Inferring networks from high-dimensional data with mixed variables

Antonino Abbruzzo, Angelo M. Mineo

Abstract We present two methodologies to deal with high-dimensional data with mixed variables, the strongly decomposable graphical model and the regression-type graphical model. The first model is used to infer conditional independence graphs. The latter model is applied to compute the relative importance or contribution of each predictor to the response variables. Recently, penalized likelihood approaches have also been proposed to estimate graph structures. In a simulation study, we compare the performance of the strongly decomposable graphical model and the graphical lasso in terms of graph recovering. Five different graph structures are used to simulate the data: the banded graph, the cluster graph, the random graph, the hub graph and the scale-free graph. We assume the graphs are sparse. Our finding, in the simulation study, is that the strongly decomposable graphical model shows, generally, comparable or better performance both in low and high-dimensional case. Finally, we show an application on mixed data.

1 Introduction

Graphical models are useful to infer conditional independence relationships between random variables. The conditional independence relationships can be visualized as a network with a graph. Graphs are objects with two components: nodes and links. Nodes are in one-to-one correspondence with random variables and links represent relations between genes. If a link between two genes is absent this means that these two genes are conditional independent given the rest. Pairwise, local and

Antonino Abbruzzo
Dipartimento Scienze Economiche, Aziendali e Statistiche, University of Palermo, Viale delle Scienze Ed. 13, 90128 Palermo, Italy e-mail: antonino.abbruzzo@unipa.it

Angelo M. Mineo
Dipartimento Scienze Economiche, Aziendali e Statistiche, University of Palermo, Viale delle Scienze Ed. 13, 90128 Palermo, Italy e-mail: angelo.mineo@unipa.it

global Markovian properties are the connections between graph theory and statistical modeling [1, 2, 3].

Applications of graphical models include among others the study of gene regulatory networks where expression levels of large number of genes are collected, simultaneously [4]. A microarray is a collection of microscopic DNA spots attached to a solid surface. Understanding how genes work together as a network could i) hold the potential for new treatments and preventive measures in disease, ii) add a new level of complexity to scientists' knowledge of how DNA works to integrate and regulate cell functionality. Many of the works on trying of inferring gene regulatory networks have focus on penalized Gaussian graphical models. The idea is to penalize the maximum likelihood function, for example with the ℓ_1-norm, to produce sparse solutions. The main assumption of these models is that the networks are sparse, which means many of the variables are conditionally independent from the others. In this setting, Meinshausen and Bühlmann [5] proposed to select edges for each node in the graph by regressing the variable on all the other variables using ℓ_1 penalized regression. Penalized maximum likelihood approaches using the ℓ_1 penalty have been considered in [6, 7] where different algorithms for estimating sparse networks have been proposed. The most known algorithm to estimate sparse graphs is probably the graphical lasso (glasso) proposed by Friedman *et al.* [8]. These models cannot deal with high-dimensional data with mixed variables. However, the need of statistical tools to analyze and extract information from such data has become crucial. For example, the most recent task in DREAM8 challenge [9] is related to predict the response of Rheumatoid Arthritis patients to anti-TNF therapy based on genetics and clinical data.

In their seminal paper Lauritzen and Wermuth [10] introduced the problem of dealing with mixed variables. Recently, Hoff [11] proposed a semiparametric Bayesian copula graphical model to deal with mixed data (binary, ordinal and continuous). The semiparametric Bayesian copula graphical model uses the assumption of Gaussianity on the multivariate latent variables which are in one-to-one correspondance with the observed variables. Conditional dependence, regression coefficients and credible intervals can be obtained from the analysis. Moreover, copula Gaussian graphical models allow to impute missing data. However, the Bayesian copula approach is infeasible for higher-dimensional problems due to its computational complexity and problem of convergence to the proposal distribution.

In this paper, we present two classes of graphical models, namely strongly decomposable graphical models [12] and regression-type graphical models [13], which are classes of models that can be used for analyzing high-dimensional data with mixed variables. Assuming that the conditional distribution of a variable A given the rest depends on any realization of the remaining variables only through the conditional mean function, the regression models are useful to find the matrix weights which can be further employed to recover the network. The aim here are i) to give some insight on the use of decomposable models for recovering graph structure; ii) to connect this model with the use of regression-type graphical lasso; iii) to provide a simulation study to compare graphical lasso, which is a penalized approach, to strongly decomposable graphical models.

The rest of this paper is organized as follows. In Section 2, we briefly recall the methodologies used to infer decomposable graphical models and regression-type graphs for mixed data. In Section 3 we show a simulation study in which we compare several type of graphs. In Section 4, we show an application of the methodology to a real dataset which contains mixed variables that are the expression level of genes collected in a microarray experiment and some clinical information of the patients.

