SN Computer Science (2020) 1:64
https://doi.org/10.1007/s42979-020-0064-2

REVIEW ARTICLE

A Statistically Efficient and Scalable Method for Exploratory Analysis of High-Dimensional Data

Mohammad S. Rahman[1] · **Gholamreza Haffari**[1]

Received: 29 May 2019 / Accepted: 9 January 2020 / Published online: 7 February 2020

Abstract
Discovering associations among variables is an important data mining task. The associations can be considered as statistical dependencies among random variables, expressed as the structure of an underlying probabilistic graphical model. Current methods for graphical model structure discovery either do not scale well to datasets with large sample sizes, or suffer from high false discovery rates when the number of dimensions is much larger than the sample size. In this paper, we propose a scalable and statistically efficient approach for graphical model structure discovery for multivariate data involving continuous variables. Our approach uses a minimum message length (MML)-based objective, for which we design a greedy algorithm where the best edges maximising improvements to the MML-based score are added incrementally to the graphical model. We present extensive empirical results on synthetic data with different sample, variable, clique and inverse correlation coefficient and show that our method outperforms strong baselines in terms of both speed and the accuracy of the predicted associations among the random variables in the graphical model. We also report that our method performs significantly very well in AML, BRCA cancer data and other real-life datasets.

Keywords Associations · Minimum message length · Gaussian graphical models

Introduction

Many real-life problems involve continuous valued random variables, where it is critical to uncover the associations among the variables from the sample data. Depending on the problem, the size of the sampled data may be much smaller or bigger compared to the number of dimensions, i.e. the number of variables describing each sampled data point. For example, one may be interested in recovering the gene interaction network over large number of genes based on only a handful of collected gene expression samples [33], or recovering the associations of stock performance in a set of companies based on large number of samples [34]. These problem scenarios make association discovery challenging due to different reasons. In the former, the method needs to be statistically efficient to accurately discover associations among many variables when the size of the sampled data is not large. In the latter, the association discovery method needs to be computationally efficient (i.e. scalable) to handle large amount of sampled data.

Association discovery among data variables can be casted as discovering statistical dependencies among the random variables, expressed by the structure of an underlying probabilistic graphical model [35]. However, current methods for graphical model structure discovery with continuous variables either do not scale well to datasets with large sample sizes [27]; or poor objective function [27]; or suffer from high false discovery rates when the number of dimensions is much larger than the sample size [3]; or sacrificing too much the computational cost [3, 27].

In this paper, we propose a scalable and statistically efficient approach for undirected graphical model structure discovery for exploratory analysis of high-dimensional continuous data. Starting from the null graph, our approach incrementally adds the best edge maximising a test statistic using the graphical model. We start from the log-likelihood ratio test and note that it leads to small number of edges in the estimated graph, hence, missing a large number of true associations. We then present a novel test statistic based on the minimum message length

✉ Mohammad S. Rahman
mohammad.rahman@monash.edu

Gholamreza Haffari
gholamreza.haffari@monash.edu

[1] Clayton School of information Technology, Monash University, Clayton, VIC 3800, Australia

(MML) principle for statistical inference, where candidate models compete based on the length of their lossless compression of the data. An integral part of our test statistics is the maximum likelihood estimate for the parameters of the competing models. As we desire our method to be computationally efficient, we restrict the structure of the competing models to *chordal* graphs. They characterise decomposable probabilistic graphical models, which enjoy analytical solution for the maximum likelihood estimates of their parameters. As such, chordal graphs have been popular in probabilistic graphical models, e.g. see [6, 12, 21].

We call our method ContChordalysis, naming it after Chordalysis [35] which is a method for chordal graphical model structure discovery for discrete-valued random variables. We present extensive empirical results on synthetic and real-life datasets, and show that our method outperforms strong baselines in terms of both speed and the accuracy of the predicted associations between the random variables in the graphical model.

Related Work

It is well know that the dependency structure among the variables corresponds to non-zero entries of the inverse covariance matrix in the corresponding Gaussian graphical model (GGM) [23]. Buhlmann and Geer [9] related the discovery of the structure of inverse covariance matrix in GGMs to the coefficients of a collection of regression problems, consisting of the prediction of each (target) random variable from the rest. Many methods have built on this relationship using different approaches to regression, e.g. to induce sparsity using Lasso. Meinshausen and Buhlmann [31] is one of the first methods that used Lasso to discover the graphical models. It can be viewed as a pseudo-likelihood approximation of the full likelihood. Moreover, their method recovered the presence or absence of an edge in penalized Gaussian graphical models. However, their method did not use stability selection to estimate the covariance matrix and estimated covariance matrix is not positive definite. Yuan and Lin [44] improved the GGM discovery by assuming that the observations are suitably centred and scaled, and the diagonal elements of the sample precision matrix are set to one to maintain the positive definiteness, which is more natural for estimating the precision matrix. At the same, [4] used block coordinate descent algorithm to estimate the covariance matrix instead of precision matrix. Both methods used penalized likelihood as the objective function to estimate the optimal GGM by considering the parameters for sparsity of all variables are the same. However, the level of sparsity of each variable should be different [16]. Hence, the objective function suffers from the inaccurate estimation of covariance matrix. [16] and [10] improved the objective function by introducing different regularization parameters for the variables and estimated covariance matrix using coordinate descent algorithm and linear programming, respectively. They call their method Graphical Lasso or GLasso and CLIME, respectively. But, both GLasso and CLIME suffer from the non-scalability issue.

However, [39] proposed a new neighbourhood selection approach using de-sparsified Lasso to resolve the positive definiteness problem. Moreover, [29] and [27] estimated the precision matrix in a column-by-column fashion using scaled Lasso and SQRT Lasso to make the methods faster. Aforementioned methods suffer from the low discovery rate (i.e. recall) because the penalized likelihood objective functions use a single constant regularization parameter to control sparsity for all variables. Eigenvalues of the covariance matrix spread more when the dimension of the data is larger than the sample size [26], which may affect the stability of GLasso estimation. To improve the stability, [3] propose using a k-root of the sample covariance matrix, with $k \geq 1$, to attain less spread eigenvalues and, therefore, obtain a more accurate estimation of covariance matrix without sacrificing too much the computational cost.

