Markov Network Structure Learning via Ensemble-of-Forests Models

Eirini Arvaniti[1] and Manfred Claassen[1†]

[1]Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland.
[†]Electronic correspondence: claassen@imsb.biol.ethz.ch

Abstract

Real world systems typically feature a variety of different dependency types and topologies that complicate model selection for probabilistic graphical models. We introduce the *ensemble-of-forests* model, a generalization of the *ensemble-of-trees* model of Meilă and Jaakkola (2006). Our model enables structure learning of Markov random fields (MRF) with multiple connected components and arbitrary potentials. We present two approximate inference techniques for this model and demonstrate their performance on synthetic data. Our results suggest that the ensemble-of-forests approach can accurately recover sparse, possibly disconnected MRF topologies, even in presence of non-Gaussian dependencies and/or low sample size. We applied the ensemble-of-forests model to learn the structure of perturbed signaling networks of immune cells and found that these frequently exhibit non-Gaussian dependencies with disconnected MRF topologies. In summary, we expect that the ensemble-of-forests model will enable MRF structure learning in other high dimensional real world settings that are governed by non-trivial dependencies.

1 INTRODUCTION

This work presents the ensemble-of-forests model for approximate structure learning in Markov random fields (MRF). As opposed to most existing MRF structure learners that either work with specific types of potentials (e.g. discrete, Gaussian) or assume connected MRF topology (Lin et al., 2009), our approach is applicable for MRFs with arbitrary potentials and topology, including disconnected topologies, and is therefore suited to accommodate a wide range of real world settings.

Markov random fields (MRF) are undirected probabilistic graphical models specifying conditional independence relations among a set of random variables. Learning MRFs involves parameter inference and model selection, i.e. learning the underlying graph structure. For general MRFs, exact parameter inference is difficult due to the necessity to evaluate the intractable partition sum and therefore is addressed by approximate inference approaches. Structure learning is an even more difficult task. The naive method of enumerating all possible topologies is prohibitively expensive and, thus, alternative approaches have been proposed based on independence tests or approximate score-based methods Koller and Friedman (2009).

Currently, the prevalent approach to model continuous random variables is to assume Gaussianity. Under this hypothesis, the Gaussian Markov random field (GMRF) structure can be directly read from the inverse covariance matrix (Koller and Friedman, 2009): zero entries exactly correspond to conditional independence statements of the Markov random field. Sparse inverse covariance selection constitutes a convex relaxation of the structure learning task for GMRFs that can be solved efficiently (Banerjee et al., 2006; Friedman et al., 2008).

Random variables of real world systems typically exhibit unusual dependency types (Trivedi and Zimmer, 2005; Berkes et al., 2008) that are not appropriately captured by the Gaussian potentials of GMRFs. *Copula* potentials constitute a more general and expressive alternative to deal with non-Gaussian dependency types. Copulas are multivariate distributions that encode the dependencies among random variables. Copula models are very flexible, as they enable researchers to independently specify the marginal distributions of random variables and their dependency structure. Liu et al. (2009) define MRFs with semi-parametric Gaussian copula potentials. Approximate structure learning in this model is tractable because the dependency type is Gaussian and, thus, parameter inference is easy and model selection can also be efficiently approximated by resorting to sparse inverse covariance estimation. However, in MRFs with general copula potentials, even parameter estimation is difficult because of the intractable partition sum. This situation entails that structure learning is also difficult.

The intractability of exact inference for MRFs with general copula potentials has motivated alternative approaches based on approximate inference. Meilă and Jaakkola (2006) introduced the *ensemble-of-trees* (ET) model that enables approximate inference for both parameter estimation and structure learning of general MRFs. A Markov network is represented as a mixture model whose components are tree-structured distributions defined over all possible spanning trees of the underlying graph. Despite the super-exponential number of such trees, the model remains tractable by defining conveniently decomposable priors over the structure and parameters of tree-distributions. Recently, Kirshner (2008) presented a tree-averaged density model based on tree structured MRFs with copula potentials. The tasks of parameter estimation and structure learning are jointly expressed as a single (non-convex) objective, which is optimized via Expectation-Maximization. Lin et al. (2009) utilize the ET model for structure learning of GMRFs and empirically demonstrate superior performance compared to sparse inverse covariance selection for limited sample size. Above considerations render copula MRFs as attractive models because they are more general than GMRFs and efficient learning approaches exist for them.

Real world systems with many random variables are frequently best represented by MRFs that decompose into several connected components. In biology, for instance, a specific stimulus might activate competing, independent signaling pathways each including its own MRF component (Johnstone et al., 2008). However, the ET structure learning approach is not able to recover disconnected topologies since it is averaging over ensembles of spanning trees. It is desirable to generalize the ET approach in order to overcome this limitation and, thereby, still benefit from the expressiveness of copula MRFs in these real world settings.

The main contribution of this work is the generalization of the ET model to the *ensemble-of-forests* (EF) model that explicitly accounts for graph topologies with multiple connected components. In the proposed model, a Markov network is represented as a mixture of forests, i.e. collections of tree-structured MRFs. An implementation of the exact model is intractable, as the averaging over all possible forests results in a hard combinatorial problem. Instead, we present approximate formulations of the structure learning task. The rest of this paper is organized as follows. In Sections 2 – 3 we formally introduce the methods that we build upon. Then, in Sections 4 – 6 we describe the *ensemble-of-forests* model and present benchmark results on synthetic datasets. In Sections 7 – 8 we apply our method to plant microarray and immune cell perturbation data. Finally, Section 9 concludes with a short discussion.

2 COPULA MODELS

This section reviews the application of copulas to describe general multivariate distributions and/or potentials in MRFs. Copulas are multivariate continuous distributions defined on the unit hypercube, $C : [0,1]^d \to [0,1]$, with uniform univariate marginals. Let $X_1, \ldots, X_d$ be real random variables with joint cumulative distribution function (cdf) $F(\mathbf{x})$ and marginally distributed as $F_1(x_1), \ldots, F_d(x_d)$ respectively. Then, the random variables $U_1 = F_1(x_1), \ldots, U_d = F_d(x_d)$ are uniformly distributed on $[0,1]$. This property forms the basis for Sklar's theorem, according to which any joint distribution $F(x_1, \ldots, x_d)$ with continuous marginals can be uniquely expressed as

$$F(x_1, \ldots, x_d) = C(F_1(x_1), \ldots, F_d(x_d)). \qquad (1)$$

The converse is also true: arbitrary univariate marginals $\{F_i\}$ can be combined using a copula function C to uniquely construct a valid joint distribution with marginals $\{F_i\}$. The copula function C exclusively encodes the dependencies among random variables.