2 Methodology

A graph is a couple $G = (V,E)$ where V is a finite set of nodes and $E \subset V \times V$ is a subset of ordered couples of V. Nodes are in one-to-one correspondence with random variables. Links represent interactions between the nodes. In this paper, we are interested in links which represent conditional independence between two random variables given the rest. Suppose we have d discrete and q continuous nodes and write the sets of nodes as Δ and Γ, where $V = \{\Delta \cup \Gamma\}$. Let the corresponding random variables be $(\mathbf{X},\mathbf{Y})$, where $\mathbf{X} = (X_1,\ldots,X_d)$ and $\mathbf{Y} = (Y_1,\ldots,Y_q)$, and a typical observation be $(\mathbf{x},\mathbf{y})$. Here, $\mathbf{x}$ is a d-tuple containing the values of the discrete variables, and $\mathbf{y}$ is a real vector of length q. We will denote with $P(\mathbf{z})$ a joint probability distribution for the random variables $(\mathbf{X},\mathbf{Y})$.

2.1 Decomposable graphical models for high-dimensional data

Finding a conditional independence graph from data is a task that requires the approximation of the joint probability distribution $P(\mathbf{z})$. A product approximation of $P(\mathbf{z})$ is defined to be a product of several of its component distributions of lower order. We consider the class of second-order distribution approximation, i.e.:

$$P_a(\mathbf{z}) = \prod_{i=1}^{p} P(z_i, z_{j(i)}), \qquad 0 \le j(i) \le p$$

where $(j_1,\ldots,j_p)$ is an unknown permutation of integers $(1,2,\ldots,p)$, where $p = d+q$.

For discrete random variables, Chow and Liu [14] proved that the problem of finding the goodness of approximation between $P(\mathbf{x})$ and $P_a(\mathbf{x})$ considering the minimization of the closeness measure

$$I(P,P_a) = \sum_{\mathbf{x}} P(\mathbf{x}) \log \frac{P(\mathbf{x})}{P_a(\mathbf{x})},$$

where $\sum_{\mathbf{x}} P(\mathbf{x})$ is the sum over all levels of the discrete variables, is equivalent to maximizing the total branch weight $\sum_{i=1}^{p} I(x_i, x_{j(i)})$, where

$$I(x_i, x_{j(i)}) = \sum_{x_i, x_{j(i)}} P(x_i, x_{j(i)}) \log \left(\frac{P(x_i, x_{j(i)})}{P(x_i) P(x_{j(i)})} \right). \quad (1)$$

Calculating the total branch weight for each of the p^{p-2} trees would be computationally too expensive even for moderate p. Fortunately, several algorithms can be used to solve the problem of finding dependence tree of maximum weight, such as Kruskal's algorithm, Dijkstra's algorithm, Prim's algorithm. These algorithms start from a square weighted matrix p by p, where a weight for a couple of variables (X_i, X_j) is given by the mutual information $I(x_i, x_j)$. So, the problem is reduced to calculating $p(p-1)/2$ weights. Consider, now a real application where probability distributions are not given explicitly. Let $\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N$ be N independent samples of a finite discrete variable $\mathbf{x}$. Then, the mutual information can be estimated as follows:

$$\hat{I}(x_i, x_j) = \sum_{u,v} f_{uv}(i,j) \log \frac{f_{uv}(i,j)}{f_{u(i)} f_{v(j)}},$$

where $f_{uv(i,j)} = \frac{n_{uv}(i,j)}{\sum_{uv} n_{uv}(i,j)}$, and $n_{uv}(i,j)$ is the number of samples such that their ith and jth components assume the values of u and v, respectively. It can be shown that with this estimator we also maximize the likelihood for a dependence tree.

This procedure can be extended to data with both discrete and continuos random variables [12]. The distributional assumption is that random variables $\mathbf{Z}$ are conditionally Gaussian distributed, i.e. the distribution of $\mathbf{Y}$ given $\mathbf{X} = \mathbf{x}$ is multivariate normal $N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ so that both the conditional mean and covariance may depend on i. We refer to the homogenous or heterogenous case if $\boldsymbol{\Sigma}$ does or does not depend on i, respectively. More details on this conditional Gaussian distribution can be found in [10]. To apply the Kruskal's algorithm, in the mixed case, we need to find an estimator of the mutual information $I(x_u, y_v)$ between each couple of variables. For a couple of variables (X_u, Y_v) we can write the sample cell counts, means, and variances as $\{n_i, \bar{y}_v, s_i^{(v)}\}_{i=1,\ldots|X_u|}$. An estimator of the mutual information, in the homogenous case, is

