Selection of the Regularization Parameter in Graphical Models using Network Characteristics

Citation for published version:
Caballe, A, Bochkina, N & Mayer, C 2018, 'Selection of the Regularization Parameter in Graphical Models using Network Characteristics' Journal of Computational and Graphical Statistics, vol. 27, no. 2, pp. 323-333. DOI: 10.1080/10618600.2017.1366910

Digital Object Identifier (DOI):
10.1080/10618600.2017.1366910

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Journal of Computational and Graphical Statistics

Selection of the Regularization Parameter in Graphical Models using Network Characteristics

Adria Caballe Mestres *[1,2], Natalia Bochkina†[1,3], and Claus Mayer‡[2]

[1]University of Edinburgh & Maxwell Institute, EH9 3FD, Scotland, UK.

[2]Biomathematics & Statistics Scotland, EH9 3FD, Scotland, UK.

[3]The Alan Turing Institute for Data Science, British Library, London, NW12DB, UK.

Abstract

Gaussian graphical models represent the underlying graph structure of conditional dependence between random variables which can be determined using their partial correlation or precision matrix. In a high-dimensional setting, the precision matrix is estimated using penalized likelihood by adding a penalization term which controls the amount of sparsity in the precision matrix and totally characterizes the complexity and structure of the graph. The most commonly used penalization term is the L1 norm of the precision matrix scaled by the regularization parameter which determines the trade-off between sparsity of the graph and fit to the data. In this paper we propose several procedures to select the regularization parameter in the estimation of graphical models that focus on recovering reliably the appropriate network structure of the graph. We conduct an extensive simulation study to show that the proposed methods produce useful results for different network topologies. The approaches are also applied in a high-dimensional case study of gene expression data with the aim to discover the genes relevant to colon cancer. Using this data, we find graph structures which are verified to display significant biological gene associations. **Note:** supplementary material is available online.

Keywords: sparse precision matrix, high dimension, clustering, gene expression, graphical lasso, hyperparameter estimation

*a.caballe@sms.ed.ac.uk

†n.bochkina@ed.ac.uk

‡c.mayer@abdn.ac.uk

1 Introduction

In recent years, the study of undirected graphical models (Lauritzen, 1996) has been the focus of attention of many authors. The increasing volume of high-dimensional data in different disciplines makes them a useful tool in order to determine conditional dependence between random variables. For instance, graphical models have been applied to gene expression data sets to find biological associations across genes in Dobra et al. (2004) and Schäfer and Strimmer (2005), as well as in other biological networks (Newman, 2003) and in social networks (Goldenberg, 2007). In Gaussian graphical models, which are often used for finding associations between genes using high throughput genomic data, the dependence between the genes is fully characterized by the non-zero elements of the precision matrix Ω (see Section 2.1).

However, in a high dimensional framework where the number of variables p is larger than the number of observations n, there is not enough information in the data available to estimate Ω, and hence the underlying conditional dependence (CD) graph. To address this problem, alternative estimators have been proposed in the last two decades using additional information about Ω such that the estimated covariance matrix and its inverse are of full rank. Typically, three classes of estimators of Ω have been used: thresholding (Bickel and Levina, 2008), shrinkage (Ledoit and Wolf, 2004; Daniels and Kass, 2001) and penalized log likelihood (Tibshirani, 1996). In this paper we consider the latter kind of estimators, the graphical Lasso penalization method (defined in Section 2.1) which adds the penalty $\lambda||\Omega||_1$ with a tuning parameter λ in the maximum likelihood. The penalized maximum likelihood optimization problem is solved using recursive algorithms, for instance we find that three of the most efficient and commonly employed ways to solve it are GLasso by Friedman et al. (2007), Neighborhood selection by Meinshausen and Bühlmann (2006) and Tuning-Insensitive Graph Estimation and Regression by Liu and Wang (2012). The choice of the tuning parameter λ represents the trade-off between close fit to the data and sparsity of Ω, and its selection for estimation of the corresponding CD graph structure is the topic of this paper.

Methods such as Cross Validation (CV), Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) have been widely used to select tuning parameters when p is small. However, they fail once dealing with high-dimensional problems by over-fitting the graph structure of Ω (Liu et al., 2010; Wasserman and Roeder, 2009).

Liu et al. (2010) proposed the selection of λ by controlling the desirable approximated variability in the estimated graphs using a subsampling approach (StARS). This method contrasts with the usual variable selection statistics since it only considers the estimated CD graph structure. Even though the method is promising and gives an alternative to AIC and BIC, it has a major drawback: another tuning parameter is needed in order to set the maximum variability across samples which can be unknown a priori in many applications. Our simulations show that the default values can lead to overestimation of the network size in certain graph topologies. Meinshausen and Bühlman (2010) presented a stability selection approach which controls the graph edges false discovery rate. The authors estimate Ω by an average subsampling graphical Lasso method such that the effect of the choice of λ is very low. However, the trade-off between false positive and true positive edges of the selected network by their subsampling approach is worse than the one given by a network with the same number of edges using all the data due to considering smaller effective sample sizes than the original n for estimation. To the best of our knowledge, there is no other relevant approach in the literature that only employs the graph structure to select the tuning parameter λ in graphical models.

We have applied the following methods for selecting λ popular in statistical literature to estimate CD graph structures in microarray data: AIC, BIC and StARS. However, the graphs we have obtained were rather dense and very difficult to interpret to a biologist, namely to extract groups of genes acting together and possibly interacting. In the biological literature, the most commonly used approaches to construct gene networks are based on clustering. This is informed by the expected presence of distinct strongly interconnected clusters in biological networks (Eisen and Spellman, 1998; Yi et al., 2007). This gave us the motivation to find λ such that the corresponding graph has a clustering structure which can be interpreted by a biologist without restricting it to a block diagonal structure and hence missing potentially important interactions.

Our aim is to select the hyperparameter λ such that (a) it produces reliable estimates of the edges of the graph (b) the corresponding CD graph structure is interpretable in terms of network characteristics and (c) works well for networks that arise in biological systems.

In this paper, we propose several such approaches to selecting λ, in the framework of a general two-step procedure. The main novelty with respect to classical approaches such as AIC or BIC is that we use only the graph structure of the GLasso estimator to tune the regularization parameter λ. The first proposed approach, Path connectivity (PC), uses the average geodesic distance of estimated networks to find the graph that corresponds to the biggest change of the number of connections and is associated with splitting of clusters. The second method, Augmented mean square error (A-MSE), similarly to the StARS approach, controls the variability of the estimated networks in terms of graph dissimilarity coefficients using subsampling. The main difference from StARS is the additional bias term to avoid having a tuning parameter. We consider the bias with respect to an initial estimated graph structure which contains a desirable global network characteristic. For instance, we use the AGNES hierarchical clustering coefficient (Kaufman and Rousseeuw, 2009), which is the third proposed method to choose λ, to select the graph that presents the highest clustering structure. Although clustering methods exist in the literature, the novelty here is that we use them to select the penalty parameter λ in Graphical Lasso estimation.

We compare performance of the proposed approaches as well as of the StARS algorithm and of the standard AIC and BIC on both simulated and real data. The data is a microarray gene expression data set generated by the TCGA Research Network: `http://cancergenome.nih.gov/`. It contains 154 samples for patients with colon tumor and about 18k genes. We are particularly interested in finding significant complex gene interactions reliably and relating the observed associations to pathway databases which describe known biochemistry connections between genes. Simulations and real data analysis are performed using the free statistical software R (R Core Team, 2015).

The rest of the paper is organized as follows. In Section 2 we introduce the tuning parameter selection methodology and in Section 3 we give their main algorithmic and computational information. In Section 4 we compare the performance of the methods using simulated data and then apply them to a gene expression dataset in Section 5.