However, all of the Lasso-based methods use the L_1-norm penalty, which is proportional to the first moment of the weighted degree[1] distribution. In a small world network, it is important to capture at least second moment of the weighted degree of nodes [19].

In parallel to the Lasso based methods, a significant research has been made using the greedy-based approach. [12] proposed a fast forward selection greedy approach to estimate the covariance matrix using decomposable models and maximum likelihood estimates (MLE). However, they did not present any empirical results nor they presented a method for handling continuous variables with GGMs. [20] and [28] discovered the structure of a GGM using a forward–backward algorithm and an L_1 penalized log-likelihood as objective function. In the paper, we present a forward selection greedy-based method with two new objective functions which resolves the major drawbacks of the existing Lasso and greedy-based methods mentioned above to discover the structure more accurately and efficiently.

Structure Discovery in Decomposable Gaussian Graphical Models

Let $\mathcal{D} = \{X_1, \ldots, X_n\}$ be a training set consisting of n data points where $X_i \in \mathbb{R}^d$ and d is the number of dimensions (equivalently attributes, or random variables). Our aim is

[1] Weighted degree is the sum of the weights of all edges attached to a node. Here weight is the amount of the association between two nodes.

to discover the unobserved undirected graphical structure based on the observed/sampled vectors in $\mathcal{D}$.

We are interested in the undirected graphical structure $G = (V, E)$, where V is the set of vertices each of which corresponds to a random variable (or a dimension of the input vectors), and E is the set of edges capturing the statistical associations between random variables. A parameterisation of the model corresponds to multivariate functions assigned to subset of variables in maximal cliques of the graph. The probability density function corresponding to the graph is defined as

$$f(X) \propto \prod_{C \in \mathcal{C}} f_C(X_C),$$

where $\mathcal{C}$ is the set of maximal cliques, and $f_C(X_C)$ is a clique-specific function defined on the subset of variables appearing in a clique C. In any distribution resulting from the graphical model, two random variables are statistically independent conditioned on the variables in a cut separating the two.

In this paper, we assume that the observed input vectors have been generated from a multivariate Gaussian distribution $X \sim \mathcal{N}_d(\mu, \Sigma)$, which means the cliques are also parameterised by Gaussian distributions. Therefore, our aim is to discover the structure of the so-called Gaussian graphical model. For computational convenience, we work with chordal graphical structures, leading to decomposable models covered in the next subsection.

Decomposable Models

Decomposable models is a subclass of undirected graphical models which provides a usefully constrained representation in which model selection and parameter estimation can be done efficiently, which makes it suitable for large-scale problems. According to [25], a graphical model G is decomposable if it is equivalent to

- G is chordal. A chordal graph is one in which all cycles of four or more vertices have a chord, which is an edge that is not part of the cycle but connects two vertices of the cycle.
- All maximal prime subgraphs of G are cliques.
- G admits a perfect numbering.
- Every minimal (α,β)-separator are complete.

Let $\mathcal{M}$ be a decomposable model, and $f_{\mathcal{M}}$ be the probability density function of a Gaussian distribution corresponding to $\mathcal{M}$. It can be shown that [24]

$$f_{\mathcal{M}}(X) = \frac{\prod_{C \in \mathcal{C}} f(X_C)}{\prod_{S \in \mathcal{S}} f(X_S)}, \tag{1}$$

where $\mathcal{C}$ is the set of maximal cliques and $\mathcal{S}$ is the set of minimal separators corresponding to the chordal graph of the model $\mathcal{M}$. The importance of this result is that it relates the Gaussian distribution over all variables to those on the subsets of variables, i.e. Gaussian distributions over the variables involved in maximal cliques $f(X_c)$ or minimal separators $f(X_s)$. This amounts to a closed form solution for the maximum likelihood estimate (MLE) of the covariance matrix $\hat{\Sigma}$ of the Gaussian graphical model $f_{\mathcal{M}}$, through the MLE of the covariance matrices of the component models.

Theorem 1 [24] *For a Gaussian graphical model corresponding to a chordal graph $G = (V, E)$, the maximum likelihood estimate of the covariance matrix is*

$$\hat{\Sigma}^{-1} = n\left\{ \sum_{C \in \mathcal{C}} \left[(ssd_C)^{-1} \right]^V - \sum_{S \in \mathcal{S}} \left[(ssd_S)^{-1} \right]^V \right\}, \tag{2}$$

where $[A]^V$ denotes extending a small matrix A defined on a subset of variables V to a larger matrix on all variables by setting extra entries to zero. ssd (Sum of Squared Distance) is defined as

$$ssd = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)^T,$$

where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the empirical mean. The determinant of the estimate can be calculated as [24]

$$\det\ \hat{\Sigma}^{-1} = n^{|V|} \frac{\prod_{S \in \mathcal{S}} \det\ (ssd_S)^{\nu(S)}}{\prod_{C \in \mathcal{C}} \det\ ssd_C}, \tag{3}$$

where $\nu(S)$ is multiplicities of separators. If separator S contains more than one vertex, $\nu(S) = 1$; otherwise, $\nu(S) = 2$.

The inverse of the covariance matrix is called the *precision* matrix $K = \Sigma^{-1}$. Interestingly, the non-association between the variables expressed in the graph $G = (V, E)$ of the Gaussian graphical model translates to pattern of zeros in the precision matrix, i.e. K has zero for all entries where there is no edge between the corresponding pairs of vertices in E.