$$\hat{I}(x_u, y_v) = \frac{N}{2} \log \left(\frac{s_0}{s} \right),$$

where $s_0 = \sum_{k=1}^{N} (y_v^{(k)} - \bar{y}_v)^2 / N$ and $s = \sum_{i=1}^{|X_u|} n_i s_i / N$. There are $k_{x_u, y_v} = |X_u| - 1$ degree of freedom. In the heterogeneous case, an estimator of the mutual information is

$$\hat{I}(x_u, y_v) = \frac{N}{2} \log(s_0) - \frac{1}{2} \sum_{i=1,\ldots,|X_s|} n_i \log(s_i)$$

with $k_{x_u, y_v} = 2(|X_u| - 1)$ degrees of freedom.

Note that the algorithm will always stop when it has added the maximum number of edges, i.e. $p-1$ for an undirected tree. Edwards et al. [12] suggested to use either $\hat{I}^{AIC} = \hat{I}(x_i, x_j) - 2k_{x_i, x_j}$ or $\hat{I}^{BIC} = \hat{I}(x_i, x_j) - \log(n) k_{x_i, x_j}$, where k_{x_i, x_j} are the degrees of freedom, to avoid inclusion of links not supported by the data.

The class of tree graphical models can be too restrictive for real data problem. However, we can start from the best spanning tree and determine the best strongly decomposable graphical model. A strongly decomposable graphical model is a graphical model whose graph neither contains cycles of length more than three nor forbidden path. A path exists between nodes A and B if one can reach A from B in a finite number of steps. A forbidden path is a path between two not adjacent discrete nodes which passes through continuous nodes. The distributional assumption is that the random variables are conditional Gaussian distributed. This procedure would be NP-hard without the following result.

If $M_0 \subset M_1$ are decomposable models differing by one edge $e = (v_i, v_j)$ only, then e is contained in one clique C of M_1 only, and the likelihood ratio test for M_0 versus M_1 can be performed as a test of $v_i \perp v_j | C_{\setminus\{v_i,v_j\}}$. These computations only involve the variables in C. It follows that for likelihood-based scores such as AIC or BIC, score differences can be calculated locally which is far more efficient then fitting both M_0 and M_1. This leads to considerable efficiency gains.

To summarize, strongly decomposable model is an important class of model that can be used to analyze mixed data. This class restricts the class of possible interaction models which would be to huge to be explored. Moreover, we have the important results that for strongly decomposable graphical models closed-form estimator exists.

2.2 Regression-type graphical models

Recently, Edwards et al. [13] proposed to estimate stable graphical models with random forest in combination with stability selection using regression models. Their main idea is motivated by the following theorem.

Assume that, for all $j = 1, \ldots, p$ the conditional distribution of Z_j given $\{Z_h; h \neq j\}$ is depending on any realization $\{z_h; h \neq j\}$ only through the conditional mean function:

$$\mu_j(\{z_h; h \neq j\}) = E[Z_j | z_h; h \neq j].$$

Assume the conditional mean exists, then

$$Z_j \perp Z_i | \{Z_h; h \neq j, i\}$$

if and only if

$$\mu_j(\{z_h; h \neq j\}) = \mu_j(\{z_h; h \neq j, i\})$$

does not depend on z_i for all $\{z_h; h \neq j\}$. Suppose the network is composed by variables some of which are predictors and some of which are response variables. We use this theorem to determine the weight importance of each predictor on the response variable. To establish the importance of each predictor regression coefficients need to be comparable, i.e. standardized regression coefficients need to be

used. These coefficients can also be interpreted as elasticity. i.e. how much we can change the regressor, by attempting to exogenously change one of the predictor.

2.3 Simple Example

In this section, we show a simple example on simulated mixed data. The aim is to recover the graph in Figure 1 with a decomposable graphical model and to evaluate the relative importance of each predictor to the regressors with regression-type graphical models. In particular, in Figure 1 we represents five variables with some of them that are regressor variables. These variables are those one having at least an incoming link. Table 1 shows distributions, models and conditional means of each variable. Regression coefficients are given in Table 2.