2 Regularization parameter selection

2.1 Gaussian graphical model

We assume that the data are i.i.d. observations from a Gaussian model: $X_i \sim N_p(0, \Omega^{-1})$, $i = 1, \ldots, n$ independently, assuming, without a loss of generality, that the mean is zero. Conditional dependence is totally characterized by the inverse covariance matrix Ω, also called the precision matrix. Two Gaussian random variables X_i and X_j are said to be conditionally independent given all the remaining variables if the coefficient $\Omega_{i,j}$ is zero. This is often expressed with a graph structure G in which each node represents a random variable and there is an edge connecting two different nodes if the correspondent element in the inverse covariance matrix is non-zero.

The corresponding log likelihood function for Ω is $\ell(\Omega) = \log\det\Omega - \operatorname{tr}(S\Omega)$ where $S = n^{-1}\sum_{i=1}^{n} X_i^2$. If S^{-1} exists ($p < n$ is a necessary condition), the MLE of Ω is given by S^{-1}. However, in a high dimensional framework where the number of variables p is larger than the number of observations n, the matrix S is singular and so cannot be inverted.

We make an additional assumption that the CD (conditional dependence) graph is sparse, and hence that the precision matrix Ω is sparse. Ideally, we would like to use a penalized likelihood estimator, with the penalty proportional to the number of non-zero elements in Ω. However, such optimization problem is non-convex and thus is very computationally intensive. In practice, a likelihood estimator with a convex penalty term proportional to the ℓ_1 norm of Ω, a Graphical Lasso, is commonly used instead:

$$\hat{\Omega}^{\lambda}_{PML} = \arg\max_{\Omega \succ 0} \; [\log\det\Omega - \operatorname{tr}(S\Omega) - \lambda||\Omega||_1], \tag{1}$$

where $||\Omega||_1 = \sum_{i,j=1}^{p} |\Omega_{ij}|$ is the element-wise ℓ_1 norm of the matrix Ω and PML stands for penalized maximum likelihood. For small λ, the corresponding penalized estimator of Ω tends to be dense and in the extreme ($\lambda = 0$) we are back to the initial Maximum Likelihood problem which may not have unique solution when p/n is large (Pourahmadi, 2011). As we increase λ, the matrix becomes more and more sparse until we get a diagonal matrix. Therefore, the choice of λ has a crucial effect on the estimated CD graph structure.

2.2 General two step procedure to select the tuning parameter

The ℓ_1 penalized maximum likelihood estimator defined in (1) requires selection of a regularization parameter λ. If the ℓ_1 penalization genuinely represented our true prior knowledge about Ω then one of the standard methods such as the maximum marginal likelihood or cross validation for the elements of Ω could be used. However, the ℓ_1 penalty here is used due to its computational convenience, replacing the ℓ_0 penalty, so these methods are not appropriate. It is well known for the problem of estimating sparse vectors in high dimensions with the Lasso penalty, that the variable selection part, with an appropriate λ, is consistent, however, the estimation of the non-zero values usually has some bias (Wasserman and Roeder, 2009; Gu et al., 2013). This can be due to the convex relaxation of the desired ℓ_0 penalty to the computationally efficient ℓ_1 penalty. Thus, in this paper we propose to employ methods that use only the variable selection part from the GLasso, $\widehat{G}^{\lambda}$, for tuning the hyperparameter λ.

We propose the following two step procedure for estimating λ:

1. Set $\hat{\Omega}^{\lambda}_{PML}$ as in equation (1) for all $\lambda \in \Lambda$, $\Lambda \subset [0, \lambda_{max}]$, $\lambda_{max} > 0$.
2. Choose $\hat{\lambda} = \arg\min_{\lambda} R(\lambda, \widehat{G}^{\lambda})$

using risk functions R that are based only on CD graphs $\widehat{G}^{\lambda}$. This procedure combines computational efficiency of the Lasso algorithm with the choice of λ that optimizes relevant characteristics of the CD graph such as connectivity, clustering structure, etc.

2.3 Graph notation and distances

Before introducing the risk functions, we give some basic definitions and properties of networks (Costa and Rodrigues, 2007; Estrada, 2011) which will be used to select the regularization parameter.

A graph $G(V, E)$ is a set of nodes V, with connections between them, called edges E. The graph structure is often represented by a $p \times p$ matrix, called adjacency matrix and denoted by A_G. In the estimation of graphical models, the off-diagonal elements of A_G are determined by the precision matrix (0 if $\Omega_{ij} = 0$ and 1 otherwise) and the diagonal elements are set to zero. Note that graphical models are undirected which means that the correspondent A_G is always symmetric.

The distance between a pair of nodes V_i and $V_j \in G(V, E)$ (also known as the geodesic distance) defines the shortest number of edges connecting node V_i to the node V_j, and it is denoted by g_{ij}. If there is no path linking the two nodes, then $g_{ij} = \infty$. The correlation coefficient σ_{ij} between two nodes V_i, $V_j \in G(V, E)$ and the corresponding dissimilarity measure d_{ij} are given by

$$\sigma_{ij} = \eta_{ij}/\sqrt{\kappa_i \kappa_j}, \qquad \text{with} \qquad d_{ij} = 1 - \sigma_{ij}, \quad D = [d_{ij}] \tag{2}$$

where η_{ij} is the number of neighbors shared by the nodes V_i and V_j and κ_i is the degree of the node V_i defined as the number of nodes that are directly connected to V_i.

2.4 Proposed risk functions

We propose several risk functions to select λ that monitor network characteristics of the conditional dependence graphs that can be applicable to genomic data. It has been observed (Yi et al., 2007) that molecules in a cell work together in groups, with some – usually less strong – interaction between the groups. This motivates our choice of risk functions to encourage a clustering structure in the estimated graphs.

2.4.1 Path connectivity risk function

To motivate the first proposed risk function, we observe the following obvious property of the graph $\hat{G}^\lambda$ that corresponds to the penalized estimator $\hat{\Omega}^\lambda$ defined by (1): for small λ, the likelihood term dominates and the estimator $\widehat{G}^\lambda$ is usually a dense graph with $\hat{\Omega}^\lambda$ closely fitting the data, and for large λ, the penalty term dominates and the corresponding estimate is a very sparse graph with $\hat{\Omega}^\lambda$ not fitting the data well. Thus, for growing values of λ, there is a decrease in graph complexity, and the aim of the method we propose here is to capture the value of λ that corresponds to the largest change in the complexity of the graph.

For simplicity, we consider a grid of values of λ, $\Lambda = (\lambda_k)_{k=1}^M$ such that $\lambda_k - \lambda_{k-1} = h$, $k = 2, \dots, M$, and the underlying estimated graphs $\widehat{G}^\lambda$ for all $\lambda \in \Lambda$. We propose Path connectivity (PC) which is a novel approach to find λ that finds the biggest change in graph complexity between the graphs $\hat{G}^\lambda$ corresponding to two consecutive values of $\lambda \in \Lambda$. In this case the measure of graph complexity is calculated by the *geodesic distance mean* statistic