To discover the optimal decomposable graphical structure from a given training data, typically one of the following strategies is employed [12, 35]:

- Forward selection: starting with the simplest model with no edge (i.e. $E = \emptyset$). Edges are added incrementally, as long as the new hypothesised models are not rejected according to an appropriate test statistics.
- Backward elimination: starting with the complete graph over the $|V|$ vertices, edges are deleted incrementally, as long as the new hypothesised models are not rejected according to an appropriate test statistics.

In this paper, we adopt the forward selection strategy, and add the edges incrementally. As we want the resulting model to be decomposable, the addition of an edge has to be done with care.

Structure Discovery by Hypothesis Testing

Let $\mathcal{S}^+(G)$ denotes all positive definite matrices whose zero patterns are consistent with the graph G, i.e. they have zero for all entries corresponding to non-existent edges of the graph G. Let G' be a candidate graph resulting from adding an edge (a, b) to the graph G, that is $G' = G \cup \{(a, b)\}$. In the forward selection strategy, we test the hypothesis that $K \in \mathcal{S}^+(G)$ under the assumption that $K' \in \mathcal{S}^+(G')$.

Theorem 2 [24] *The exact deviance test for testing a decomposable model $K \in \mathcal{S}^+(G)$ assuming a decomposable model $K' \in \mathcal{S}^+(G')$ can be performed by rejecting the small values of*

$$r = \frac{\det \hat{K}}{\det \hat{K}'} \tag{4}$$

which is distributed as a beta distribution $\mathcal{B}(\frac{|V|-|C_{ab}|}{2}, \frac{1}{2})$, where C_{ab} is the maximal clique that contains the newly added edge (a, b). $\hat{K}$ and $\hat{K}'$ are the maximum likelihood estimates for the precision matrix of the Gaussians corresponding to G and G', respectively.

We will discuss how the test statistics can be computed efficiently in "Efficient Computation of the Test Statistics". As the graphical model is learnt incrementally by adding one edge at a time, we make intensive use of statistical testing. Multiple hypothesis testing is prone to many false discoveries. This is critical in our approach where we need to do a lot statistical testing due to the large size of the search space, which may lead to accepting modifications of the models more often than needed. This can be avoided using layered critical values [42], a variant of the Bonferroni correction that increases the number of significant patterns discovered while still maintaining strict control over the risk of false discoveries. Given the p value threshold α (usually $\alpha = 0.1$), the layered p value at iteration t of the algorithm is

$$\alpha_t = \frac{\alpha}{2^t |\mathcal{G}_t|}, \tag{5}$$

where t is the number of edges in the current best model, and $\mathcal{G}_t$ is the number of chordal graphs that can be formed by adding an edge to the current model.

The resulting structure discovery algorithm is presented in Algorithm 1. We call our algorithm ContChordalysis to highlight that it is for structure discovery of chordal graphs for continuous valued variables.

Algorithm 1 ContChordalysis

Input: Dataset $D = \{X_i\}_{i=1}^n$, Significance level α
Output: The graph $G = (V, E)$
Initialise G to be the graph without any edges
$t \leftarrow 1$
repeat
  $E^c \leftarrow$ CandidateEdge(G) ▷ see Section 3.3
  for $e \in E^c$ **do**
    $r_e =$ testStatistic(D, G, e) ▷ based on eqn (6)
  end for
  $e^* \leftarrow \arg\min_{e \in E^c} r_e$
  $pval = \frac{\alpha}{2^t |E^c|}$
  if $r_{e^*} \leq pval$ **then**
    $G \leftarrow$ addEdge(G, e^*) ▷ see Section 3.3
  end if
  $t \leftarrow t + 1$
until $(r_{e^*} > pval)$ **or** $(t > \frac{|V|(|V|-1)}{2})$

Efficient Computation of the Test Statistics

We now turn to the question of how to efficiently compute the test statistics in Eq. (4), which is particularly important for large-scale datasets.

Deshpande et al. [12] characterise the edges that can be added to a decomposable model while retaining its decomposability. Furthermore, it presents an efficient algorithm to enumerate all such edges in $O(|V|^2)$. This is achieved by a data structure called the *clique graph*, which keeps track of the maximal cliques $\mathcal{C}$ and minimal separators $\mathcal{S}$. Adding an edge to the graph and updating the underlying data structures also takes $O(|V|^2)$.

Theorem 3 [12] *If two decomposable models $\mathcal{M} \subset \mathcal{M}'$ differ only in one edge (a, b) (i.e. $(a, b) \in \mathcal{M}'$ and $(a, b) \notin \mathcal{M}$), then the maximal cliques and the minimal separators (C, S) and (C', S') in these two models differ as follows*:

- if $C_a \not\subset C_{ab}$ and $C_b \not\subset C_{ab}$, then $C' = C + C_{ab}$ and $S' = S + C_{ab} \cap C_a + C_{ab} \cap C_b - S_{ab}$,
- if $C_a \subset C_{ab}$ and $C_b \not\subset C_{ab}$, then $C' = C + C_{ab} - C_a$ and $S' = S + C_{ab} \cap C_b - S_{ab}$,
- if $C_a \not\subset C_{ab}$ and $C_b \subset C_{ab}$, then $C' = C + C_{ab} - C_b$ and $S' = S + C_{ab} \cap C_a - S_{ab}$,
- if $C_a \subset C_{ab}$ and $C_b \subset C_{ab}$, then $C' = C + C_{ab} - C_a - C_b$ and $S' = S - S_{ab}$,

where C_{ab} and S_{ab} are the maximal clique and minimal separators for the nodes a and b, and C_a and C_b are the maximal cliques including each of these nodes.

Thus, the change in the determinant of the MLE estimates of the precision matrix after adding an edge (a, b) is only dependent on the minimal separator of the two vertices S_{ab}, the newly formed clique C_{ab}, and the newly formed separators $C_{ab} \cap C_a$ and $C_{ab} \cap C_b$. This means we only have to compute the determinant terms relevant to the candidate edges that can be added to the current model. This immediately leads to the following theorem.