Distribution	Model	Conditional mean
Gaussian	$Y_1 \sim N(\mu_1, \sigma^2 = 1)$	$\mu_1 = \sum_{j=1}^{5} \beta_{j1} y_j$
Gaussian	$Y_2 \sim N(\mu_2, \sigma^2 = 1)$	$\mu_2 = \sum_{j=1}^{5} \beta_{j2} y_j$
Binomial	$Y_3 \sim Binom(1, \pi_3)$	$\pi_3 = \frac{\exp(\sum_{j=1}^{5} \beta_{j3} y_j)}{1+\exp(\sum_{j=1}^{5} \beta_{j3} y_j)}$
Binomial	$Y_4 \sim Binom(1, \pi_4)$	$\pi_4 = \frac{\exp(\sum_{j=1}^{5} \beta_{j4} y_j)}{1+\exp(\sum_{j=1}^{5} \beta_{j4} y_j)}$
Gaussian	$Y_5 \sim N(\mu_5, \sigma^2 = 1)$	$\mu_5 = 0$

Table 1 Model assumption for random variables represented in the DAG in Figure 1. There are three continuos Gaussian random variables and two binomial random variables. Regression coefficients are given in Table 2.

To generate $N = 100$ independent samples with structure given in Figure 1 and conditional mean and distribution given in Table 1, we consider the following procedure:

- Generate Y_5 from a normal with mean zero and variance one. Then, calculate π_4 and π_3 and generate Y_4 and Y_3.
- Calculate μ_2 and generate Y_2. Then, calculate μ_1 and generate Y_1.
- Repeat the process 100 times.

Table 3 shows the relative importance according to AIC ranking (AIC-ranking column) and the score calculated according to standard regression coefficients (SC column). There are 10 possible links for an undirected graphical model. According to AIC ranking, the first link to be drawn in the tree is the link between variables Y_3

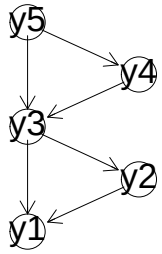

Fig. 1 Directed graph.

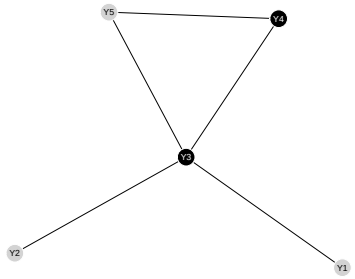

Fig. 2 Recovered graph.

and Y_5. The selected strongly decomposable graphical model is shown in Figure 2. It seems that ranking the links according to regression coefficients can give a more information on the relative importance of each link. In fact, from column SC in Table 3 we can see that regression-type graphical model would order the coefficients almost in the same order as the original coefficients.

β	1	2	3	4	5
1	0	0	0	0	0
2	0.01	0	0	0	0
3	0.31	0.45	0	0	0
4	0	0	0.98	0	0
5	0	0	0.69	0.72	0

Table 2 Regression coefficients

	V_i	V_j	AIC-ranking	SC
1	3	5	192.36	0.31
2	4	5	148.53	0.24
3	3	4	144.53	1.17
4	2	3	58.82	0.16
5	1	3	34.34	0.18
6	2	5	33.52	0.06
7	2	4	30.07	0.086
8	1	2	15.68	0.094
9	1	5	5.55	0.015
10	1	4	-1.60	0.07

Table 3 Graph edge ordering and standardized regression coefficients

3 Simulation Study

We perform a simulation study to compare the performance of graphical lasso to decomposable graphical models, in terms of recovering of the graph. The support recovery of the graph is evaluated by the following scores:

$$\text{PPR} = \frac{TP}{TP+FP}, \qquad \text{Sensitivity} = \frac{TP}{TP+FN},$$

and

$$\text{MCC} = \frac{(TP \times TN) - (FN \times FP)}{\sqrt{(TP \times TN)(FN \times FP)}},$$

where TP are the true positive, FP are the false positive, TN are the true negative and FN are the false negative. The larger the score value, the better the classification performance.

The "best" graph structures are estimated in terms of AIC (minForest-aic) and BIC (minForest-bic) for the decomposable graphical models. Whereas, for the graphical lasso we select the graph according to stability selection procedure [15].

We consider five models as follows:

- Model 1. A banded graph with bandwidth equal to 1.
- Model 2. A cluster graph where the number of cluster is about $p/20$ if $p > 40$ and 2 if $p \leq 40$. For cluster graph, the value 3/d is the probability that a pair of nodes has an edge in each cluster.
- Model 3. An hub graph where the number of hubs is about $p/20$ if $p > 40$ and 2 if $p \leq 40$.
- Model 4. A random graph where the probability that an edge is present between two nodes is $3/p$.
- Model 5. A scale-free graph where an edge is present with probability 0.9.