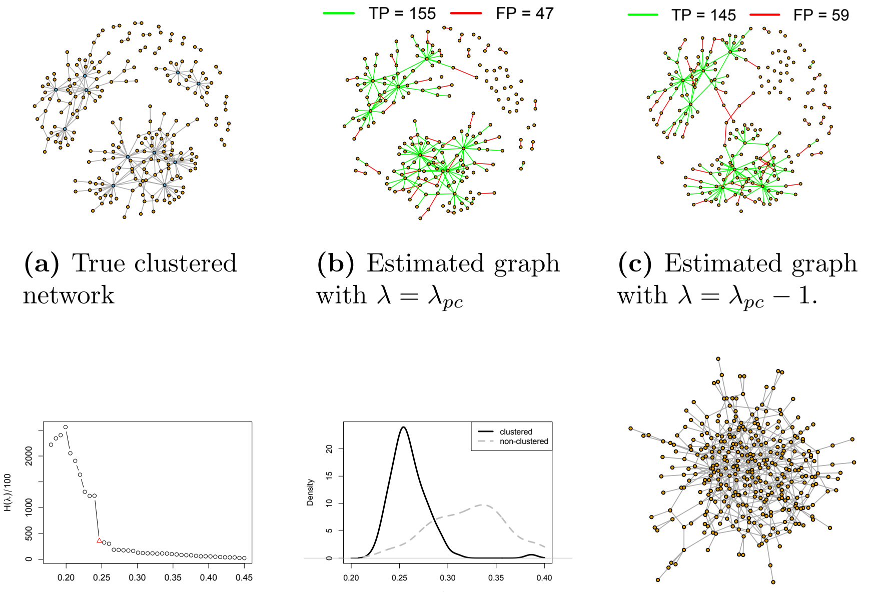
$$H(\lambda) = \frac{2}{p\,(p-1)} \sum_{i<j} \hat{g}_{ij}(\lambda) I(\hat{g}_{ij}(\lambda) < \infty), \tag{3}$$

where $\hat{g}_{ij}(\lambda)$ are the dissimilarity coefficients for the graph $\hat{G}^{\lambda}$. To find the largest change in $H(\lambda)$, we consider the first order differences of $H(\lambda)$ by $D_h(\lambda) = \Delta_h H(\lambda)$, where Δ_h refers to the difference operator with bandwidth h. The regularization parameter selection by PC is given by the λ that produces the most rapid relative descent in the number of graph connections:

$$\lambda_{pc} = \arg\max_{\lambda_k \in \Lambda} R_{PC}(\lambda_k) = \arg\max_{\lambda_k \in \Lambda} \left| D_h(\lambda_k)/\bar{D}_h(\lambda_k) \right|, \tag{4}$$

where λ_k is the k-th ordered element in Λ and $\bar{D}_h(\lambda_k)$ is the running average defined as the average of elements $D_h(\lambda)$ with $\lambda \in \{\lambda_1, \ldots, \lambda_k\}$. The difference of the geodesic distance mean is divided by $\bar{D}_h(\lambda_k)$ in (4) to favor big jumps for larger λ_k (and sparser $\hat{G}^{\lambda}$) in comparison to the jumps for smaller λ_k which correspond to more dense graphs.

In Figure 1 we illustrate the motivation of using the PC selection of λ in simulated data (see Section 4 for details). The true CD graph structure defined by three non-overlapping clusters is plotted in Figure 1(a). We show the geodesic distance mean as function of λ for graph estimations in Figure 1(d). This presents a few big jumps which are related to the separation of clusters. The last one gives the selected graph by PC and is due to the partition of two clusters (see Figure 1(b) for the selected $\lambda_{pc} = \lambda_k$ and Figure 1(c) for the previous graph structure defined by λ_{k-1}). This is a generally observed behaviour in both simulated and real gene expression datasets. In Figure 1(e) we show the density estimates of λ_{pc} using 100 i.i.d. datasets with $n = 200$, $p = 350$ and two theoretical graph structures: hubs-based clustered graph as shown in Figure 1(a) and non-clustered/random graph structure as shown in Figure 1(f). We can see the clear peak around $\lambda = 0.25$ for the clustered data against a flatter empirical distribution for the non-clustered data.



(a) True clustered network

(b) Estimated graph with $\lambda = \lambda_{pc}$

(c) Estimated graph with $\lambda = \lambda_{pc} - 1$.

(d) Geodesic distance mean

(e) Densities of λ_{pc}

(f) True non-clustered network

Figure 1. Path connectivity regularization parameter selection (PC) using the clustered graph structure in (a) to generate the data. Figure (b) shows the selected network by PC and (c) its previous estimated network. In both networks, true positive edges are in green whereas false positives are in red. The graphical structure in (b) differs from the one in (c) since the two clusters in the bottom are no longer connected by a (false positive) edge. Figure (d) shows the geodesic distance mean statistic over several values for λ in which the triangle point is λ_{pc}. Figure (e) illustrates the empirical distribution of λ_{pc} over 100 i.i.d. instances of data with true graph structure in (a), with black solid line, and true graph structure in (f), with grey dashed line. The first concentrates the values to a peak at 0.25 whereas the second is more disperse leading to values of λ_{pc} ranging from 0.27 to 0.35.

2.4.2 A-MSE risk function

The idea explored in this section is to use a risk function based on network characteristics such as dissimilarities of the graph defined by (2). Ideally, we would like to find λ^{oracle} that minimizes

$$R_{MSE}(\lambda) = \mathbf{E}(\sum_{i>j} |d_{ij} - \hat{d}_{ij}(\lambda)|^q), \tag{5}$$

for some $q \geq 1$ where d_{ij} are the dissimilarities of the true graph defined by (2) and $\hat{d}_{ij}(\lambda)$ are the dissimilarities of the CD graph estimated by (1) for a given tuning parameter λ. $R_{MSE}(\lambda)$ depends on the unknown true graph structure of Ω; in practice, an unbiased estimator of $R_{MSE}(\lambda)$ is used, commonly obtained by subsampling (bootstrap, cross validation) by comparing estimated values to observations. However, the problem in this setting is that direct observations of d_{ij} are not available.

To overcome this problem we propose to use an initial graph estimate $\tilde{G}$ and its dissimilarities coefficients $[\tilde{d}_{ij}]$ in place of observed data. Thus, we propose to use the following choice of λ:

$$\lambda_{amse} = \arg\min_{\lambda\in\Lambda} \hat{R}_{AMSE}(\lambda) = \arg\min_{\lambda\in\Lambda} \sum_{i>j} \hat{\mathbf{E}}|\tilde{d}_{ij} - \hat{d}_{ij}(\lambda)|^q, \tag{6}$$

where $\hat{\mathbf{E}}$ indicates the estimation of the expected value using subsampling, and it is obtained as presented in Section 3.2. We find that λ_{amse} can approximate well λ^{oracle} in our simulated data (see Section 5 in Supplementary material).

For $q = 2$, this risk function can be written as a sum of the variance term and the sum of the squared differences between the initial and the current estimator (the "bias" term); see equation (8) in Section 3.2. Note that the first summand in (8), the variance of the estimated distances, gives a stability measure similar to the one proposed in StARS (the latter uses the adjacency matrix instead of the dissimilarities). However, we add a bias term for the distance estimator which allows us to avoid the selection of the power tuning parameter β that controls the desired variability in the StARS approach (Liu et al., 2010).

The proposed $R_{AMSE}(\lambda)$ risk can be applied to other network characteristics. By the definition of graph dissimilarities, $d_{ij} = 1$ if nodes i and j are neither directly nor indirectly (share neighbor) connected. Defining $h_{ij} = 0$ if $\sigma_{ij} = 1 - d_{ij} = 0$ and $h_{ij} = 1$ if $\sigma_{ij} > 0$, for sparse networks, there are many $h_{ij} = 0$ and only few $h_{ij} = 1$. Applying the $R_{AMSE}(\lambda)$ to $[h_{ij}]$ instead of $[d_{ij}]$, we obtain

$$R^h_{AMSE}(\lambda) = \mathbf{E}\sum_{i<j}(h_{ij} - \hat{h}_{ij}(\lambda))^2 = C_h + \mathbf{E}\sum_{(ij)\in\theta(\lambda)}(1 - 2h_{ij}) = C_h + \mathbf{E}[TP(\lambda) - FP(\lambda)]$$

where $\theta(\lambda) = \{(i,j);\, i < j \,\&\, \hat{h}_{ij}(\lambda) = 0\}$, $\mathrm{FP}(\lambda) = \sum_{i<j} I[h_{ij} = 0,\, \hat{h}_{ij}(\lambda) = 1]$, $\mathrm{TP}(\lambda) = \sum_{i<j} I[h_{ij} = 1,\, \hat{h}_{ij}(\lambda) = 1]$ and C_h is independent of λ. Minimizing $R^h_{AMSE}(\lambda)$ is the same as maximizing the TP and FP differences (also known as Youden indices).