Furthermore, copula density functions $c(\mathbf{u}) = \dfrac{\partial^d C(\mathbf{u})}{\partial u_1 \ldots \partial u_d}$ can be expressed in terms of probability density functions as

$$c(u_1, \ldots, u_d) = \frac{f(x_1, \ldots, x_d)}{\prod_{i=1}^d f_i(x_i)}. \qquad (2)$$

A large number of copula functions have been proposed in the literature (Nelsen, 1999), especially for the bivariate case. Commonly used examples are the Clayton, Gumbel, Frank, Gaussian and Student's t parametric copula families. In **Figure** 1, we present contour plots of six distributions with standard Gaussian marginals but different types of dependencies between the marginals. In each case, the dependency structure is specified via a different copula function.

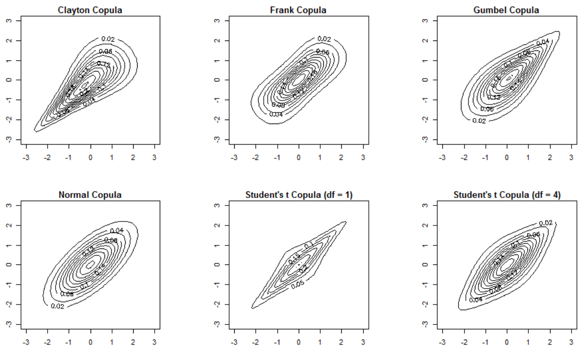

Figure 1: Contour plots of six joint distributions defined using standard Gaussian marginals and different dependency structures specified by different copulas.

Bivariate copulas are typically used to model strong extreme-value dependencies in financial data (Embrechts et al., 2003; Trivedi and Zimmer, 2005). Recently, the probabilistic graphical model framework has been successfully employed for the construction of copula-based high-dimensional models. A review on this topic can be found in (Elidan, 2013).

3 ENSEMBLE-OF-TREES MODELS

Here we introduce the ensemble-of-trees (ET) method for approximate parameter inference and structure learning of MRFs. This method forms the basis for the ensemble-of-forests method, the main conceptual contribution of this paper. From here on, we adopt the following notation: we consider a Markov network encoded by a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the set of nodes corresponding to random variables $\mathcal{X} = \{X_1, \ldots, X_d\}$ and $\mathcal{E}$ is the set of edges.

The ensemble-of-trees model of Meilă and Jaakkola (2006) is an approximate inference approach to carry out structure learning for MRFs with "inconvenient" potentials. It constitutes a mixture model over all possible spanning trees of the complete graph over the nodeset $\mathcal{V}$. A prior distribution over spanning tree structures T is defined as

$$p_\beta(T) = \frac{1}{Z_\beta} \prod_{e_{uv} \in T} \beta_{uv} \tag{3}$$

where each parameter $\beta_{uv} = \beta_{vu} \geq 0$, for all $u \neq v$, $u, v \in \mathcal{V}$ can be interpreted as a weight for edge e_{uv}, directly proportional to the probability of appearance of that edge.

$Z_\beta = \sum_T \prod_{e_{uv} \in T} \beta_{uv}$ is a normalizing constant, ensuring that the prior constitutes a valid probability distribution. It turns out that Z_β can be efficiently computed. Defining the matrix $\mathbf{Q}(\boldsymbol{\beta})$ as the first $d-1$ rows and columns of the Laplacian matrix

$$L_{uv} = \begin{cases} -\beta_{uv} & \text{if } u \neq v, \\ \sum_k \beta_{uk} & \text{if } u = v \end{cases} \tag{4}$$

Meilă and Jaakkola (2006) generalize Kirchhoff's Matrix-Tree theorem for binary weights and show that

$$Z_\beta = \sum_T \prod_{e_{uv} \in T} \beta_{uv} = |\mathbf{Q}(\boldsymbol{\beta})|. \tag{5}$$

This result makes the averaging over all possible (d^{d-2}) spanning tree structures computationally tractable.

Assuming a prior tree structure T, the conditional distribution of a data sample $\mathbf{x}$ can be expressed as

$$p(\mathbf{x}|T, \boldsymbol{\theta}) = \prod_{v \in \mathcal{V}} \theta_v(x_v) \prod_{e_{uv} \in T} \frac{\theta_{uv}(x_u, x_v)}{\theta_u(x_u)\theta_v(x_v)} \tag{6}$$

where the parameter vector $\boldsymbol{\theta}$ consists of univariate $\theta_v(x_v)$ and bivariate $\theta_{uv}(x_u, x_v)$ marginal densities defined, respectively, over the nodes and the edges of the tree (Meilă and Jaakkola, 2006). These distributions are assumed invariant for all tree structures.

Finally, after introducing the notation $w_{uv}(\mathbf{x}) = \frac{\theta_{uv}(x_u, x_v)}{\theta_u(x_u)\theta_v(x_v)}$, $w_0(\mathbf{x}) = \prod_{v \in \mathcal{V}} \theta_v(x_v)$ and applying twice the generalized Matrix-Tree theorem we have

$$\begin{aligned} p_\beta(\mathbf{x}) &= \sum_T p_\beta(T) p(\mathbf{x}|T, \boldsymbol{\theta}) \\ &= \frac{w_0(\mathbf{x})}{Z_\beta} \sum_T \prod_{e_{uv} \in T} \beta_{uv} w_{uv}(\mathbf{x}) \\ &= w_0(\mathbf{x}) \frac{|\mathbf{Q}(\boldsymbol{\beta} \otimes \mathbf{w}(\mathbf{x}))|}{|\mathbf{Q}(\boldsymbol{\beta})|} \end{aligned} \tag{7}$$

where the symbol $\otimes$ denotes element-wise multiplication.