Theorem 4 *If two decomposable models $\mathcal{M} \subset \mathcal{M}'$ differ only in one edge (a, b) (i.e. $(a,b) \in \mathcal{M}'$ and $(a,b) \notin \mathcal{M}$), then*

$$\frac{\det\ \hat{K}}{\det\ \hat{K}'} = \frac{\det\ ssd_{C_{ab}} \cdot \det\ ssd_{S_{ab}}}{\det\ ssd_{C_{ab}\cap C_a} \cdot \det\ ssd_{C_{ab}\cap C_b}}. \tag{6}$$

Proof From Theorem 2, we need to compute the test statistics $r = \frac{\det\ \hat{K}}{\det\ \hat{K}'}$. From Eq. (3) and Theorem 3, the test statistics is calculated as

$$\frac{\det\ \hat{K}}{\det\ \hat{K}'} = \frac{n^{|V|}\frac{\prod_{S\in\mathcal{S}} \det\ ssd_S}{\prod_{C\in\mathcal{C}} \det\ ssd_C}}{n^{|V|}\frac{\prod_{S\in\mathcal{S}'} \det\ ssd_S}{\prod_{C\in\mathcal{C}'} \det\ ssd_C}} = \frac{\frac{\prod_{S\in\mathcal{S}} \det\ ssd_S}{\prod_{C\in\mathcal{C}} \det\ ssd_C}}{\frac{\prod_{S\in\mathcal{S}+C_{ab}\cap C_a+C_{ab}\cap C_b-S_{ab}} \det\ ssd_S}{\prod_{C\in\mathcal{C}+C_{ab}} \det\ ssd_C}}$$

which immediately gives the test statistics in Eq. 6. We have considered the first case in Theorem 3 above; moreover, considering the other three cases results in the same expression for the test statistics. □

Time Complexity Analysis

We now turn to the time complexity analysis of ContChordalysis in Algorithm 1. At the step t of the main loop of the algorithm, all candidate next graphs have t edges. The computation of the test statistics in line 8 is upper bounded by the evaluation of the maximal clique that can be formed by adding the tth edge to the graph. In the extreme case, where all t edges form one clique, the maximum size of the clique would be[2] $k = \frac{1+\sqrt{1+8t}}{2}$. Hence, the time complexity of computing the test statistics is $O(k^3) = O(t^{\frac{3}{2}})$, assuming computing the determinant of a k-by-k matrix is $O(k^3)$. The time complexity of enumerating the candidate edges in line 6 and adding a selected edge in line 13 is $O(|V|^2)$. Therefore, the time complexity of one pass over the main loop is $O(|V|^2 + |E^c|t^{\frac{3}{2}}) = O(|V|^2 t^{\frac{3}{2}})$ since the number of candidate edges E^c is upper bounded by the number of edges of the complete graph $\frac{|V|(|V|-1)}{2}$.

[2] A clique of k nodes contains $\frac{k(k-1)}{2}$ edges. Setting the number of edges to t and solve the resulting quadratic equation yields $k = \frac{1+\sqrt{1+8t}}{2}$, [35].

Information Theory-Based Objective Function for Structure Discovery

ContChordalysis is based on the maximum likelihood estimation, and uses multiple test correction to reduce the false discovery rate. However, the low rate of false discoveries comes at a price: it requires many samples to accept correct hypotheses. Moreover, ContChordalysis has a major functional drawback: it relies on the existence of the maximum likelihood estimates, which may not exist if the number of samples is less than the size of the largest clique in the graph. To overcome these drawbacks, we propose a new test statistic based on the minimum message length (MML).

The MML criterion provides an information-theoretic objective for statistical inference to find the best hypothesis for the observed data [40]. It controls the false discovery rate, requiring far fewer samples to accept true hypotheses. MML relies on quantifying the amount of information required to convey losslessly the observed data in an explanation message. The best hypothesis is the one that can convey the entire data set in the shortest possible explanation message.

Let us consider a hypothesis (or model) $\mathcal{M}$ that offers an explanation of the observed data $\mathcal{D}$. Based on the fundamental rules of probability:

$$p(\mathcal{M}, \mathcal{D}) = p(\mathcal{M}) \times p(\mathcal{D}|\mathcal{M}) = p(\mathcal{D}) \times p(\mathcal{M}|\mathcal{D}),$$

where $p(\mathcal{M})$ is the prior over hypotheses/models, $p(\mathcal{D}|\mathcal{M})$ is the likelihood, $p(\mathcal{D})$ is the prior probability of data, and $p(\mathcal{M}|\mathcal{D})$ is the posterior of $\mathcal{M}$ given $\mathcal{D}$. Using Shannon's communication theory, the amount of information for explaining $\mathcal{D}$ with $\mathcal{M}$ is

$$I(\mathcal{M}, \mathcal{D}) = I(\mathcal{M}) + I(\mathcal{D}|\mathcal{M}) = I(\mathcal{D}) + I(\mathcal{M}|\mathcal{D}), \tag{7}$$

where $I(a) = -\log(p(a))$ gives the optimal code length to convey an event a whose probability is $p(a)$. This results in an objective criterion to compare two competing models $\mathcal{M}_1$ and $\mathcal{M}_2$ given the same data D:

$$I(\mathcal{M}_1|\mathcal{D}) - I(\mathcal{M}_2|\mathcal{D}) = I(\mathcal{M}_1) + I(\mathcal{D}|\mathcal{M}_1) - I(\mathcal{M}_2) - I(\mathcal{D}|\mathcal{M}_2). \tag{8}$$

For our structure discovery setting, the encoding of the model in the message consists of the encoding of the chordal graph's topology G and the associated model parameters, which we elaborate in the rest of this section.