In practice, biologists often use clustering algorithms to discover groups of genes. Hence, we propose to use the output of a hierarchical clustering algorithm as an initial estimate of the graph to characterize global structure for the dissimilarities $[d_{ij}]$. We have investigated several clustering algorithms on real and simulated data, and we have not found much difference in the resulting graph estimate. Below we present the algorithm based on AGNES clustering method.

2.4.3 AGNES risk function

Clustering of features using a dissimilarity measure has been intensively studied in the literature. Here we focus on the algorithm AGNES (AGglomerative NESting) which is presented in Kaufman and Rousseeuw (2009, chap. 5) and is implemented in the R package `cluster` (Rousseeuw et al., 2013). AGNES finds clusters iteratively joining groups of nodes with the smallest average dissimilarity coefficient. This average is found by considering the dissimilarity coefficients between all possible pairs of nodes from two different clusters. Moreover, AGNES proposes an agglomerative coefficient (AC) that measures the average distance between a node in the graph and its closest cluster of nodes. We propose to choose λ that maximizes the AC coefficient

$$\lambda_{ac} = \arg\max_{\lambda\in\Lambda} \hat{R}_{AGNES}(\lambda) = \arg\max_{\lambda\in\Lambda} AC(\lambda). \tag{7}$$

The details of the AGNES algorithm and the definition of the coefficient AC can be found in Section 3.3.

The matrix of dissimilarities D obtained by (2) gives a good representation of the complexity of a given graph, so, in addition to being applied as an initial estimate for the A-MSE method described above, AGNES can also be used as a method of choosing λ.

2.5 Comparison of the methods

In Table 1 we give some of the main properties of the 6 risk functions we want to compare which are the three proposed methods, as well as StARS, AIC and BIC. Likelihood-based risk functions to select λ such as AIC and BIC are useful to compromise between goodness of fit to the data and model over-fitting. The additional AIC penalty (given by $p\,(p-1)$) is smaller than BIC (given by $p\,(p-1)\log(n)/2$) even for very small n. Hence, the selection of λ by AIC results in a denser CD graph structure of Ω than by BIC. StARS gives a good alternative to select λ when estimating graph structures. It transforms the selection of λ problem to the choice of the maximum expected variability that we allow in the graph. Even though such a choice is more intuitive than the direct selection of λ, we find it difficult to use without any prior information; our simulations show that using the default value of the tuning parameter results in high number of false positive edges (see Section 4.4).

We provide two computationally fast approaches, AGNES and PC, and the slightly more computationally challenging A-MSE method due to subsampling. The AGNES selection tends to find the most clustered graph possible such that different groups of nodes can be interpreted and analyzed. This is found to be a good choice of λ to recover global graph structure characteristics when the true precision is block diagonal (See Section 4 in the supplementary material). The A-MSE selection uses the AGNES estimator as the initial graph structure with the aim to improve estimations of local network characteristics. The value of λ selected by A-MSE is at least as large as the one given by the initial estimator (AGNES), and it is used to stabilize the trade-off between false positive and true positive edges in the original estimator (AGNES) when n is small (for details see Section 4.4). Moreover, as the sample size increases, the value of λ chosen by the A-MSE method tends to the original estimator of λ (AGNES). We use Path connectivity as the initial good choice of λ to find the most sparse graph that is easy to interpret. Starting from the sparsest graph and proceeding to denser graph structures, the PC method monitors the first big change in connectivity of the estimated networks, which is frequently associated with cluster agglomerations.

Table 1. Main characteristics of six risk functions that can be separated between statistics that use the likelihood expression (BIC, AIC) and statistics that only use the graphical structure of the estimated precision matrices (PC, A-MSE, AGNES, StARS).

method	penalized likelihood	uses network characteristics.	subsampling	fully automatic	fast	very sparse graph estimates
PC		✓		✓	✓	✓
A-MSE		✓	✓	✓		✓
AGNES		✓		✓	✓	
StARS		✓	✓			
BIC	✓			✓	✓	✓
AIC	✓			✓	✓	

3 Algorithms

3.1 Path connectivity regularization parameter selection

The procedure to select λ by Path connectivity is detailed in Algorithm 1. It is generally fast and straightforward, i.e. does not require any additional tuning.

Algorithm 1 Path connectivity algorithm

```
procedure R_{PC}($\lambda$)
Set $\Lambda = (\lambda_k)_{k=1}^M$ with $\lambda_k - \lambda_{k-1} = h$, $k = 2, \dots, M$.
    for k in 1 until M do:
        Estimate the graph $\hat{G}^{\lambda_k}$ using (1) and calculate its geodesic distance matrix
         $[\hat{g}_{ij}]$ as in (2).
        Calculate geodesic distance mean $H(\lambda_k) = m^{-1} \sum_{i<j} \hat{g}_{ij}(\lambda_k) I(\hat{g}_{ij}(\lambda_k) < \infty)$
         with $m = p(p-1)/2$.
Calculate $D_h(\lambda_k) = H(\lambda_k) - H(\lambda_{k-1})$ and the running average $\bar{D}_h(\lambda_k) = 1/(M - k - 1) \sum_{j=k}^M D_h(\lambda_j)$ for $(\lambda_k)_{k=2}^M$.
Return $D_h(\lambda_k)/\bar{D}_h(\lambda_k)$, $k = 2, \dots, M$.
```

3.2 A-MSE regularization parameter selection

For $q = 2$, the risk function $R_{AMSE}(\lambda)$ presented in (6) can be decomposed by the sum of the variance and the squared bias, with the corresponding approximation given by

$$\hat{R}_{AMSE}(\lambda) = \sum_{i>j} [\hat{\mathbf{E}}(\hat{\mathbf{E}}[\hat{d}_{ij}(\lambda)] - \hat{d}_{ij}(\lambda))^2 + (\hat{\mathbf{E}}[\hat{d}_{ij}(\lambda)] - \hat{d}_{ij}(\lambda_{ac}))^2]. \qquad (8)$$

Here $\hat{\mathbf{E}}(\hat{\mathbf{E}}[\hat{d}_{ij}(\lambda)] - \hat{d}_{ij}(\lambda))^2$ and $\hat{\mathbf{E}}[\hat{d}_{ij}(\lambda)] - \hat{d}_{ij}(\lambda_{ac})$ are estimators of the variance of $\hat{d}_{ij}(\lambda)$ and the bias of $\hat{d}_{ij}(\lambda)$ with respect to $\hat{d}_{ij}(\lambda_{ac})$ using subsampling. The subsampling procedure to select λ_{amse} is presented in Algorithm 2. Following Meinshausen and Bühlman (2010) we choose the effective sample size $B = 0.5n$ since the procedure gets the closest to bootstrap. Nevertheless, other effective sizes could be used. For instance, Liu et al. (2010) use $B = 10\sqrt{n}$.