The structure learning task in the ET model can be approximated by an empirical estimation of $\boldsymbol{\beta}$, as in (Lin et al., 2009), where $\boldsymbol{\beta}$ is used to approximate the MRF adjacency matrix: non-zero entries β_{uv} correspond to edges in the graph. In our model, we adopt this interpretation of $\boldsymbol{\beta}$.

3.1 ET MODELS WITH DISCONNECTED SUPPORT GRAPH

A mixture model over spanning trees is based on the implicit assumption that the *support graph* of the model is connected. The support graph is a graph that contains exactly the edges corresponding to positive entries in $\boldsymbol{\beta}$. The case of disconnected support graphs is considered by Meilă and Jaakkola (2006) only for *a priori* defined connected components. That is, certain patterns of zero entries in the parameter set $\boldsymbol{\beta}$ predefine a partitioning of nodes into different connected components and these assignments to components cannot be changed e.g. during the course of a structure learning procedure. In this case, each connected component can be treated independently from all others. Assuming k connected components that partition $\mathcal{V}$ into $\{V^1, \ldots, V^k\}$ and introducing the notation

$$\boldsymbol{\beta}_{V^i} = \{\beta_{uv},\ u \neq v,\ u, v \in V^i\}$$

equation (7) is generalized as

$$p_\beta(\mathbf{x}) = w_0(\mathbf{x}) \frac{\prod_{i=1}^k |\mathbf{Q}(\boldsymbol{\beta}_{V^i} \otimes \mathbf{w}_{V^i}(\mathbf{x}))|}{\prod_{i=1}^k |\mathbf{Q}(\boldsymbol{\beta}_{V^i})|} \tag{8}$$

4 ENSEMBLE-OF-FORESTS MODELS

Here we introduce the main contribution of our work, that is the *ensemble-of-forests* (EF) model. This model constitutes an approximate inference approach for structure learning of MRFs with multiple connected components that are not known *a priori*. We assume a nodeset $\mathcal{V}$ of size d and a partition thereof $\mathbf{V} = \{V^1, \dots, V^k\}$. Then, a *maximal forest* or *forest* of size k is a collection of spanning trees $\{T^i\}_{i=1,\dots,k}$, one for each V^i. Extending the ensemble-of-trees model, we introduce a mixture model over all possible forests up to a certain size, i.e. allowing for disconnected structures with a maximal number of k connected components. The limiting cases are $k = 1$, corresponding to the ET model, and $k = d$, corresponding to a model that allows for any possible arrangement of connected components.

The prior probability of a collection of spanning trees $\mathcal{F} := \{T^1, \dots, T^k\}$ is defined as

$$p_\beta(\mathcal{F}) = \frac{1}{Z_\beta} \prod_{T^i \in \mathcal{F}} \prod_{e_{uv} \in T^i} \beta_{uv} \tag{9}$$

where $\beta_{uv} = \beta_{vu} \geq 0$, for all $u \neq v,\ u, v \in \mathcal{V}$. Now, in order to normalize over all possible forests that consist of at most k connected components, the partition function is computed via

$$\begin{aligned} Z_\beta &= \sum_{\mathbf{V} \in part(\mathcal{V})} \sum_{\mathcal{F} \in f(\mathbf{V})} \prod_{T^i \in \mathcal{F}} \prod_{e_{uv} \in T^i} \beta_{uv} \\ &= \sum_{\mathbf{V} \in part(\mathcal{V})} \prod_{V^i \in \mathbf{V}} |\mathbf{Q}(\boldsymbol{\beta}_{V^i})| \end{aligned} \tag{10}$$

where the outer summation $\sum_{\mathbf{V} \in part(\mathcal{V})}$ is performed over all possible partitions of $\mathcal{V}$ into k subsets and the inner summation $\sum_{\mathcal{F} \in f(\mathbf{V})}$ is performed over all maximal forests defined on a specific node partition $\mathbf{V}$. Partitions where some of the subsets V^i are empty are allowed and correspond to graphs with less than k connected components. For example, the partition $\{\mathcal{V}, \emptyset, \dots, \emptyset\}$ represents a fully connected graph. In order to treat such partitions without changing our notation, we define $\mathbf{Q}(\boldsymbol{\beta}_\emptyset) = 1$.

Ignoring the constant term $w_0(\mathbf{x})$, the negative log-likelihood of the model given a dataset $\mathcal{D} = \{x^{(1)}, \dots, x^{(N)}\}$ is written as

$$\begin{aligned} \mathcal{L}(\mathcal{D}\,; \boldsymbol{\beta}) = N \log & \sum_{\mathbf{V} \in part(\mathcal{V})} \prod_{V^i \in \mathbf{V}} |\mathbf{Q}(\boldsymbol{\beta}_{V^i})| \\ & - \sum_{j=1}^{N} \log \sum_{\mathbf{V} \in part(\mathcal{V})} \prod_{V^i \in \mathbf{V}} |\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}^{(j)}_{V^i})| \end{aligned} \tag{11}$$

where $\boldsymbol{\beta}\mathbf{w}^{(j)}_{V^i}$ is a shorthand for $\boldsymbol{\beta}_{V^i} \otimes \mathbf{w}_{V^i}(\mathbf{x}^{(j)})$.

5 LEARNING IN THE EF MODEL

In this section, we describe two approaches for structure learning of Markov networks based on the EF model, namely the *EF-cuts* and *EF-λ* methods. Additionally, we describe common features of the two methods, such as the choice of MRF potentials and the optimization algorithm used for minimizing the learning objective.

5.1 SELECTION OF EDGE POTENTIALS

The first step in learning the EF model is concerned with the choice of the edge potentials $w_{uv}(\mathbf{x})$. Here, we consider continuous distributions as edge potentials. Although we do not explicitly consider discrete distributions in the following, we want to emphasize that learning in the EF model easily extends to this class of potentials. In order to keep our model as generic as possible, we have chosen to use copula-based potentials. Note from Equation (2) that the potentials $w_{uv}(\mathbf{x})$ exactly correspond to bivariate copula densities. In our analysis, we have used the bivariate Clayton, Frank, Gumbel, Gaussian and Student's t copula as candidate parametric families. These copulas have one single parameter to be estimated.

