelihood and found that the algorithms tend to stop late and therefore result in relatively high false-positive rate and, of course, also high sensitivity. On the other hand, the BIC criteria treating the number of non-zero off-diagonal elements as the degrees of freedom often stop the iterations early and result in low false-positive rates and also low sensitivity. How to stop the TGD iteration to obtain an optimal trade-off between sensitivity and false discovery rate deserves further investigation. Some biological knowledge about the networks can help. An alternative is to choose the graphs based on the power law of the numbers of the neighbors which were often observed for biological or genetic networks (Barabasi and Oltvai, 2004).

An important area of future research is to improve the networks identified by the TGD–EBT procedure by incorporating other biological information such as gene ontology or known biochemical pathways and to develop methods that allow for non-linear relationships among the variables. One way of extending the proposed method in order to incorporate prior known pathways information is to perform gradient-based thresholding only on elements with uncertain edges. Suppose that we have known a certain genetic pathway or network involving a set of p_s genes, denoted by V_s. The prior knowledge about the underlying

genetic or biological networks can be rephrased as certain edges in the graphs involving these p_s genes are definitely present. In order to ensure that the known edges are included in the graphs identified, we can modify the TGD algorithm by setting thresholds on the negative gradients of the elements only in the precision matrix that correspond to uncertain edges to ensure that the edges corresponding to the known pathways have non-zero partial correlations. Since the proposed method is mainly aimed to identify gene coexpression networks based on gene expression data, it would be interesting to extend the ideas in this paper for integrating various sources of genomic data, such as sequences and transcriptional factors binding data, with microarray gene expression data in order to obtain better understanding of complex genetic networks, including genetic transcriptional networks.

The TGD regularization method was originally developed by Friedman and Popescu (2004) in the context of classification and linear regression when sample size is large or the dimension of predictors is high and was further extended for the Cox regression (Gui and Li, 2005). To our knowledge, this paper is the first attempt to extend this procedure to estimate a sparse precision matrix in Gaussian graphical models. As indicated by Friedman and Popescu (2004), for linear regression and classification problems, the TGD procedure with $\tau = 1$ gives similar results as Tibshirani's lasso. However, for the estimation of sparse Gaussian models, our proposed TGD procedure is very different from the lasso approach proposed by Meinshausen and Buhlmann (2006), where they proposed to perform lasso for each gene using the rest of the genes as predictors in a linear regression setting. In contrast, our TGD procedure considers the sparse nature of the numbers of neighbors for all the genes simultaneously. The connection between these two procedures is not clear.

In conclusion, we have proposed a TGD regularization procedure for estimating sparse Gaussian precision models and have demonstrated its application in generating gene networks based on microarray gene expression data. This procedure will be quite useful in studying the associations among genes based on gene expression data.

ACKNOWLEDGMENTS

This research is supported by the National Institutes of Health grant ES009911 and a grant from the Pennsylvania Department of Health. We thank the two reviewers for many insightful comments, which greatly improved the contents of the paper and Edmund Weisberg, MS, for editorial help.

REFERENCES

BARABASI, A. L. AND OLTVAI, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics* **5**, 101–113.

DEMPSTER, A. P. (1972). Covariance selection. *Biometrics* **28**, 157–175.

DOBRA, A., JONES, B., HANS, C., NEVIS, J. AND WEST, M. (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis* **90**, 196–212.

DRTON, M. AND PERLMAN, M. D. (2003). A SINful approach to model selection for Gaussian precision graphs. *Technical Report*. University of Washington.

EDWARDS, D. (2000). *Introduction to Graphical Modelling*, 2nd edition. New York: Springer.

FRIEDMAN, N. (2004). Inferring cellular networks using probabilistic graphical models. *Science* **30**, 799–805.

FRIEDMAN, J. H. AND POPESCU, B. E. (2004). Gradient directed regularization. *Technical Report*. Stanford University.

GARDNER, T. S., DI BERNARDO, D., LORENZ, D. AND COLLINS, J. J. (2003). Inferring genetic networks and identifying compound mode of action via expression profiling. *Science* **301**, 102–105.

GUI, J. AND LI, J. (2005). Threshold gradient descent method for censored data regression, with applications in pharmacogenomics. *Pacific Symposium on Biocomputing* **10**, 272–283.

IDEKER, T., THORSSON, V., RANISH, J. A., CHRISTMAS, R., BUHLER, J., ENG, J. K., BUMGARNER, R., GOODLETT, D. R., AEBERSOLD, R., AND HOOD, L. (2001). Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* **292**, 929–934.

JEONG, H., MASON, S. P., BARABASI, A. L., AND OLTVAI, Z. N. (2001). Lethality and centrality in protein networks. *Nature* **411**, 41–42.

JOHNSTONE, I. M. AND SILVERMAN, B. W. (2004). Needles and hay in haystacks: empirical Bayes estimates of possibly sparse sequences. *Annals of Statistics* **32**, 1594–1649.

LIN, S. P. AND PERLMAN, M. D. (1985). A Monte Carlo comparison of four estimators of a covariance matrix. In Krishnaish, P. R. (ed), *Multivariate Analysis*, Volume 6. Amsterdam: North-Holland, pp. 411–429.

MEINSHAUSEN, N. AND BUHLMANN, P. (2006). Consistent neighbourhood selection for high-dimensional graphs with the lasso. *Annals of Statistics* (in press).

SCHAFER, J. AND STRIMMER, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21**, 754–764.

SEGAL, E., SHAPIRA, M., REGEV, A., PE'ER, D., BOTSTEIN, D., KOLLER, D., AND FRIEDMAN, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nature Genetics* **34**, 166–176.

TAVAZOIE, S., HUGHES, J. D., CAMPBELL, M. J., CHO, R. J. AND CHURCH, G. M. (1999). Systematic determination of genetic network architecture. *Nature Genetics* **22**, 281–285.

TEGNER, J., YEUNG, M. K., HASTY, J., AND COLLINS, J. J. (2003). Reverse engineering gene networks: integrating genetic perturbations with dynamical modeling. *Proceedings of the National Academy of Science of the United States of America* **100**, 5944–5949.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B* **58**, 267–288.

WILLE, A., ZIMMERMANN, P., VRANOVA, E., FURHOLZ, A., LAULE, O., BLEULER, S., HENNIG, L., PRELIC, A., VON ROHR, P., THIELE, L. *et al.* (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biology* **5**, 1–13.

ZOU, H., HASTIE, T. AND TINSHIRANI, R. (2004). On the "degrees of freedom" of the lasso. *Technical Report*. Department of Statistics, Stanford University.

[*Received May 31, 2005; revision November 23, 2005; accepted for publication November 29, 2005*]