Algorithm 2 Subsampling approach to approximate (8)

```
procedure R$_{AMSE}(\lambda)$
Set $\Lambda = (\lambda_k)_{k=1}^{M}$ and number of subsampling replicates T.
    for t in 1 until T do:
        Subsample $B \subset \{1 : n\}$ and set $X_B = (X_j, j \in B)$.
        Estimate the graphs $\hat{G}^t(\lambda_k)$ for all $\lambda_k \in \Lambda$ using $X_B$.
        Find dissimilarities of $\hat{G}^t(\lambda_k)$ by $\hat{d}^t_{ij}(\lambda_k) = 1 - \eta^t_{ij}(\lambda_k)/\sqrt{\kappa^t_i(\lambda_k)\kappa^t_j(\lambda_k)}$.
Estimate the average $\bar{d}_{ij}(\lambda_k)$ over all $T$ iterations.
Return $T^{-1}\sum_{t=1}^{T}(\bar{d}_{ij}(\lambda_k) - \hat{d}^t_{ij}(\lambda_k))^2$ for all $\lambda_k \in \Lambda$.
```

3.3 AGNES regularization parameter selection

Below is the AGNES iterative clustering algorithm, including the agglomeration coefficient that is used to select λ. The input to the algorithm is a dissimilarity matrix $D = [d_{ij}] = \hat{D}(\lambda)$ based on the graph $\hat{G}^{\lambda}$ corresponding to the estimator $\hat{\Omega}^{\lambda}$ defined by (1). AGNES performs hierarchical clustering by iteratively joining groups of nodes with the smallest average dissimilarity coefficient, starting with individual nodes as single clusters and finishing with a single cluster of all p variables. Let $(C_1^{(t)}, \ldots, C_p^{(t)})$ be a partition of $(1 : p)$ at iteration

t, and let $\delta_{k,\ell}^{(t)}$ denote a dissimilarity between clusters $C_k^{(t)}$ and $C_m^{(t)}$. We also record the dissimilarity for each node when it merges with another cluster or node for the first time, denoting it by $\delta_j^\star$, $j = 1, \ldots, p$, and the distance $\delta_{\max}^\star$ between the two clusters merged at the last step into the single cluster. The procedure is detailed in Algorithm 3.

Algorithm 3 AGNES clustering algorithm

```
1: procedure $\mathrm{R}_{AGNES}(\lambda)$
2: Initialization: take each node as an individual cluster, i.e. set $C_k^{(0)} = \{k\}$, $k = 1, \ldots, p$,
   and $\delta_{k,\ell}^{(0)} = d_{k,\ell}$ - dissimilarity between nodes $k$ and $\ell$.
3: At iteration $t \geq 0$:
4:     Find pair of clusters $(h, k)$ $(h < k)$ with the smallest dissimilarity, i.e.
           $(h, k) = \arg\min_{i<j} \delta_{i,j}^{(t)},$
       merge them, i.e. set $C_k^{(t+1)} = \{C_k^{(t)}, C_h^{(t)}\}$ and remove cluster $h$: $C_h^{(t+1)} = \varnothing$.
       Remaining clusters are unchanged: set $C_j^{(t+1)} = C_j^{(t)}$ for $j \neq k, h$.
5:     The dissimilarities change to
           $\delta_{j,h}^{(t+1)} = \delta_{h,j}^{(t+1)} = \infty, \quad \delta_{k,j}^{(t+1)} = \delta_{j,k}^{(t+1)} = \frac{1}{2}\left[\delta_{k,j}^{(t)} + \delta_{j,h}^{(t)}\right], \quad \forall j \neq k, h.$
       If $|C_k^{(t)}| = 1$, set $\delta_k^\star = \delta_{k,h}^{(t)}$; if $|C_h^{(t)}| = 1$, set $\delta_h^\star = \delta_{k,h}^{(t)}$.
6:     If the number of non-empty sets (clusters) in the newly formed partition $(C_j^{(t+1)})$
       is more than 1, then set $t = t + 1$ and go to step 3; otherwise set $\delta_{\max}^\star = \delta_{k,h}^{(t)}$ .
7: Return
           $AC(\lambda) = \frac{1}{p} \sum_{j=1}^{p} \left(1 - \frac{\delta_j^\star}{\delta_{\max}^\star}\right).$    (9)
```

The coefficient $AC(\lambda)$ measures the average distance between a node in the graph and its closest cluster of nodes. When the dissimilarities within the clusters are small in comparison to the maximum dissimilarity, then $1 - \delta_j^\star / \delta_{\max}^\star$ is large for all j and $AC(\lambda)$ is consequently high.

The time and total memory used in the AGNES algorithm increases exponentially as p grows. In order to make computations feasible in very high dimensions, we use an approximation of the measure by a variable subset selection approach (Kohavi and John, 1997). We consider the average AC coefficient with respect to λ over several sets of variables. We validate the subsets $V \subset \{1:p\}$ of size $|V|$ using the coefficients of variation of the empirical degree distribution (κ) defined by $\mathrm{CV}_V = \mathrm{sd}_V(\kappa)/\mathbf{E}_V(\kappa)$ with $\mathbf{E}_V(\kappa) = 1/|V| \sum_{j \in V} \kappa_j$ and $\mathrm{sd}_V(\kappa) = 1/(|V|-1) \sum_{j \in V}(\kappa_j - \mathbf{E}_V(\kappa))^2$ (see Algorithm 4). We aim to find subset of variables whose number of edges is approximately proportional to those in the original matrix. In Section 4 of the supplementary material we illustrate how the variable subset approach reduces the computational time in high-dimensional simulated datasets.

Algorithm 4 Subset selection for AGNES computations

```
1: procedure S($\lambda$)
2: Input: variables $V_t = \{1:p\}$ and their degrees $\kappa = \{\kappa_1, \ldots, \kappa_p\}$.
3: Compute $\mathrm{CV}_{V_t}$.
4: Select randomly $m < p$ variables from the original data to form set $V_0 \subset V_t$.
5: Add all the nodes $V_1$ in the adjacency matrix $\hat{A}^\lambda$ which have a path to at least one node
   in $V_0$. Use $V_s = \{V_0, V_1\}$.
6: Compute $\mathrm{CV}_{V_s}$. If $|\mathrm{CV}_{V_s}/\mathrm{CV}_{V_t} - 1| > \tau$ go to step 4, otherwise return $V_s$.
```

4 Simulated data analysis

In this section we consider simulated data to test the performance of the regularization parameter selection methods using graph structures similar to what can be expected in biological networks. We analyze both the capacity to obtain the true connections and the accuracy in recovering network characteristics of the true graph.

4.1 Graph topologies in biological data

In real applications, the graph which defines causal connections between variables (e.g. genes, proteins, etc) is unknown but there is typically some knowledge about what kind of network structure can be expected (Newman, 2003). For instance, biological graph structures usually present associations in the shape of clusters, meaning that the nodes form groups that are more similar to the nodes within the group than to the nodes of other groups (Eisen and

Spellman, 1998). In addition, network patterns can be defined by the distribution of the variable p_k, which denotes the fraction of nodes in the network that has degree k. Here we consider two different graph topologies: hubs-based and power-law.

Hubs-based networks are graphs where only few nodes have a much higher degree (or connectivity) than the rest. This is a typical case in biological networks where nodes that behave as hubs may have different biological functions than the other nodes (Lu et al., 2007). Power-law networks assume that the variable p_k follows a power-law distribution

$$p_k = k^{-\alpha}/\varsigma(\alpha),$$

where $k \geq 1$, α is a positive constant and the normalizing function $\varsigma(\alpha)$ is the Riemann zeta function. Following Peng et al. (2009), $\alpha = 2.3$ provides a distribution that is close to what is expected in biological networks.

4.2 Simulated data

We generate data from multivariate normal distributions with zero mean vector and several almost-block diagonal precision matrices, where each block (or cluster) has a hubs-based or power-law underlying graph structure (defined in Section 4.1) and there are some extra random connections between blocks. Let A bet the adjacency matrix with the non-zeros of the partial correlation matrix, then the coefficients of this matrix are simulated by

$$\Omega^{(0)} = [\omega_{ij}^{(0)}], \quad \omega_{ij}^{(0)} = \begin{cases} \text{Unif}(0.5, 0.9) & \text{if } A_{ij} = 1 \text{ with prob}= 0.5 \text{ ;} \\ \text{Unif}(-0.5, -0.9) & \text{if } A_{ij} = 1 \text{ with prob}= 0.5 \text{ ;} \\ 0 & \text{if } A_{ij} = 0. \end{cases} \tag{10}$$

We regularize $\Omega^{(0)}$, which may not be positive definite, by $\Omega^{(1)} = \Omega^{(0)} + \delta I$, with δ such that the condition number of $\Omega^{(1)}$ is less than the number of nodes, so obtaining a positive definite matrix (Cai et al., 2011). Note that such precision matrices are non-singular, sparse and with the non-zero elements bounded away from 0.