In order to fit a single-parameter copula family to data, we follow a two-step procedure. As a first step, the marginal cdf for each random variable is estimated in a non-parametric approach (Kojadinovic and Yan, 2010) and the obtained estimators, known as pseudo-observations, are plugged into the copula function. Subsequently, the dependence parameter is computed by maximizing the pseudo-likelihood

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log c(\widehat{\mathbf{u}}_\mathbf{i}\,; \boldsymbol{\theta}) \tag{12}$$

where $\widehat{\mathbf{U}}_\mathbf{i}$ is the vector of estimators for the marginals and n is the sample size. The best-fitting copula for each variable pair is selected via cross-validation, where the cross-validation score is based on the pseudo-likelihood of the left-out samples.

5.2 THE EF-cuts HEURISTIC

Graphs with two connected components constitute an important subclass of disconnected networks. Even when restricting ourselves to a maximum of two connected components, it is computationally prohibitive to use the exact ensemble-of-forests model of Equation (11) for sets of random variables of non-trivial size due to the super-exponential number of possible node partitions $part(\mathcal{V})$. Therefore, we resort to heuristic approaches for choosing partitions that are most likely to allow us to recover the true graph structure. For a given parameter configuration $\boldsymbol{\beta}$, we aim to identify a number of high scoring partitions of the nodeset and then average over these partitions only.

Our heuristic is based on the intuition that edges e_{uv} with small β_{uv} are assigned a low prior probability and, therefore, are expected to be most likely not present in the true

MRF. Therefore, we would like to prioritize partitions generated by dropping these low-weight edges. Following that intuition, we derive a scoring system based on systematic enumeration of minimum cuts.

A *cut* of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is a partition of $\mathcal{V}$ into subsets A, $B = \mathcal{V} - A$. The weight of a cut is the sum of the weights of all edges crossing the cut. Starting with the minimum-weight cut, we want to enumerate a ranked set of graph cuts of increasing weight. An efficient algorithm (Vazirani and Yannakakis, 1992) exists for this task. In our case, edge weights correspond to the structural parameters $\boldsymbol{\beta}$. Let (A, B) denote a cut and let $\mathcal{C}$ denote the set of M minimum-weight cuts in the graph. Since we are only considering graphs with at most two connected components, a forest $\mathcal{F}$ consists of two spanning trees T_A, T_B. To simplify our notation, we include the case of connected graphs as a special case where $A = \mathcal{V}$ and $B = \emptyset$. This is a special cut of zero weight and is always included in $\mathcal{C}$. We perform structure learning by minimizing the negative log-likelihood of the model with respect to $\boldsymbol{\beta}$. The respective objective is derived from Equation (11) by setting $k = 2$ and only considering partitions that belong to the set $\mathcal{C}$. The optimization problem can be formulated as

$$\min_{\boldsymbol{\beta}} N \log \sum_{(A,B)\in\mathcal{C}} |\mathbf{Q}(\boldsymbol{\beta}_A)||\mathbf{Q}(\boldsymbol{\beta}_B)| - \sum_{j=1}^{N} \log \sum_{(A,B)\in\mathcal{C}} |\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}_A^{(j)})||\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}_B^{(j)})| \qquad s.t. \quad \beta_{uv} \geq 0 \quad u, v \in \mathcal{V}, \quad u \neq v. \tag{13}$$

Let us denote $\mathcal{C}'$ the set of partitions where nodes u, v belong to the same connected component. The set of partitions where u, v belong to different components has no contribution to the gradient $(\nabla_\beta f)_{uv}$. Without loss of generality, we will assume that if nodes u, v belong to the same partition set, then this is set A and the other set is $B = \mathcal{V} - A$. Then the gradient of the objective (13) follows as

$$(\nabla_\beta f)_{uv} = N \frac{\sum\limits_{(A,B)\in\mathcal{C}'} M_{uv}(\boldsymbol{\beta}_A)|\mathbf{Q}(\boldsymbol{\beta}_A)||\mathbf{Q}(\boldsymbol{\beta}_B)|}{\sum\limits_{(A,B)\in\mathcal{C}'} |\mathbf{Q}(\boldsymbol{\beta}_A)||\mathbf{Q}(\boldsymbol{\beta}_B)|} - \sum_{j=1}^{N} w_{uv}^{(j)} \frac{\sum\limits_{(A,B)\in\mathcal{C}'} M_{uv}(\boldsymbol{\beta}_A)|\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}_A^{(j)})||\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}_B^{(j)})|}{\sum\limits_{(A,B)\in\mathcal{C}'} |\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}_A^{(j)})||\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}_B^{(j)})|} \tag{14}$$

where M is defined as in (Meilă and Jaakkola, 2006)

$$M_{uv} = \begin{cases} Q^{-1}_{uu} + Q^{-1}_{vv} - 2Q^{-1}_{uv} & \text{if } u \neq v, u \neq w, v \neq w, \\ Q^{-1}_{uu} & \text{if } u \neq v, v = w, \\ Q^{-1}_{vv} & \text{if } u \neq v, u = w, \\ 0 & \text{if } u = v. \end{cases} \tag{15}$$

With w we denote the index of the row and column that are removed from the Laplacian matrix of Equation (4) in order to obtain Q.

The min-cut heuristic is a feasible approximation to structure learning of MRFs with disconnected topologies. However, it is practically restricted to graph structures with at most two connected components. Furthermore, the approach does not scale with increasing node or sample size due to the complicated objective and gradient functions. These considerations limit its applicability to real world scenarios.

5.3 THE EF-λ HEURISTIC

In the following, we introduce the EF-λ heuristic that scales well with dimensionality and number of connected components of the underlying MRF. The starting point is again equation (11), but now we drop the summation $\sum_{\mathbf{V}\in part(\mathcal{V})}$ over possible node partitions. Instead, we only consider a single partition $\mathbf{V}$. Additionally, we impose an L_1 penalty term on the structural parameters $\boldsymbol{\beta}$ to encourage sparse solutions. The new optimization task is expressed as

$$\min_{\boldsymbol{\beta}} N \sum_{V^i\in\mathbf{V}} \log|\mathbf{Q}(\boldsymbol{\beta}_{V^i})| - \sum_{j=1}^{N} \sum_{V^i\in\mathbf{V}} \log|\mathbf{Q}(\boldsymbol{\beta}\mathbf{w}_{V^i}^{(j)})| + \lambda\|\boldsymbol{\beta}\|_1 \qquad s.t.\ \beta_{uv} \geq 0 \quad u, v \in \mathcal{V},\ u \neq v. \tag{16}$$

An iterative optimization procedure is employed to minimize the objective (16). At each iteration step, summation is performed over maximal forests defined for the single node partition $\mathbf{V}$ that is induced by the current iterate $\boldsymbol{\beta}$. The number of connected components does not need to be fixed. The penalty term has the critical role of controlling sparsity and, thus, allowing structures with multiple connected components to be considered.