Encoding of the Graph

We now describe the encoding of the graphical structure G associated with the model $\mathcal{M}$ based on [1] and [36]. For this purpose, it is sufficient to send the edges of the graph: (a) the number of edges $|E|$, and (b) the particular combination of

the edges that the graph exhibits if we have an enumeration of all possibilities. We do not need to encode the variables since it is common across all models, hence does not change the outcome of comparing messages.

We need $\log(|E_{Complete}|+1)$ to encode the number of edges,[3] where $|E_{Complete}| = \frac{|V|\times|V-1|}{2}$. For a given number of the edges, we ideally need to index and send only the chordal graphs. However, we are not aware of an analytical expression for the number of chordal graph with a fixed number of edges. Hence, we use the number of all graphs as an upper-bound, which results in sending more bits than necessary. The number of all possible graphs with a fixed number of edges $|E|$ is $\log \binom{|E_{Complete}|}{|E|}$. Hence, the length of encoding the graph's topology is

$$I(G) = \log(|E_{Complete}|+1) + \log \binom{|E_{Complete}|}{|E|}. \tag{9}$$

Encoding of the Parameters and the Data

Once the graph's topology has been encoded, we encode the parameters of the model as well as the data. To encode model parameters, we encode the parameters of all maximal cliques and minimal separators and then combine them. Let $k << |V|$ be the number of nodes in a maximal clique C (or alternatively, a minimal separator). Let $\mathcal{D}^C$ be the part of data set $\mathcal{D}$ corresponding to the variables in C. According to [40], the MML encoding of $\mathcal{D}^C$ and the parameters of the multivariate Gaussian distribution corresponding to a maximal clique (or minimal separator) C is

$$I(C,\mathcal{D}^C) \overset{\text{def}}{=} \underbrace{\frac{m_C}{2}\log(q) + \frac{m_C}{2} - \overbrace{\log(h(\theta_C))}^{\text{Prior}} + \overbrace{\log\sqrt{|\mathcal{F}(\theta_C)|}}^{\text{Fisher information}}}_{I(C)} \underbrace{- \overbrace{\mathcal{L}(\mathcal{D}^C|\theta_C)}^{\text{log-likelihood}}}_{I(\mathcal{D}^C|C)}, \tag{10}$$

where $m_C = \frac{k^2+3k}{2}$ is the number of free parameters, q is the lattice quantisation constant to reduce the quantisation error,[4] and $\theta_C = (\mu_C, \Sigma_C)$. In what follows, we compute various components of $I(C,\mathcal{D}^C)$ in Eq. (10), i.e. the prior probability, the Fisher information matrix, and the likelihood.

Prior Probability of the Parameters

Following [13], we use a flat prior for μ_C and a conjugate inverted Wishart prior for Σ_C. Hence, the prior joint density over the parameters is $h(\theta_C) \propto |\Sigma_C|^{-\frac{k+1}{2}}$, where k is the size of the clique C.

[3] This includes zero for the null graph.

[4] Quantisation error results from limited precision in machines when representing real numbers.

Likelihood

The log-likelihood $\mathcal{L}(\mathcal{D}^C|\mu_C, \Sigma_C)$ of the relevant part of the data based on the multivariate Gaussian distribution corresponding the maximal clique C is

$$-\frac{nk}{2}\log 2\pi - \frac{n}{2}\log|\Sigma_C| - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu_C)\Sigma_C{}^{-1}(X_i-\mu_C)^T, \tag{11}$$

where k is the size of the clique $|C|$. As seen in "Structure Discovery in Decomposable Gaussian Graphical Models", the MLE estimates are given by

$$\hat{\mu}_C = \frac{1}{n}\sum_{i=1}^{n}X_i^C \;\;,\;\; \hat{\Sigma}_C = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^C-\hat{\mu}_C)(X_i^C-\hat{\mu}_C)^T. \tag{12}$$

Fisher Information of the Parameters

We need to evaluate the second-order partial derivatives of $-\mathcal{L}(D^C|\mu_C, \Sigma_C)$ for computing the Fisher information for the parameters [40]. Let $|\mathcal{F}(\mu_C, \Sigma_C)|$ represent the determinant of the Fisher information matrix which is the product of $|\mathcal{F}(\mu_C)|$ and $|\mathcal{F}(\Sigma_C)|$, i.e. the determinant of the Fisher information matrices of μ_C and Σ_C, respectively.

Taking the second-order partial derivatives of $-\mathcal{L}(\mathcal{D}_C|\mu_C, \Sigma_C)$ with respect to μ_C, we get $-\nabla^2_{\mu_C}\mathcal{L} = n\Sigma_C^{-1}$. So the determinant of the Fisher information matrix for μ_C is $|\mathcal{F}(\mu_C)| = n^k|\Sigma_C|^{-1}$.

To compute $|\mathcal{F}(\Sigma_C)|$, Magnus and Neudecker [30] derived an analytical expression using the theory of matrix derivatives based on matrix vectorization:

$$|\mathcal{F}(\Sigma_C)| = n^{\frac{k(k+1)}{2}}2^{-k}|\Sigma_C|^{-(k+2)}. \tag{13}$$

Hence, the determinant of the Fisher information matrix for μ_C and Σ_C is

$$|\mathcal{F}(\mu_C, \Sigma_C)| = n^{\frac{k(k+3)}{2}}2^{-k}|\Sigma_C|^{-(k+2)}. \tag{14}$$

Putting it All Together

Substituting the prior probability, Fisher information and log-likelihood into Eq. (10), the encoding of the parameters and data of a maximal-clique/minimal-separator is

$$I(C,\mathcal{D}^C) = \frac{n-1}{2}\log(|\Sigma_C|) + \frac{1}{2}\sum_{i=1}^{n}(x_i-\mu_C)\Sigma_C^{-1}(x_i-\mu_C)^T + c, \tag{15}$$

where c is a constant.