We consider precision matrices with p = 50, 170, 290 and 500 and sample sizes n = 50, 100, 200, 500. Different number of hubs, degree of hubs, and sparsity levels are considered in

60 simulated datasets for each combination of p and n. Full specification of simulated data is given in the supplementary material.

We use the R package `huge` (Zhao et al., 2012) to estimate CD graph structures by GLasso and Neighborhood selection (MB). The GLasso gives the estimated partial correlation matrix but MB only provides the estimated adjacency matrix. In order to compare the proposed methods to both AIC and BIC, here we only present the results for the GLasso procedure. Nevertheless, the performance of the methods using MB estimates is shown in the supplementary material. We take a sequence of 70 equidistant points for λ going from 0.20 to 0.66 for small n and a sequence going from 0.03 to 0.40 for large n (the graphs almost have no change for λ's smaller than the lower limit with all nodes connected as well as higher than the upper limit with no edges across nodes). Then we select λ by six different approaches: 1) PC; 2) A-MSE; 3) AGNES; 4) StARS; 5) BIC and 6) AIC. StARS (with $\beta = 0.05$) produces the lowest λ for almost all the simulated datasets followed closely by AIC. The BIC results are strongly dependent on the sample size; the methods selects large tuning parameters for small n and low tuning parameters for large n in comparison to A-MSE. The AGNES selections are always larger than A-MSE but they get close when n increases. The PC λ selections do not vary much for different n and p scenarios and produce similar magnitudes to λ's selected by A-MSE.

We assess the performance of the λ selection approaches for GLasso estimates using two different measures: squared errors in both the partial correlation matrix and the dissimilarity matrix defined in (2) and graph recovery with a false positive and true positive analysis. The simulated data analysis is completed in the supplementary material where we compare for both GLasso and MB the selected graph structures and the true networks given global network characteristics as clustering, connectivity and graph topology.

4.3 Mean square errors

To measure performance of the methods we use the ranks of the average mean square errors (MSE) of the partial correlation matrix Ω (Table 2) as well as of the dissimilarity matrix D (Table 3). This second rate gives a good reference to determine if the estimated graph captures the true local structure. The lowest rank (rank = 1) is assigned to the lowest MSE and the largest rank (rank = 6) is for the largest MSE out of the six approaches. In the tables, we show the errors for the GLasso method.

Even though StARS and AIC estimate Ω well, they produce larger errors than AGNES, A-MSE, PC and BIC when minimizing the MSE of the dissimilarity matrix. Particularly, A-MSE tends to be the best selection for this loss function for large n. We find that BIC does well for small n, contrarily of what is obtained in Liu et al. (2010), but tends to be unreliable for larger sample sizes. AGNES gives fairly good ranks when n is large, and PC is almost always among the three best methods.

Table 2. Average ranks for the mean square error of the precision matrix using several sample sizes, dimension and network topologies (hubs-based and power law). The methods StARS and AIC find the best rates (lowest ranks) whereas PC and A-MSE tend to obtain the worst rates (highest ranks).

	Hubs-based				Power law			
n	50	100	200	500	50	100	200	500
				dimension p=50				
AGNES	3.66	4.00	4.00	2.38	3.71	4.17	4.56	4.73
A-MSE	5.94	5.65	5.72	4.75	5.76	5.72	5.84	5.76
PC	4.97	5.35	5.28	3.45	5.07	5.07	4.60	4.51
StARS	**1.04**	**1.70**	**1.50**	**3.42**	**1.28**	**1.67**	**1.79**	**2.00**
BIC	3.42	2.60	3.00	3.55	3.47	2.71	2.42	**2.00**
AIC	1.96	**1.70**	**1.50**	3.45	1.72	**1.67**	**1.79**	**2.00**
				dimension p=170				
AGNES	2.96	3.79	4.00	4.00	2.82	3.88	4.08	4.35
A-MSE	6.00	5.86	5.75	5.82	5.98	5.91	5.67	5.68
PC	5.00	5.14	5.25	5.18	4.89	5.09	5.25	4.97
StARS	**1.00**	**1.00**	**1.00**	**1.43**	**1.00**	**1.00**	**1.00**	**1.58**
BIC	3.92	3.21	3.00	2.83	3.98	3.12	3.00	2.83
AIC	2.12	2.00	2.00	1.74	2.33	2.00	2.00	**1.58**
				dimension p=290				
AGNES	2.67	3.62	4.00	4.00	2.33	3.79	4.00	4.12
A-MSE	5.83	5.98	5.60	5.75	6.00	5.74	5.84	5.85
PC	5.17	5.02	5.40	5.25	4.92	5.26	5.16	5.03
StARS	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
BIC	3.92	3.38	3.00	3.00	4.08	3.21	3.00	3.00
AIC	2.42	2.00	2.00	2.00	2.67	2.00	2.00	2.00
				dimension p=500				
AGNES	2.25	3.30	4.00	4.00	2.33	3.79	4.00	4.12
A-MSE	6.00	6.00	5.93	5.87	6.00	5.74	5.84	5.85
PC	4.96	5.00	5.07	5.13	4.92	5.26	5.16	5.03
StARS	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**	**1.00**
BIC	4.04	3.70	3.00	3.00	4.08	3.21	3.00	3.00
AIC	2.75	2.00	2.00	2.00	2.67	2.00	2.00	2.00

Table 3. Average ranks for the mean square error of the dissimilarity matrix using several sample sizes, dimension and network topologies (hubs-based and power law). A-MSE tends to be the method with the best rates (lowest ranks). BIC does well for small sample sizes but fails when the sample size increases.

	Hubs-based				Power law			
n	50	100	200	500	50	100	200	500
dimension p=50								
AGNES	3.31	3.00	3.00	2.88	2.89	2.65	2.36	2.33
A-MSE	**1.24**	**1.35**	**1.32**	**1.23**	**1.64**	**1.29**	**1.18**	**1.32**
PC	1.91	1.65	1.68	1.89	2.33	2.35	2.47	2.34
StARS	5.96	5.30	5.50	5.12	5.72	5.33	5.29	5.00
BIC	3.54	4.40	4.00	4.77	3.17	4.04	4.42	5.00
AIC	5.04	5.30	5.50	5.12	5.25	5.33	5.29	5.00
dimension p=170								
AGNES	4.11	3.21	3.00	3.00	4.10	3.03	2.75	2.48
A-MSE	**1.42**	**1.09**	**1.17**	**1.18**	**1.65**	**1.16**	**1.43**	**1.35**
PC	1.72	1.91	1.83	1.82	2.02	1.92	1.82	2.17
StARS	6.00	6.00	6.00	5.74	6.00	6.00	6.00	5.50
BIC	2.97	3.79	4.00	4.00	2.64	3.88	4.00	4.00
AIC	4.79	5.00	5.00	5.26	4.58	5.00	5.00	5.50
dimension p=290								
AGNES	4.32	3.36	3.00	3.00	4.73	3.21	3.00	2.62
A-MSE	**1.35**	**1.17**	**1.40**	**1.25**	**1.28**	**1.39**	**1.38**	**1.57**
PC	1.80	1.85	1.60	1.75	1.80	1.61	1.62	1.81
StARS	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00
BIC	2.97	3.62	4.00	4.00	2.92	3.79	4.00	4.00
AIC	4.57	5.00	5.00	5.00	4.27	5.00	5.00	5.00
dimension p=500								
AGNES	4.73	3.70	3.00	3.00	4.96	3.41	3.00	2.65
A-MSE	2.03	**1.32**	**1.10**	**1.13**	1.71	**1.50**	**1.2**	**1.36**
PC	**1.29**	1.68	1.90	1.87	**1.54**	**1.50**	1.80	1.99
StARS	6.00	6.00	6.00	6.00	6.00	6.00	6.00	6.00
BIC	2.71	3.30	4.00	4.00	2.80	3.59	4.00	4.00
AIC	4.23	5.00	5.00	5.00	3.99	5.00	5.00	5.00

4.4 Graph recovery

In order to quantify how well the algorithms recover the non-zero elements in Ω we compare the true discovery rate (TDR), which can be defined by $TDR = TP/(TP + FP)$ with

$$TP = \sum_{i<j} I(\hat{\Omega}_{ij} \neq 0 \text{ and } \Omega_{ij} \neq 0), \quad FP = \sum_{i<j} I(\hat{\Omega}_{ij} \neq 0 \text{ and } \Omega_{ij} = 0),$$

for each of the estimated networks. In Figure 2, we show the average TDR in the 60 simulation data instances for all considered combinations of n and p. The TDR increases with n for AGNES, A-MSE and PC whereas for AIC and BIC it goes down. In this analysis we can see the limitations of the BIC method whose main goal is not the graph recovery of

Ω. BIC passes from selecting very sparse graphs with more TP than FP when n is small to selecting much denser graphs with many more FP than TP when n is large.