A similar L_1-regularized approach cannot be employed for the ET model, because the ET objective is not defined for all sparsity patterns in $\boldsymbol{\beta}$. Therefore, there is effectively no sparsity induction by an L_1 penalty in ET. Furthermore, for some iterative optimization procedures, numerical instabilities might occur if $\boldsymbol{\beta}$ is temporarily set to an invalid value.

The gradient of the objective for the EF-λ takes a simple form. Considering the non-negativity of $\boldsymbol{\beta}$, the L_1-norm $\|\boldsymbol{\beta}\|_1$ is equal to $\sum_{u,v\in\mathcal{V},\, u\neq v} \beta_{uv}$. Thus, the objective is differentiable at all points. Assuming that $\boldsymbol{\beta}$ induces a partitioning of $\mathcal{V}$ into $\{V^1, \dots, V^k\}$, the gradient of the objective can be expressed as

$$(\nabla_{\beta} f)_{uv} = N M_{uv}(\boldsymbol{\beta}_{V^i}) - \sum_{j=1}^{N} w_{uv}^{(j)} M_{uv}(\boldsymbol{\beta} \mathbf{w}_{V^i}^{(j)}) + \lambda \tag{17}$$

for $u, v \in V^i$ and is equal to 0 otherwise.

The choice of the regularization parameter λ is an important aspect of the EF-λ approach. We optimize the EF-λ objective using different penalty parameters $\lambda = \exp(-\rho)$, where ρ takes values in the interval $[3, 6]$ with a step of 0.1. The optimal λ is selected so as to minimize the extended Bayesian Information Criterion (eBIC) (Foygel and Drton, 2010) defined as

$$eBIC = 2\mathcal{L} + |E| \log n + 4|E| \gamma \log d \tag{18}$$

where $\mathcal{L}$ is the negative log-likelihood of the model, $|E|$ is the number of non-zero predicted $\boldsymbol{\beta}$ entries, n is the sample size, d is the number of nodes and γ is an additional penalty term imposed on more complex structures. The classical Bayesian Information Criterion is obtained as a subcase for $\gamma = 0$. We performed simulations with different values of γ in the interval $[0, 1]$ and resulted in using $\gamma = 0.5$.

5.4 OPTIMIZATION OF THE LEARNING OBJECTIVE

The objectives (13) and (16) to fit the EF model are nonconvex functions. Therefore, there is no guarantee of convergence to a global optimum and the initial point for optimization has to be carefully chosen. Lin et al. (2009) initialize $\boldsymbol{\beta}$ with an upper-bound obtained by optimizing a convex sub-expression of the full objective. Our preliminary experiments confirmed that this method yielded significantly better optima than random initializations. Therefore, we adopted this choice for initialization. As for the main optimization task, we have used the Spectral Projected Gradient (SPG) algorithm (Varadhan and Gilbert, 2009), a gradient-based method that allows for simple box constraints.

6 BENCHMARK ON SIMULATED DATA

In this section, we evaluate the empirical performance of our proposed EF approximations via comparison to the ET (Lin et al., 2009) and glasso (Friedman et al., 2008) algorithms on synthetic Gaussian and non-Gaussian data. We use the glasso implementation from the R-package `huge` (Zhao et al., 2012). The glasso regularization term is obtained via Stability Approach to Regularization Selection (StARS) (Liu et al., 2010), a criterion based on variability of the graphs estimated by overlapping subsamplings. We employ this criterion, since it achieves the best performance in our simulations. For the ET and EF approaches we use Gaussian copula or Student's t-copula potentials and optimize the corresponding objective via SPG. For the EF-cuts method, we consider the first 50 minimum-weight cuts.

6.1 RESULTS ON GAUSSIAN MRF DATA

We first aim at confirming that the EF model achieves comparable performance to state-of-the-art methods for MRF structure learning. To this end, we generated Gaussian MRF data following the procedure described in (Lin et al., 2009). The off-diagonal entries of the precision matrix $\Omega = \Sigma^{-1}$ are sampled from $\pm(0.1 + 0.2|n|)$, where n is drawn from $\mathcal{N} \sim (0, 1)$. The diagonal entries are selected via Gershgorin's circle theorem to ensure that the matrix is positive definite. Given $\Omega = \Sigma^{-1}$, data can be easily sampled from a multivariate Gaussian distribution $\mathcal{N} \sim (\mathbf{0}, \Sigma)$.

We first generate random connected graphs of $d = 25$ nodes with an average of 2 neighbours/node. For a given graph, we draw 500 samples from the corresponding GMRF distribution and then compare the ability of different methods to retrieve the graph structure when a different sample size is available. Performance metrics for this setting, obtained from 100 repetitions, are reported in **Figure** 2A, while the average runtime for each method is given in **Table** 1. We can see that the EF-λ and EF-cuts approaches have similar accuracy as the ET, as the corresponding Hamming distances to the ground truth (i.e. number of misclassified edges) are on the same level. Notably, the number of false positive edges predicted by the EF-λ method is zero in most cases. Thus, precision is always very close to one. As a trade-off, recall is limited, especially for lower sample sizes. When 500 samples are available, recall reaches levels comparable to the baseline methods. The EF-cuts method performs very similar to the ET, while exhibiting a much higher runtime. The reported runtimes for EF-λ and glasso correspond to a complete run with 32 λ-values. The runtime for glasso is not dependent on the sample size and is mostly consumed for choosing the optimal λ. On the other hand, the runtime for EF-λ increases with sample size. However, we argue that the added runtime constitutes a reasonable trade-off for achieving superior structure learning performance.

Table 1: Average runtime (in seconds) for the experiments presented in **Figure** 2. For EF-λ and glasso the reported runtime corresponds to a complete run with 32 λ-values and choice of the optimal λ.