After encoding the data and parameters corresponding to all maximal cliques and minimal separators, we need to combine them to get the length of the message needed to be sent from the sender to the receiver. Let us start by an example graphical model consisting of $C_{\{A;B\}}$ and $C_{\{B;C\}}$ as maximal cliques where S_B is their minimal separator. To send the parameters of this simple graphical model, one may think of sending the parameters of the multivariate Gaussian distributions corresponding to the maximal cliques and the separator $\{P(A,B), P(B,C), P(B)\}$, which include their means and covariance matrices. However, this encoding has redundancy as the parameters of the separator $P(B)$ can be reconstructed by the receiver from either $P(A, B)$ or $P(B, C)$ via marginalisation. Therefore, a more efficient encoding consists of sending the parameters of $\{P(A, B), P(B, C)\}$. This idea can be pushed further to send the parameters of $P(C|B)$ instead of $P(B, C)$, as the joint can be constructed from the conditional as well as the marginal $P(B)$ which is already computable from $P(A, B)$.

Hence, a more efficient encoding may be that for sending $\{P(A,B), P(C|B)\}$ or $\{P(B,C), P(A|B)\}$.

In general, there are exponentially many non-redundant sets of conditional and joint factors, from which the original joint distribution can be constructed. To find the minimum message length, we would need to have a search over this exponential space of sets. To avoid such search, we resort to a measure which sums up the encoding needed for the parameters of the maximal cliques, and deducts the encoding of the minimal separators to remove redundancy.

We resort to the following efficiently computable expression to approximate the message length consisting of the model and the data:

$$I(\mathcal{M}|\mathcal{D}) = I(G) + \sum_{C\in\mathcal{C}} I(C, \mathcal{D}^C) - \sum_{S\in\mathcal{S}} I(S, \mathcal{D}^S), \tag{16}$$

where $\mathcal{C}$ and $\mathcal{S}$ are the set of maximal cliques and minimal separators, respectively. A similar expression has been used in [36] to encode model parameters and data for discrete-valued graphical models.

MML as Test Statistics

As mentioned earlier, we use forward selection to discover the graphical model. In forward selection, the reference model $\mathcal{M}$ and a candidate model $\mathcal{M}'$ are differed by and edge (a, b). According to MML theory, $\mathcal{M}'$ replaces $\mathcal{M}$ if encoding the message based on $\mathcal{M}'$ requires fewer bits than that of $\mathcal{M}$, i.e. $I(\mathcal{M}|\mathcal{D}, G) - I(\mathcal{M}'|\mathcal{D}, G') > 0$. Therefore, the MML score for comparing the reference and a candidate model is

$$\begin{aligned} &I(\mathcal{M}|\mathcal{D}, G) - I(\mathcal{M}'|\mathcal{D}, G') \\ &= \log\left(\frac{|E_{Complete}| - |E|}{|E_{Complete}| - |E| - 1}\right) + I(C_{ab}, \mathcal{D}^{C_{ab}}) + I(S_{ab}, \mathcal{D}^{S_{ab}}) \\ &\quad - I(C_{ab} \cap C_b, \mathcal{D}^{C_{ab}\cap C_b}) - I\big(C_{ab} \cap C_a, \mathcal{D}^{C_{ab}\cap C_a}\big). \end{aligned} \tag{17}$$

The resulting method, which we call ContChordalysis-MML, is summarised in Algorithm 2; it differs from Algorithm 1 only in lines from 8 to 11.

The time complexity of computing $I(C, \mathcal{D}^C)$ is $O(|C|^3)$ since it contains the inverse and the determinant of $|C|$-by-$|C|$ matrices. Therefore, the time complexity of Algorithm 2 is similar to that of Algorithm 1, where one pass over the main loop is $O(|V|^2 + |E^c|t^{\frac{3}{2}}) = O(|V|^2 t^{\frac{3}{2}})$ in the round t of the main loop.

Algorithm 2 ContChordalysis-MML

1: **Input:** Dataset $D = \{X_i\}_{i=1}^n$
2: **Output:** Graph $G = (V, E)$
3: $t \leftarrow 0$
4: Initialise G to be the graph without any edges
5: **repeat**
6: $E^c \leftarrow$ CandidateEdge(G) ▷ see Section 3.3
7: **for** $e \in E^c$ **do**
8: $s_e =$ MMLScore(D, G, e) ▷ based on eqn (17)
9: **end for**
10: $e^* \leftarrow \arg\max_{e\in E^c} s_e$
11: **if** $s_{e^*} > 0$ **then**
12: $G \leftarrow$ addEdge$(G, e*)$ ▷ see Section 3.3
13: **end if**
14: $t \leftarrow t + 1$
15: **until** $(s_{e^*} < 0)$ **or** $(t > \frac{|V|(|V|-1)}{2})$

Experiments and Results

We compare the performance of our methods (ContChordalysis and ContChordalysis-MML) with five baselines on both synthetic and real-life datasets.

Baselines

We compare our methods with five strong competing methods: TIGER [27], CLIME [10], Graphical Lasso (GLasso) [16], rooted Graphical Lasso (r-GLasso) [3] and a recently proposed greedy approach called FoBa-gdt [28].

All of the baselines use penalized log-likelihood as the objective function. Moreover, all of the baselines discover the Gaussian graphical model structures. For these two reasons, we compare our methods with the above-mentioned baselines to evaluate the performance.

Other Scoring Functions

In this paper, we have proposed two test statistics based on MML and p value to select an optimal solution. We compare the performance of the proposed test statistics to two scoring functions: [2]'s proposed MDL (Minimum Descriptor Length) score and BIC (Bayesian Information Criterion) [15].

$$Altmueller's\ score = -\mathcal{L}(\mathcal{D}|\theta) + \log n + \sum_{i=1}^{|\mathcal{C}|}(|E_i|\log n) + \sum_{i=1}^{|\mathcal{C}|}|V_i| - \sum_{i=1}^{|\mathcal{S}|}|V_i| \tag{18}$$

$$BIC = -2\mathcal{L}(\mathcal{D}|\theta) + |E|\log n, \tag{19}$$

where $\mathcal{L}(\mathcal{D}|\theta) = \sum_{c=1}^{|\mathcal{C}|}\mathcal{L}(\mathcal{D}^c|\theta^c) - \sum_{s=1}^{|\mathcal{S}|}\mathcal{L}(\mathcal{D}^s|\theta^s)$.