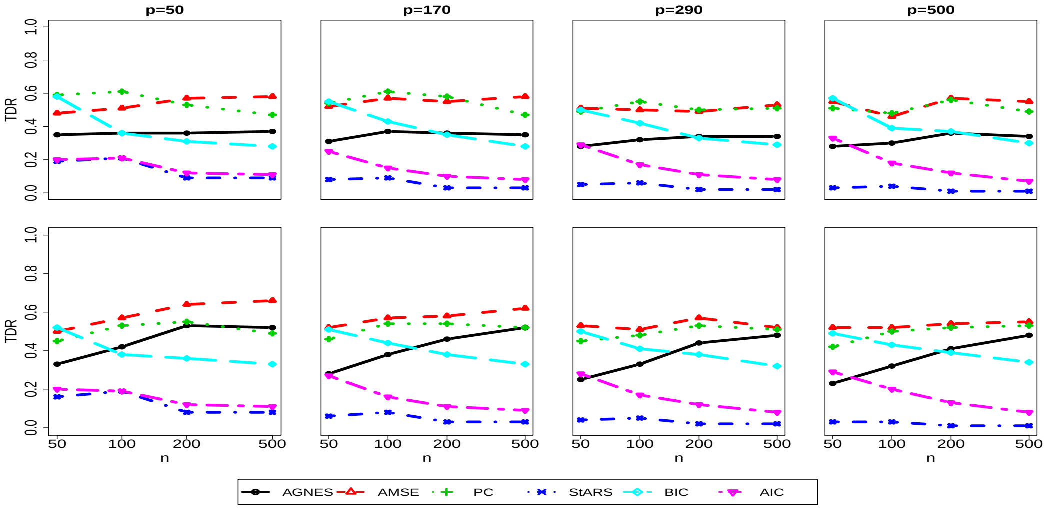


Figure 2. True discovery rate for all λ selection approaches (AGNES -black-, A-MSE -red-, PC -green-, StARS -dark blue-, BIC -cyan- and AIC -purple-) and all combinations of p and n. The top figures correspond to hub-based networks and the bottom figures are the power-law networks. The x-axis scale is $n : \log(n)$. BIC rates decrease with the sample size whereas AGNES, A-MSE and PC rates slightly increase with the sample size.

4.5 Summary

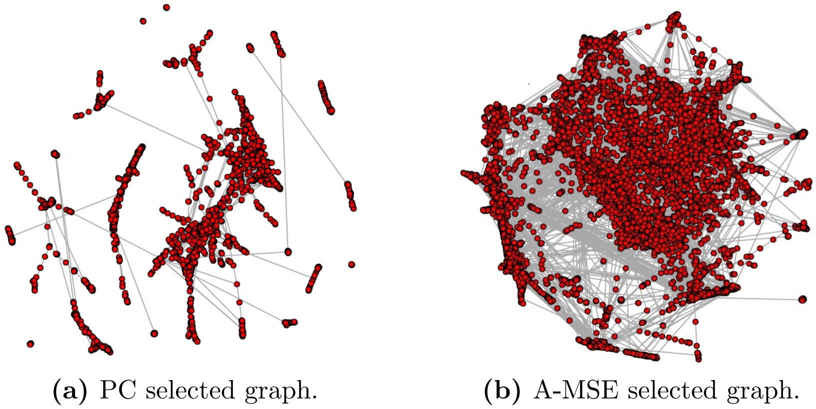
In our simulations A-MSE turned out to be the best approach to recover the CD graph structure as can be seen in Table 3. BIC is also competitive when n is small, but it is not reliable when analyzing larger sample sizes. PC is computationally the fastest method and only does slightly worse than A-MSE in Table 3. Moreover, it generally obtains simple graph structures which result in comprehensible connectivity interpretations. The AGNES procedure is usually over-performed by the augmented version A-MSE for small n. For large n, AGNES and A-MSE have similar λ selections with AGNES being significantly faster than A-MSE. AIC and StARS (using its default values) produce dense graph estimations and achieve the best results when minimizing the mean square error of Ω. Nevertheless, they fail to obtain interpretable network structures due to poor graph recovery.

5 Application to colon cancer gene expression data

We apply the methods to a case study of genomic data which contain the gene expression profile of 154 colorectal tumor samples and 17,617 genes. The data are generated by the TCGA Research Network: `http://cancergenome.nih.gov/`, and are currently available at the portal `https://gdc-portal.nci.nih.gov/`, under the TCGA cancer program and the Colon Adenocarcinoma disease type.

A reduction on the variable space is applied so that we only keep the most highly correlated genes. We use a filter for the gene's average square correlation with threshold equal to 0.04. Moreover, we add the non-filtered genes which have at least one correlation coefficient with the filtered genes larger than 0.5. This means a reduction to the 55% of the genes with a total of 9,723 genes left to analyze. We estimate CD graphs via the Neighborhood selection algorithm of Meinshausen and Bühlmann (2006). We compute 90 different graphs given an equidistant sequence of λ's between 0.35 and 0.80. Values of λ lower than 0.35 produce almost-fully connected graphs and values above 0.80 produce zero edges in the graph. We use the PC and A-MSE approaches to select one particular graph with $\lambda_{pc} = 0.69$ and $\lambda_{amse} = 0.55$. The graphical representation of the two underlying networks is presented in Figure 3. The graph by PC, with $4,819$ edges, shows a simpler structure compared to A-MSE, with $19,986$ edges.

We separate the graphs in different clusters by applying a Partitioning Around Medoids (Reynolds et al., 2006) on the shortest distance matrix. We choose the number of clusters manually by considering the largest rate of change in the within-subject and between-subject variation such that the PC graph structure contains 15 clusters and the A-MSE contains 18 clusters. To assess which biological processes may be linked to the clusters, we download 1,320 gene sets from the MSig database (Subramanian et al., 2005), which represent canonical pathways compiled from two sources: KeGG (Kanehisa et al., 2016) and Reactome (Milacic et al., 2012). For each pathway we test for a significant over-representation in a cluster by using Fisher's exact test applied to the 2×2-table defined by pathway and cluster membership with a Bonferroni correction for multiple testing. Note that we use the reduced selection of 9,723 genes here as "background", i.e. the analysis corrects for any over-representation of a pathway in that selection.



(a) PC selected graph.

(b) A-MSE selected graph.

Figure 3. Selected graphs by PC and A-MSE to describe conditional gene associations on colon cancer gene expression data. The A-MSE graph is denser than the PC graph but in both cases several clusters of genes are visible.