Sample Size:	**25**	**50**	**100**	**250**	**500**
ET	6	9	13	28	57
EF-λ	32	39	56	110	188
EF-cuts	1166	2088	3512	8151	14435
glasso	31	31	31	31	31

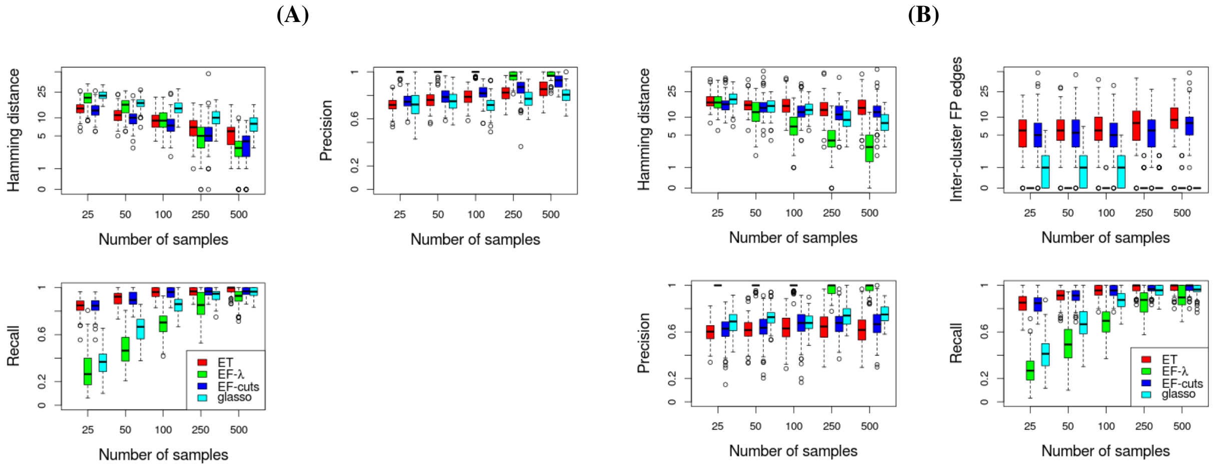

Figure 2: Comparison of the EF-λ, EF-cuts, ET and glasso algorithms on recovering the structure of **(A)** connected **(B)** disconnected sparse GMRFs from different sample sizes. Simulated graphs comprise 25 nodes with 2 neighbours/node on average. The boxplots contain results from 100 repetitions.

In a next step, we evaluated the performance of the EF model in a situation where the data is drawn from a Gaussian MRF with multiple connected components. Therefore, we generated data from GMRFs with no restriction on the number of connected components. Again, each graph comprises $d = 25$ nodes with an average of 2 neighbours/node. Performance metrics for this setting, obtained from 100 repetitions, are reported in **Figure** 2B. We can observe that the EF-λ approach outperforms the other three in terms of accuracy, as it achieves the lowest Hamming distance. As in the one-component setting, the number of false positive edges predicted by this method is zero in most cases. Thus, there are no inter-cluster false positive edges (i.e. edges that are falsely predicted to connect nodes belonging to different clusters) and precision is always very close to one. The recall achieved is inferior to the other methods. However, as the sample size grows, recall also reaches competitive levels. Again in this setting, the EF-cuts approach performs very similar to the original ET method.

We have seen that the EF-cuts method performs very similar to the original ET method, but exhibits much higher runtimes. On the other hand, the EF-λ heuristic performs very well for both connected and disconnected MRFs and is additionally faster and more generic than the the EF-cuts. Thus, we only include EF-λ in the next simulations and refer to it as simply EF.

6.2 RESULTS ON NON-GAUSSIAN MRF DATA

Here we explore the ability to learn the structure of MRFs with non-Gaussian potentials. The EF, as well as the ET approach, are applicable for arbitrary potentials and are, therefore, expected to well adapt to this situation.

We now perform simulations for a Markov network whose data dependencies are no longer Gaussian. More specifically, we generate random graphs consisting of 25 nodes that are organized in small cliques of size 3 or 4. For each clique we draw data samples of pseudo-observations (Kojadinovic and Yan, 2010) from a Student's t-copula with 1 degree of freedom. The dependencies among random variables in each clique are clearly non-Gaussian. Subsequently, we apply the Gaussian quantile function to the pseudo-observations of each random variable and, thereby, we obtain data that is marginally normally distributed. In this setting, we compare the EF approach to the ET, glasso and, additionally, to the non-paranormal model (npn) of Liu et al. (2009). The latter utilizes Gaussian copulas for structure learning. Its implementation is also available via the R-package `huge`.

The results of 100 simulations are summarized in the boxplots of **Figure** 3.The Hamming distances produced by the EF approach are considerably smaller than those produced by competing approaches. Moreover, no false positive edges are predicted by the EF method. Precision and also recall are very high. In contrast, the glasso and non-paranormal methods, that assume Gaussian dependency structures, achieve limited recall. The ET method produces higher Hamming distances and also low precision, since it introduces false positive edges that connect the cliques. Note that the Hamming distance for this method is almost equal to the number of inter-cluster false positive edges. In such a setting, the EF approach performs significantly better than all alternative methods since it naturally deals with t-copula dependencies and disconnected MRF topologies.

6.3 A HIGH-DIMENSIONAL SETTING WITH VERY LOW SAMPLE SIZE

Here, we explore structure learning on the basis of an extremely low number of samples from a comparably high dimensional MRF. This situation commonly arises in many

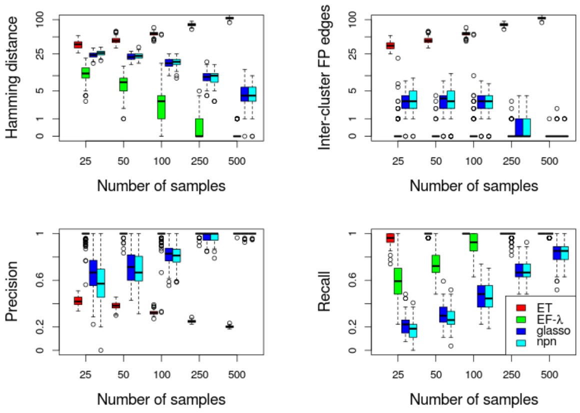

Figure 3: Comparison of the EF, ET, glasso and non-paranormal algorithms on recovering the structure of sparse MRFs with Student's t-copula (df = 1) potentials. Simulated graphs comprise 25 nodes organized in small cliques of size 3 or 4. The boxplots contain results from 100 repetitions.

real world applications, as for instance in biology where typically only few observations are available. In this situation, we do not expect to comprehensively recover the underlying MRF structure. Instead, we aim to maximize the number of recovered true MRF edges at high precision, i.e. without accumulating false positive relationships. Therefore, we generate 50 data samples from an 80-dimensional GMRF, where each node has on average 3 neighbours. The ROC curves in **Figure** 4 compare the performance of the EF and glasso approaches. We can see that, for very low sample sizes, the EF method recovers almost a double number of edges at a tolerance level of 1% FDR. In **Table** 2 we present the average runtime for EF and glasso when run with a single λ value.