We use the function `huge.generator` of the R package `huge` to generate these graphical structures [16]. We keep the structure of the graph fixed and simulate $n = 100$ independence samples from a multivariate distribution with $\mu = 0$ and $\Sigma = K^{-1}$ where zero elements in K are absent links. For each model, we generate a sample of size $n = 100$ from a multivariate normal distribution We consider different values of $p = (10, 30, 50, 100, 150)$ and 100 replicates. We report the results for the support recovery of the precision matrix together with an example of the graph structures of each of the five models in Appendix.

The main conclusion which can be drawn from the results reported in the tables is that the strongly decomposable graphical model show, generally, comparable or better performance both in lower and high-dimensional case. We would expect minForest-bic have better results than minForest-aic but this doesn't appear in our simulation study. The glasso-stars performs worse than minForest-aic and minForest-bic for banded graphs and hub graphs. This could be due to the particular structure of the graph and it should not be linked with the selection method. In other words, it seems to be a limitation of the glasso.

4 Analysis of breast cancer data

In this section we analyze a breast cancer dataset. The data come from a study performed on 62 biopsies of breast cancer patients over 59 genes. These genes were identified using comparative genomic hybridization. Continuous measures of expression levels of those 59 genes were collected. In order to link gene amplifica-

tion/deletion information to the aggressiveness of the tumors in this experiment, clinical information is available about each patient: age at diagnosis (AGE), follow-up time (Surv.Time), whether or not the patient died of breast cancer (C.Death), the grade of the tumor (C.Grade), the size of the tumor (Size.Lesion), and the Nottingham Prognostic Index (NPI). C.Death is a dichotomous variable, C.Grade is ordinal with three categories and NPI is a continuous index used to determine prognosis following surgery for breast cancer. NPI values are calculated using three pathological criteria: the size of the lesion; the number of involved lymph nodes; and the grade of the tumor. The complete dataset results in 62 units and 65 variables.

Our aim is to find a network which may underline some important relationships between the 65 variables. These variables comprise both gene expression levels and clinical variables. We use the package `gRapHD` [17] to analyse the breast cancer data. Firstly, the forest that minimizes the BIC is found by applying the function `minForest`. This result in a quite simple graph with at last 64 links. A more complex model can be found by applying the function `stepw`. This function performs a forward search strategy through strongly decomposable models starting from a given decomposable graphical model. At each step, the edge giving the greatest reduction in BIC is added. The process ends when no further improvement is possible.

Figure 3 shows the graph for the homogeneous strongly decomposable graphical model applied to the breast cancer data with starting point a minimum BIC forest with a link between C.Grade and C.Death. Black nodes indicate discrete variables while grey nodes represent continuous variables. The graph in Figure 3 indicates that Gene 4 is the connection between Surv.Time, C.Death, C.Grade, NPI and Size.Lesion and the gene expression levels. Gene 4 separates two blocks of genes the one represented in the top part of Figure 3 and the one represented in the bottom part of the same figure. The other most connected genes are Gene 12 and Gene 49 with 8 and 9 nodes, respectively. C.grade and Size.Lesion are linked to NPI as we expected and there is a short path between NPI and Survival time.

5 Discussion

In this paper, we have explored a class of graphical models, the strongly decomposable graphical models, which can be used to infer networks for high-dimensional mixed data. Results from the simulation study shows comparable or better performance in terms of graphs recovering with respect to graphical lasso. There are some limitations. The first one is due to the assumption of decomposable models, namely neither cycle of length more than 3 nor forbidden path can be estimated. The second one is due to the distributional assumption. In fact, the conditional Gaussian distribution cannot take into account dependence of a continuos variable to a discrete one. So, careful attention should be paid during the analysis of real data. In the real data analysis, in which mixed data are present, we have shown that a relation between gene expression levels and clinical conditions of the patients seems to be present. We have not dealt with parameter estimation which is indeed another challenge task