For the PC and A-MSE selected graphs, respectively, 6 out of 15 clusters of genes, and 7 out of 18 clusters of genes, overlap significantly with at least one pathway gene set (at 0.01 significant level). Besides, a total of 160 and 122 pathway sets (out of 1.320) present significant overlap with clusters of genes defined in the PC and A-MSE graphs. Among the significant lists, PLK1, NFAT, DNA replication or adaptive immune system are pathways associated with tumor cells.

6 Discussion

In this paper we study the problem of choosing the regularization parameter λ for Gaussian graphical models in high dimensional data assuming we have high level knowledge about the nature of the graph structures, namely strong clustering in the case of gene expression data (e.g. Eisen and Spellman, 1998). The methods we introduce here take this assumption into account by selecting λ so that risk functions measuring the degree of clustering (AGNES, A-MSE) or connectivity (PC) are optimized. We aim to select the sparsest graph such that the real cluster structure is maintained and at the same time it contains a good tradeoff between true and false positive edges. The proposed approaches to select the regularization parameter provide competitive results at a relatively high computational speed. They present more reliable results than the StARS approach which tends to overestimate the network size. The StARS method accounts for the stability of the estimated graphs and has been proven to work well in Liu et al. (2010). It depends, however, on another parameter which controls the maximum amount of variability in the graph. There is no straightforward choice for this parameter and our simulation study shows that using the default value of 0.05 StARS yields uninformative networks with a majority of edges being false positives.

The Path connectivity approach introduced here provides a good compromise between estimating the structure well and the number false positive edges. The main characteristic of this approach is that it relies on the shortest distance between all pairs of nodes. Interestingly, this quantity tends to show a clear changepoint when studied as a function of λ, at which the structure of the graph changes radically. It typically produces very informative graphs in all the tested simulated datasets and gives competitive results for the mean square error between dissimilarity matrices as discussed in Section 4.3. In the gene expression data set it also provides us with a clearly structured informative graph. PC gives an excellent first choice of λ if we want to find an easily interpretable graph.

The A-MSE, with initial graph structure given by the AGNES selected graph, is the best of all the approaches in terms of minimizing the MSE between the true distances and the estimated ones in the simulated data. Also, λ_{amse} is always smaller than λ_{ac} leading to less complex graphs than the ones estimated by AGNES. This is a desirable property as we assume only a small proportion of non-zero elements in Ω and thus with increasing graph density the number of false positive edges grows much faster than the number of true positives. However, if the aim is to have fewer false negatives, that is, that as many as possible true edges are included at the expense of a higher number of false positives, then algorithms like AGNES and StARS are more appropriate.

The analysis of the gene expression data underlines some interesting results. The obtained graphs present a cluster-based structure as we can see in Figure 3. Our new approach of choosing a regularization parameter, PC, leads to a sparse and clustered network that is easy to interpret. Closer investigation of the results shows that the clusters overlap significantly with a number of pre-defined gene sets and regulatory pathways which indicates that our assumption of a sparse clustered structure leads us to biologically meaningful results.

In conclusion, we find that approaches such as PC, A-MSE and AGNES, which use network characteristics for parameter selection, can be beneficial in estimating CD graph structures (sparse partial correlation matrices) for high-dimensional biological data. While maintaining good statistical properties in terms of false discovery rates and mean square error, the resulting graphs tend to be easier to interpret from a biological perspective and thus are more useful in applications compared to parameter selection methods based on penalized log likelihood such as AIC or BIC.

Acknowledgements

We acknowledge support from the Rural and Environment Science and Analytical Services Division (RESAS) of the Scottish Government. Natalia Bochkina is grateful to the Alan Turing Institute for the financial support under the EPSRC grant EP/N510129/1.

7 Supplementary Materials

Supplementary material: extension of some of the simulated data analysis (pdf file).

R-package for selection of tuning parameter in graphical models: R-package "GM-RPS" contains the functions to select the regularization parameter in graphical models as well as the functions to generate simulated data. In file "codeSimulatedDataAnalyisMainPaper.r" the main simulated data analysis can be reproducible. Other code available include "pcMotivatingExampleMainPaper.r" (for Figure 1), "AGNEStimeSuppMatPaper.r" and "lambdaOracleSuppMatPaper.r" (for supplementary material).

References

Bickel, P. and E. Levina (2008). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Cai, T., W. Liu, and X. Luo (2011). A constrained L1 minimization approach to sparse precision matrix estimation. *Journal of the American Statistical Association 106*, 594–607.

Costa, L. and F. Rodrigues (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics 56*(1), 167–242.

Daniels, M. J. and R. E. Kass (2001). Shrinkage estimators for covariance matrices. *Biometrics 57*(4), 1173–84.

Dobra, A., C. Hans, B. Jones, J. R. Nevins, G. Yao, and M. West (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis 90*(1), 196–212.

Eisen, M. and P. Spellman (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America 95*, 14863–14868.

Estrada, E. (2011). *The Structure of Complex Networks.* New York: OXFORD University press.

Friedman, J., T. Hastie, and R. Tibshirani (2007). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics 9*(3), 432–441.

Goldenberg, A. (2007). *Scalable Graphical Models for Social Networks.* Doctoral dissertation.

Gu, X., G. Yin, and J. Lee (2013). Bayesian two-step Lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemp Clin Trials 36*(2), 642–650.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research 44*(D1), D457–D462.

Kaufman, L. and P. Rousseeuw (2009). *Finding Groups in Data: An Introduction to Cluster Analysis.* New Jersey: John Wiley & sons.

Kohavi, R. and G. John (1997). Wrappers for feature subset selection. *Artificial intelligence 97*, 273–324.

Lauritzen, S. (1996). *Graphical Models* (Oxford uni ed.). New York.

Ledoit, O. and M. Wolf (2004). A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis 88*(2), 365–411.

Liu, H., K. Roeder, and L. Wasserman (2010). Stability approach to regularization selection (stars) for high dimensional graphical models. *Adv. Neural Inf. Process. Syst. 23*, 1432–1440.

Liu, H. and L. Wang (2012). TIGER : A tuning-insensitive approach for optimally estimating Gaussian graphical models. *arXiv preprint 1209.2437*.

Lu, X., V. V. Jain, P. W. Finn, and D. L. Perkins (2007). Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Molecular systems biology 3*, 98.

Meinshausen, N. and P. Bühlman (2010). Stability Selection. *Journal of the Royal Statistical Society: Series B 72*, 417–473.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the Lasso. *The Annals of Statistics 34*, 1436–1462.

Milacic, M., R. Haw, K. Rothfels, G. Wu, D. Croft, H. Hermjakob, P. D'Eustachio, and L. Stein (2012). Annotating cancer variants and anti-cancer therapeutics in Reactome. *Cancers 4*(4), 1180–1211.

Newman, M. (2003). The structure and function of complex networks. *SIAM REVIEW 45*(2), 167—-256.

Peng, J., P. Wang, N. Zhou, and J. Zhu (2009). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association 104*, 735–746.

Pourahmadi, M. (2011). Covariance estimation: the GLM and regularization perspectives. *Statistical Science 26*(3), 369–387.

R Core Team (2015). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Reynolds, A. P., G. Richards, B. De La Iglesia, and V. J. Rayward-Smith (2006). Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. *Journal of Mathematical Modelling and Algorithms 5*(4), 475–504.

Rousseeuw, P., A. Struyf, and M. Hubert (2013). cluster: Cluster Analysis Basics and Extensions.

Schäfer, J. and K. Strimmer (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics 21*(6), 754–64.

Subramanian, A., P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. a. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America 102*(43), 15545–50.

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

Wasserman, L. and K. Roeder (2009). High-dimensional variable selection. *The Annals of Statistics 37*(5A), 2178–2201.

Yi, G., S. H. Sze, and M. R. Thon (2007). Identifying clusters of functionally related genes in genomes. *Bioinformatics 23*(9), 1053–1060.

Zhao, T., H. Liu, and K. Roeder (2012). The huge package for high-dimensional undirected graph estimation in r. *The Journal of Machine Learning Research 13*, 1059–1062.