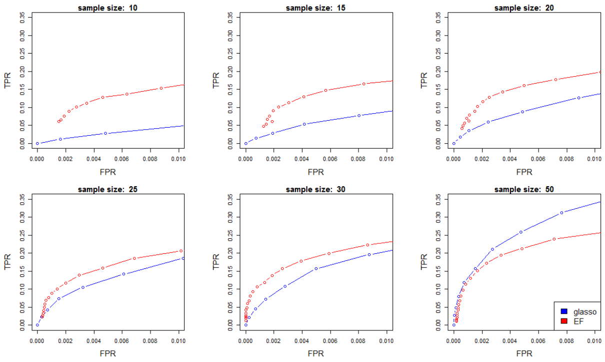

Figure 4: Comparison of the EF and glasso algorithms in a high-dimensional setting (80-node graph) with very low sample size. ROC curves for different numbers of available data replicates are presented, averaged over 100 repetitions. The curves are truncated at a tolerance level of 1% FDR.

Table 2: Average runtime (in seconds) for the simulations presented in **Figure** 4. Runtime is averaged over repetitions and λ values.

Sample Size:	**10**	**15**	**20**	**25**	**30**	**50**
EF-λ	107	153	163	182	185	248
glasso	< 1	< 1	< 1	< 1	< 1	< 1

7 RESULTS ON MICROARRAY DATA

Here we demonstrate the performance of the EF approach on a microarray dataset (Wille et al., 2004) from the isoprenoid biosynthesis pathways in *Arabidopsis thaliana*. Expression levels of 39 genes (variables) are quantified under $n = 118$ conditions (observations). EF is evaluated via comparison to glasso (Friedman et al., 2008), the state-of-the-art algorithm for learning the structure of continuous MRFs. For the EF analysis, we used the Gaussian, Gumbel, Clayton, Frank and Student's t copula as candidate parametric families. A summary of the copula selection results is presented in **Table** 4, where we can observe that a variety of different dependency types is present.

For both methods, a decreasing sequence of 40 λ-values was used. The optimal regularization parameter λ for EF was obtained via eBIC (Foygel and Drton, 2010), resulting in a sparse MRF whose graph structure is depicted in **Figure** 5A. On the contrary, the use of information criteria (eBIC, StARS (Liu et al., 2010)) for glasso yielded very dense networks, as depicted in **Figure** 5B. In order to additionally compare both approaches with respect to results at similar sparsity levels, we also selected the glasso graph with the smallest Hamming distance with respect to the graph learned via EF. To evaluate the performance of the algorithms, we used a 5-fold cross validation setting and evaluated the best-fitting model on the basis of the average per-sample held-out log-likelihood. Results are shown in **Table** 3 and demonstrate that the MRF learned via EF has better cross validation performance. Besides the performance advantage, we note that the sparse structure of EF model selection enables straightforward interpretation and further hypothesis generation by domain experts.

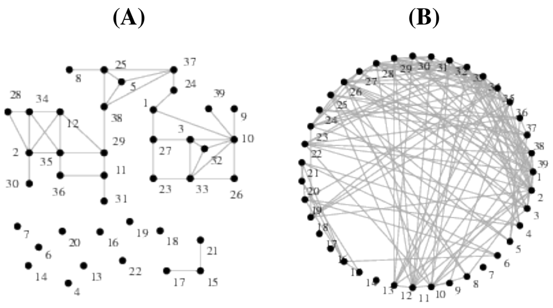

Figure 5: Optimal MRF graph structure recovered via **(A)** EF, **(B)** glasso for the microarray data. The numbering scheme legend is provided as Supplementary Material.

Table 3: Average per-sample held-out log-likelihood for the microarray data.

	Log-likelihood	**Std. error**
EF-λ	9.694	0.526
glasso (StARS)	8.522	0.418
glasso (sparse)	8.995	0.455

8 RESULTS ON IMMUNE CELL PERTURBATION DATA

Finally we apply the EF model to study the occurrence of MRFs with multiple connected components in a proteomics setting. Specifically, we analyze mass cytometry data from human peripheral blood mononuclear cells (PBMC), essentially representing all immune cells residing in the blood stream (Bodenmiller et al., 2012). Mass cytometry allows for proteomic profiling of molecular signaling events at single-cell resolution. The considered publicly available dataset recapitulates the response of PBMC populations to various molecular stimuli under several different pharmacological interventions. Signaling response has been measured by quantifying 14 phosphorylation sites (variables). For each intervention and cell type, 96 conditions were considered, where a condition consisted of an intervention strength setting and a specific stimulus.

Here we present results for interventions with the drug dasatinib. Again we observe the occurrence of a variety of non-Gaussian dependencies in this real world dataset (**Table** 4). We evaluate the performance of EF by comparing it to glasso, as we did for the microarray data. The average held-out log-likelihood per dataset is reported in the boxplots of **Figure** 6A. Different PBMC datasets are grouped together according to the stimulus used in each experiment. We can see that EF achieves constantly superior performance. Furthermore, in **Figure** 6B, separate histograms of the number of connected components for each stimulus are presented. For specific stimuli, MRF topologies with multiple components are common, reflecting the molecular impact of the intervention on the respective cellular signaling event. The EF approach is able to adapt to and recover underlying disconnected topologies even in the presence of unusual dependencies and, thus, we expect this approach to enable the probabilistic characterization of cellular signaling events and, thus, to enable molecular insights of possibly pathologically altered responses and to generate hypotheses for clinical interventions.