We employ BIC and Altmueller's MDL scores to form additional variants of ContChordalysis, and call them *ConChordalysis-BIC* and *ContChordalysis-Altmueller*, respectively.

Performance Metrics

We evaluate results using standard performance metrics: precision, recall, and FMeasure. Precision is the fraction of correctly predicted edges (i.e. associations) with respect to all predicted edges. Recall is the fraction of correctly predicted edges with respect to the correct edges. FMeasure is the harmonic mean of precision and recall, i.e. $\text{FMeasure} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}}$.

In our synthetic data experiments, the graphs used for generating synthetic data are considered the gold standard. Both ContChordalysis and ContChordalysis-MML use decomposable model to discover the GGM.

Encoding Graph Structures

In "Encoding of the Graph", we use the number of all graphs of $|V|$ as upper bound to encode the graph structure instead of the number of all chordal graphs. In this section, we empirically compare the number of bit requires to encode graphs with respect to the number of all graphs and of all chordal graphs. [22] and [43] estimated the number of all chordal graphs of $|V|$ which we use to compute the number of chordal graphs of $|V|$ in the experiments is as below:

$$|E_{chordal}| = \sum_{i=0}^{|V|}\binom{|V|}{i}2^{i(|V|-i)}. \tag{20}$$

In Eq. (17), we compute the MML difference between reference and candidate models to select the appropriate model. In this equation, we find that bit difference to encode the reference and candidate models are very small. Hence, in the experiment, we compare the bit difference between reference and candidate models with respect to the number of all graphs and chordal graphs. To do this, we modify the graph encoding part of Eq. (17) for the number of chordal graphs as below:

$$\log\left(\frac{|E_{chordal}| - |E|}{|E_{chordal}| - |E| - 1}\right). \tag{21}$$

Table 1 Empirical comparison between the number of all graphs and chordal graphs to use to encode the graph structures

$\|E\|$	$\|V\| = 100$		$\|V\| = 1000$	
	$\|E_{Complete}\|$	$\|E_{chordal}\|$	$\|E_{Complete}\|$	$\|E_{chordal}\|$
1	0.00010	0.00000	0.00000010	0.00000
10	0.00010	0.00000	0.00000010	0.00000
100	0.00010	0.00000	0.00000010	0.00000
1000	0.00011	0.00000	0.00000010	0.00000
10000	0.00011	0.00000	0.00000011	0.00000
100000	–	–	0.00000011	0.00000

Table 1 compares the bit difference of encoding the reference and candidate models using the number of all graphs and the number of chordal graphs. The bit difference between the MML scores by assuming the number of all nodes and chordal graphs are not significant which is even less than 0.001. Therefore, we can say that there is no significant difference between the number of bits to encode a graph structure using the number of all graphs instead of using the number of chordal graphs.

Synthetic Data

Clauset et al. [11] mentioned that real-world networks maintain the following properties:

- Many small nodes are connected with few hubs. It is known as power law property.
- Short path exists between two nodes.
- New nodes prefer to attach to well-connected nodes over less well-connected nodes which is known as preferential attachment property.
- Every new vertex is born with a link or several links.

Beeri et al. [5] proposed an model to generate scale-free graphs having the above-mentioned properties. We use Barabási–Albert (BA) model to generate the graph structures with the properties of real-world networks. This model facilities us by controlling the number of nodes and the degree which controls the edge density of the graph. As we use decomposable models to discover the graphical

Table 2 Average recall (*Re*), precision (*Pr*), FMeasure (*FM*) and computational times (in seconds) of ContChordalysis-MML and its variants, TIGER, CLIME, GLasso, r-GLasso and FoBa-gdt for small world networks by varying the number of samples n while $V = 100$, $|C| = 3$, and $S = 50$

Methods	$n = 100$				$n = 1000$				$n = 50,000$			
	Re	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)
ContChordalysis-MML	0.51	0.62	0.56	231.00	0.59	0.72	0.65	246.00	0.63	0.79	0.7	1284.00
ContChordalysis	0.34	0.59	0.43	195.00	0.34	0.61	0.44	211.32	0.36	0.68	0.47	1068.01
ContChordalysis-Altmueller	0.24	0.6	0.34	140.40	0.28	0.63	0.39	165.12	0.31	0.69	0.43	844.20
ContChordalysis-BIC	0.16	0.54	0.25	106.68	0.17	0.55	0.26	109.68	0.24	0.68	0.35	615
TIGER	0.48	0.62	0.54	451.26	0.53	0.65	0.58	498.18	0.59	0.70	0.64	2607.05
CLIME	0.49	0.59	0.54	593.4	0.48	0.63	0.54	653.71	0.54	0.69	0.61	3396.00
GLasso	0.44	0.52	0.48	441.36	0.50	0.53	0.51	464.64	0.54	0.59	0.56	2534.4
r-GLasso	0.47	0.56	0.51	229.68	0.53	0.58	0.55	529.88	0.56	0.61	0.58	1327.2
FoBa-gdt	0.49	0.61	0.54	378.42	0.52	0.62	0.57	436.81	0.59	0.66	0.62	2235.45

structure, we add an additional condition that the generated graph would be chordal. We use candidate edge selection process of ContChordalysis algorithm to maintain the chordality of generated graph.