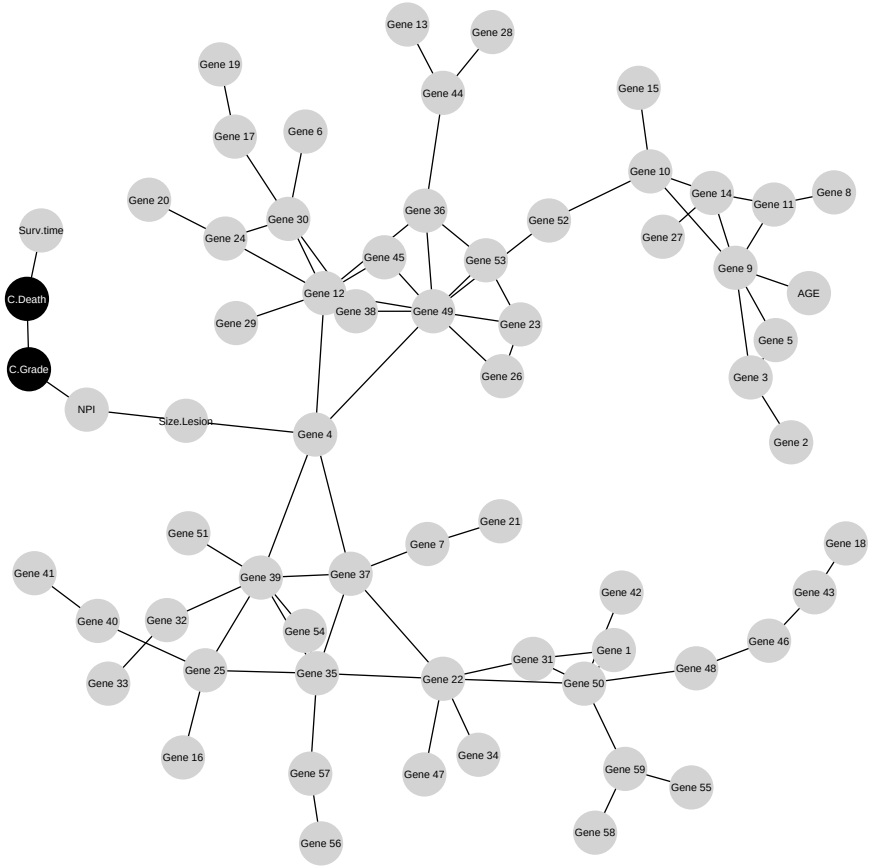

Fig. 3 Graph obtained by applying the homogeneous strongly decomposable graphical model to breast cancer data with starting point a minimum BIC forest with a link between C.Grade and C.Death. Black dot nodes indicate discrete variables while circle grey nodes represent continuous variables.

for high-dimensional data. To conclude, the main advantages of using strongly decomposable graphical models we have illustrated in this paper are: i) their feasibility for high-dimensional setting; ii) the facility to communicate the results by showing the graph; iii) the possibility to catch patters in terms of clustering, hubs, important variables from the conditional independent graph. Moreover, regression-type graphical models can give some insight on the ordering of importance for some of the regressors.

References

1. Lauritzen S.L.: Graphical models. Oxford University Press (1996)
2. Edwards D.: Introduction to graphical modelling. Springer (2000)
3. Whittaker J.: Graphical models in applied multivariate statistics. Wiley Publishing (2009)
4. De Jong, H.: Modeling and simulation of genetic regulatory systems: a literature review. Journal of computational biology **9**, 67–103 (2002)
5. Meinshausen, N., Buhlmann, P. High-dimensional graphs and variable selection with the lasso. The Annals of Statistics **34**, 1436–1462 (2006)
6. Yuan, M., Lin, Y. Model selection and estimation in the Gaussian graphical model. Biometrika **94**, 19–35 (2007)
7. Banerjee, O., El Ghaoui, L., d'Aspremont, A. Sparse inverse covariance estimation with the graphical lasso. The Journal of Machine Learning Research **9**, 485–516 (2008)
8. Friedman, J., Hastie, T., Tibshirani, R. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Biostatistics **9**, 432–441(2008)
9. DREAM8 (2014). Rheumatoid Arthritis Responder Challenge. https://www.synapse.org. Cited 11 Feb 2014.
10. Lauritzen, S.L., Wermuth, N. Graphical models for associations between variables, some of which are qualitative and some quantitative. The Annals of Statistics 31–57(1989)
11. Hoff, P.D. Extending the rank likelihood for semiparametric copula estimation. The Annals of Applied Statistics **1**, 265–283 (2007)
12. Edwards, D., de Abreu, G., Labouriau, R. Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. BMC bioinformatics 11:18 (2010)
13. Fellinghauer, B., Bühlmann, P., Ryffel, M., Von Rhein, M., Reinhardt, J.D. Stable graphical model estimation with Random Forests for discrete, continuous, and mixed variables. Computational Statistics & Data Analysis **64**, 132–152 (2013)
14. Chow, C., Liu, C. Approximating discrete probability distributions with dependence trees. Information Theory, IEEE Transactions on **14**, 462–467 (1968)
15. Wasserman, L., Roeder, K. Liu, H. Stability approach to regularization selection (stars) for high dimensional graphical models. Advances in Neural Information Processing Systems (2010)
16. Zhao, T., Liu, H., Roeder, K., Lafferty, J., Wasserman, L.. Huge: High-dimensional Undirected Graph Estimation R Package. http://CRAN.R-project.org/package=huge Cited 11 Feb 2014.
17. Abreu,G.C.G., Edwards, D., Labouriau, R. High-Dimensional Graphical Model Search with the gRapHD R Package. Journal of Statistical Software **37**, 1–18 (2010)