9 DISCUSSION

We have introduced the ensemble-of-forests model to approximate structure learning for MRFs with arbitrary potentials and connected components. Additionally, we have presented two approximate inference techniques for this model and compared their structure learning performance with state-of-the-art methods on a comprehensive set of synthetic data.

Table 4: Frequencies of selected copula families during the analysis of plant microarray and PBMC mass cytometry data.

	Gumbel	*Frank*	*Clayton*	*Gaussian*	*t* (df=1)
Micro.	0.28	0.06	0.13	0.51	0.02
PBMC	0.20	0.06	0.35	0.23	0.16

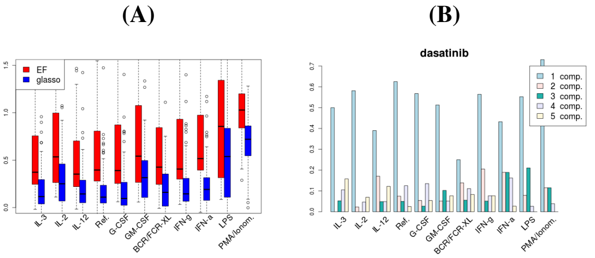

Figure 6: **(A)** Comparison of the EF and glasso algorithms. Boxplots of average held-out log-likelihood for different cell-type / stimulus combinations. **(B)** Histograms of the number of MRF connected components predicted by EF when applied to PBMC mass cytometry data. Separate histograms are given for each stimulus, indicated on the x-axis. Frequencies on the y-axis are normalized to sum up to 1 for each stimulus.

ET and EF models are appealing structure learning approaches when unusual MRF potentials are to be expected. Indeed, our simulation results confirm that the EF method can accurately reconstruct non-Gaussian dependencies that are a priori accounted for.

Disconnected dependency structures frequently arise in real world applications. However, the ET model is conceptually not able to handle such cases. We have extended the ET to the EF model to the end of accommodating multiple-component situations. Our simulation results confirm that we are able to faithfully recover MRF topologies with one as well as with multiple connected components. The study of the plant microarray and PBMC mass cytometry data furthermore confirms the ubiquitous occurrence of the multiple-component situation in cell biology and further emphasizes the need for structure learning approaches that are able to deal with this situation.

We also assessed how the EF model performs for limited sample size, again a typical case for real world applications. Our approach seems ideal for low-sample situations, where we aim to maximize the number of recovered true MRF edges at high precision.

In summary, we expect the EF model to enable MRF structure learning for many real world applications since this approach naturally deals with low sample size, unusual dependency types and disconnected dependency topologies.

References

Banerjee, O., Ghaoui, L. E., d'Aspremont, A., and Natsoulis, G. (2006). Convex Optimization Techniques for Fitting Sparse Gaussian Graphical Models. In *Proceedings of the 23rd International Conference on Machine Learning*, pages 89–96. ACM.

Berkes, P., Wood, F., and Pillow, J. W. (2008). Characterizing neural dependencies with copula models. In *Advances in Neural Information Processing Systems*, pages 129–136.

Bodenmiller, B., Zunder, E. R., Finck, R., Chen, T. J., Savig, E. S., Bruggner, R. V., Simonds, E. F., Bendall, S. C., Sachs, K., Krutzik, P. O., and Nolan, G. P. (2012). Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nature biotechnology*, 30(9):858–867.

Elidan, G. (2013). Copulas in machine learning. In *Copulae in Mathematical and Quantitative Finance*, Lecture Notes in Statistics, pages 39–60. Springer Berlin Heidelberg.

Embrechts, P., Lindskog, F., and McNeil, A. (2003). Modelling dependence with copulas and applications to risk management. *Handbook of heavy tailed distributions in finance*, 8(1):329–384.

Foygel, R. and Drton, M. (2010). Extended Bayesian Information Criteria for Gaussian Graphical Models. In *Advances in Neural Information Processing Systems 23*, pages 604–612.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Johnstone, R. W., Frew, A. J., and Smyth, M. J. (2008). The TRAIL apoptotic pathway in cancer onset, progression and therapy. *Nature Reviews Cancer*, 8(10):782–798.

Kirshner, S. (2008). Learning with Tree-Averaged Densities and Distributions. In *Advances in Neural Information Processing Systems*, pages 761–768.

Kojadinovic, I. and Yan, J. (2010). Modeling multivariate distributions with continuous margins using the copula R package. *Journal of Statistical Software*, 34(9):1–20.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press.

Lin, Y., Zhu, S., Lee, D. D., and Taskar, B. (2009). Learning Sparse Markov Network Structure via Ensemble-of-Trees Models. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 360–367.

Liu, H., Lafferty, J., and Wasserman, L. (2009). The Nonparanormal: Semiparametric Estimation of High Dimensional Undirected Graphs. *The Journal of Machine Learning Research*, 10:2295–2328.

Liu, H., Roeder, K., and Wasserman, L. (2010). Stability Approach to Regularization Selection (StARS) for High Dimensional Graphical Models. In *Advances in Neural Information Processing Systems 23*, pages 1432–1440.

Meilă, M. and Jaakkola, T. (2006). Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 16(1):77–92.

Nelsen, R. B. (1999). *An introduction to copulas*. Springer.

Trivedi, P. K. and Zimmer, D. M. (2005). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1(1):1–111.

Varadhan, R. and Gilbert, P. (2009). Bb: An R Package for Solving a Large System of Nonlinear Equations and for Optimizing a High-Dimensional Nonlinear Objective Function. *Journal of Statistical Software*, 32(4):1–26.

Vazirani, V. and Yannakakis, M. (1992). Suboptimal cuts: Their enumeration, weight and number. In *Automata, Languages and Programming*, pages 366–377. Springer.

Wille, A., Zimmermann, P., Vranova, E., Furholz, A., Laule, O., Bleuler, S., Hennig, L., Prelic, A., von Rohr, P., Thiele, L., Zitzler, E., Gruissem, W., and Buhlmann, P. (2004). Sparse graphical Gaussian modeling of the isoprenoid gene network in Arabidopsis thaliana. *Genome Biol.*, 5(11):R92.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge Package for High-dimensional Undirected Graph Estimation in R. *The Journal of Machine Learning Research*, 98888:1059–1062.