Appendix

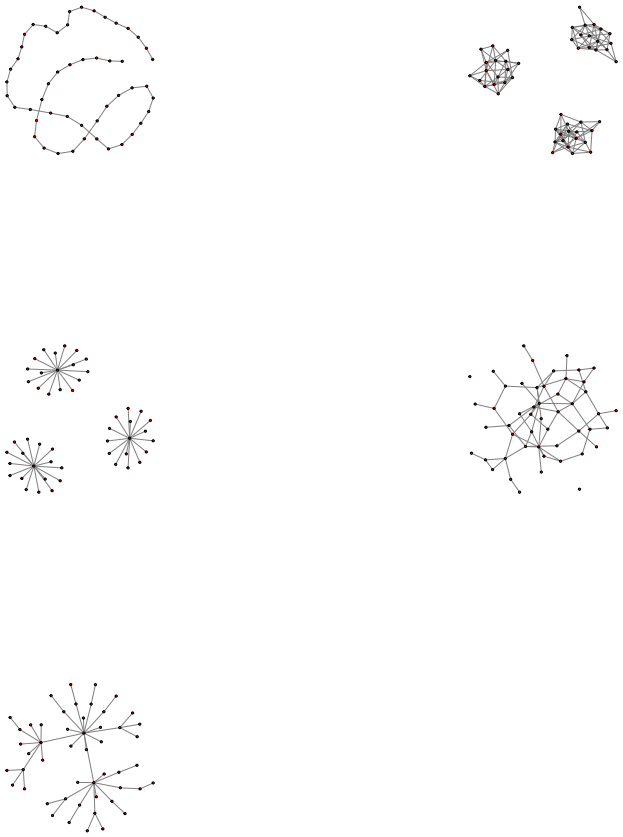

Fig. 4 Model structures from which we generate data. These graphs are described as models in this section and are named banded graph, cluster hub, random and scale free. All the graphs are sparse.

Table 4 Model 1 - Banded Graph

	glasso-stars	*minForest-aic*	*minForest-bic*
p		**PPV**	
10	88.88 (23.81)	98.38 (4.14)	96.18 (5.77)
30	67.55 (5.31)	96.72 (3.27)	94.06 (4.35)
50	56.04 (3.86)	95.69 (2.71)	93.16 (3.47)
100	40.99 (1.92)	94.23 (1.92)	91.35 (2.52)
150	32.23 (1.25)	93.42 (1.86)	90.53 (2.22)
		Sensitivity	
10	60.56 (35.57)	98.67 (3.63)	99.00 (3.20)
30	99.31 (1.70)	97.03 (3.02)	97.24 (2.90)
50	99.49 (1.10)	95.84 (2.61)	96.00 (2.44)
100	99.55 (0.62)	94.40 (1.86)	94.63 (1.79)
150	99.56 (0.50)	93.57 (1.84)	93.83 (1.70)
		MCC	
10	66.87 (27.41)	98.14 (4.77)	96.91 (5.02)
30	80.38 (3.63)	96.65 (3.32)	95.30 (3.58)
50	73.38 (2.84)	95.58 (2.76)	94.33 (2.86)
100	62.91 (1.55)	94.20 (1.92)	92.82 (2.05)
150	55.82 (1.12)	93.41 (1.87)	92.06 (1.88)

Table 5 Model 2 - Cluster

	glasso-stars	*minForest-aic*	*minForest-bic*
p		**PPV**	
10	90.50 (29.04)	90.98 (3.64)	96.73 (4.57)
30	74.29 (5.11)	79.99 (5.98)	77.77 (5.64)
50	67.43 (3.68)	82.37 (4.26)	78.55 (4.47)
100	52.52 (2.20)	73.99 (3.94)	71.14 (3.73)
150	43.84 (1.51)	74.20 (2.89)	71.60 (2.95)
		Sensitivity	
10	10.75 (6.90)	56.25 (12.48)	93.00 (12.23)
30	36.73 (5.72)	27.24 (2.35)	32.92 (3.72)
50	54.39 (4.02)	31.27 (1.73)	35.45 (2.52)
100	50.04 (2.31)	25.57 (1.38)	28.85 (1.63)
150	52.63 (1.98)	26.56 (1.12)	29.36 (1.40)
		MCC	
10	23.11 (10.26)	57.90 (11.32)	91.23 (12.05)
30	44.68 (4.55)	40.46 (4.19)	43.78 (4.53)
50	56.35 (3.13)	47.71 (2.86)	49.41 (3.00)
100	48.30 (1.96)	41.71 (2.41)	43.35 (2.28)
150	45.79 (1.48)	43.25 (1.83)	44.61 (1.89)

Table 6 Model 3 - Hub

	glasso-stars	*minForest-aic*	*minForest-bic*
p		**PPV**	
10	89.54 (16.62)	88.14 (3.84)	89.08 (8.42)
30	65.20 (6.42)	84.73 (5.74)	75.94 (6.05)
50	51.22 (3.92)	79.59 (4.61)	71.52 (4.78)
100	35.98 (2.18)	75.73 (4.36)	66.75 (3.56)
150	28.38 (1.29)	73.95 (3.42)	65.48 (2.87)
		Sensitivity	
10	74.12 (31.99)	98.88 (3.60)	99.25 (2.98)
30	90.50 (6.28)	88.57 (5.81)	90.61 (5.46)
50	91.00 (4.51)	83.64 (4.90)	85.77 (4.78)
100	92.56 (3.20)	79.60 (4.70)	82.34 (4.61)
150	92.75 (2.53)	78.09 (3.65)	80.29 (3.90)
		MCC	
10	75.89 (22.22)	91.84 (4.31)	92.53 (6.18)
30	74.85 (5.05)	85.68 (6.13)	81.63 (5.46)
50	66.66 (3.56)	80.84 (4.92)	77.36 (4.55)
100	56.53 (1.97)	77.19 (4.61)	73.57 (3.88)
150	50.36 (1.56)	75.67 (3.57)	72.11 (3.20)

Table 7 Model 4 - Random

	glasso-stars	*minForest-aic*	*minForest-bic*
p		**PPV**	
10	92.40 (22.25)	95.61 (7.02)	90.15 (8.96)
30	68.28 (5.92)	77.75 (5.37)	79.31 (6.02)
50	62.82 (3.90)	81.79 (4.41)	79.96 (5.06)
100	48.74 (2.66)	74.87 (3.84)	71.91 (4.05)
150	36.80 (1.71)	62.68 (3.80)	60.62 (3.82)
		Sensitivity	
10	21.19 (10.09)	54.25 (4.34)	57.31 (5.69)
30	83.94 (5.82)	69.09 (5.26)	72.00 (6.35)
50	80.86 (4.76)	61.88 (3.31)	62.88 (3.41)
100	73.56 (3.68)	46.54 (2.33)	47.39 (2.44)
150	65.00 (3.17)	38.41 (2.32)	39.19 (2.28)
		MCC	
10	35.57 (11.24)	63.06 (7.41)	61.32 (6.98)
30	73.40 (4.74)	71.24 (5.64)	73.61 (5.74)
50	69.43 (3.73)	69.78 (4.00)	69.48 (4.16)
100	58.25 (2.71)	58.00 (3.06)	57.29 (3.11)
150	47.43 (2.09)	48.21 (3.02)	47.84 (2.98)

Table 8 Model 5 - Scale free

	glasso-stars	*minForest-aic*	*minForest-bic*
p		**PPV**	
10	92.46(23.94)	95.51 (6.41)	88.63 (9.22)
30	68.44(7.04)	72.87 (8.67)	62.23 (6.45)
50	48.81(4.95)	54.98 (7.87)	47.40 (4.47)
100	27.23(2.55)	33.57 (5.44)	30.78 (3.49)
150	17.75(1.78)	23.86 (4.06)	22.97 (3.12)
		Sensitivity	
10	41.00(24.96)	95.89 (6.25)	96.56 (5.63)
30	68.93(9.54)	73.97 (8.84)	76.83 (9.18)
50	61.47(7.60)	55.67 (8.14)	58.88 (8.98)
100	51.39(6.03)	33.93 (5.65)	37.40 (6.51)
150	45.70(5.99)	24.06 (4.11)	27.48 (5.00)
		MCC	
10	55.44(22.77)	94.61 (7.79)	90.39 (8.27)
30	66.28(7.36)	71.50 (9.36)	66.65 (7.61)
50	52.59(5.66)	53.45 (8.33)	50.59 (6.40)
100	35.67(3.78)	32.39 (5.65)	32.43 (4.83)
150	26.98(3.24)	22.93 (4.14)	24.01 (4.00)