Therefore, we generate synthetic data with the following four parameters:

- V: the number of variables. It varies in 10, 100 and 1000.
- N: the number of samples. It varies in 100, 1000, 10,000, 50,000.
- $|C|$: maximum clique size. According to [5], every new node born with some edge connections with existing nodes. It produces the connected graph. Therefore, the minimal clique size would be 2. But we also investigate the performance of the model for disconnected graph. For disconnected graph, the maximum clique size would be 1. Therefore, it varies in between 1 and 6.
- S: inverse correlation coefficient between the nodes (where $\rho = \frac{1}{S}$). It varies in $\{1,\ 5,\ 10,\ 50,\ 100,\ 500\}$.

Having the graph structure, we generate the covariance matrix $\Sigma \sim (1/S) \cdot \mathcal{U}(0,5) \cdot \mathcal{U}(0,5) \cdot adj(i,j)$. Here adj be the adjacency matrix. Finally, we generate data using $\mathcal{D} = \mathcal{D} \sim \mathcal{N}(0, \Sigma)$. Moreover, to assess the performance of our models, we vary each of the above-mentioned parameters in turn, having set the others to $V = 100$, $N = 1000$, $|C| = 3$, and $S = 50$. For each configuration, we generate 50 datasets from the corresponding multivariate Gaussian distribution, and report the average results over these 50 artificially generated datasets.

In synthetic data experiments, we make use of fivefold cross validation, thereby dividing the dataset into five partitions. We take any of these five partitions as the test set and use the other four partitions as the training sets to learn the regularization parameters of the competitive methods.

Varying the Number of Samples (n)

Table 2 compares our methods with the baselines on synthetic data. ContChordalysis-MML outperforms the other methods in all configurations in terms of precision, recall, and FMeasure. As expected, for a particular dimension size, the FMeasure of all methods improves as the sample size n increases. Likewise, in almost all cases, the FMeasure decreases for a particular sample size as the number of dimensions increases. In other words, ContChordalysis-MML requires fewer samples compared to other methods to discover meaningful associations between the variables.

As our method ContChordalysis-MML detects more true edges than other methods to maintain chordality, both recall and precision of ContChordalysis-MML are higher than others. Interestingly, ContChordalysis-MML outperforms ContChordalysis based on the empirical results in Table 2. In ContChordalysis, the threshold α_t is geometrically decreased as new edges are added to the graph. Hence, after a few steps, the threshold becomes very small, stopping the addition of new edges. Therefore, the number of edges in graphs discovered by ContChordalysis is small, which leads to missing a large number of true edges. This is confirmed by inspecting the number of edges in the graphs discovered by different methods.

MML extensively uses the covariance matrix to predict the association between the random variables, whereas MDL [2] and BIC [15] use the number of edges instead of the covariance matrix to resolve the overfitting problem of MLE and ignore the effect of covariance matrix. Moreover, [17] points out the limitation of BIC on high-dimensional data and performs well for data having small number of variables. This consequently affects its recall, precision and FMeasure. Altmueller and Haralick [2]'s score is a variant of BIC. Therefore, [2]'s score performs similar to

Table 3 Average recall (*Re*), precision (*Pr*), FMeasure (*FM*) and computational times (in seconds) of ContChordalysis-MML and its variants, TIGER, CLIME, GLasso, r-GLasso and FoBa-gdt for small world networks by varying the number of variables V while $n = 1000, |C| = 3$, and $S = 50$

Methods	$V = 10$				$V = 100$				$V = 1000$			
	Re	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)
ContChordalysis-MML	0.62	0.76	0.68	2.52	0.59	0.72	0.65	246	0.51	0.69	0.59	2238.41
ContChordalysis	0.44	0.63	0.52	2.28	0.34	0.61	0.44	211.32	0.27	0.56	0.36	1890.13
ContChordalysis-Altmueller	0.35	0.64	0.45	1.68	0.28	0.63	0.39	165.12	0.21	0.57	0.31	1424.76
ContChordalysis-BIC	0.3	0.64	0.41	1.32	0.17	0.55	0.26	109.68	0.09	0.39	0.15	807.29
TIGER	0.54	0.72	0.62	4.14	0.53	0.65	0.58	498.18	0.45	0.62	0.52	4550.39
CLIME	0.50	0.67	0.57	5.41	0.48	0.63	0.54	653.17	0.37	0.57	0.45	5662.17
GLasso	0.56	0.59	0.57	4.08	0.50	0.53	0.51	464.64	0.37	0.47	0.41	4266.72
r-GLasso	0.55	0.63	0.59	4.67	0.53	0.58	0.55	529.88	0.4	0.53	0.46	4842.14
FoBa-gdt	0.58	0.70	0.63	4.20	0.52	0.62	0.57	436.80	0.42	0.56	0.48	3825.46

BIC and is outperformed by both ContChordalysis-MML and ContChordalysis.

TIGER, CLIME, GLasso and r-GLasso use Lasso to estimate the precision matrix. In Lasso, the regularization parameter λ in the penalized likelihood objective functions significantly affects the precision matrix estimation process. However, TIGER, CLIME, GLasso and rGLasso uses EBIC to select the regularized parameter λ. As EBIC is not fully Bayesian and do not encode the graphical structures to penalize likelihood, TIGER, CLIME, GLasso and r-GLasso are outperformed by ContChordalysis-MML.

Similar to other baselines, FoBa-gdt is also outperformed by ContChordalysis-MML. FoBa-gdt uses a penalized likelihood as the objective function and it removes edges at backward elimination step until the objective function finds an optimal solution. Therefore, it removes many true edges which affects the recall, precision and FMeasure.

Table 2 also shows that ContChordalysis runs faster than the baselines, although it suffers from the inaccurate prediction of associations. ContChordalysis-MML runs much faster than the baseline methods: TIGER, CLIME, GLasso, r-GLasso and FoBa-gdt. Therefore, ContChordalysis-MML is a statistically efficient and scalable method for predicting associations.

Varying the Number of Variables (*V*)

We perform three experiments by varying the number of random variables. In these experiments, recall, precision and FMeasure results are reported in Table 3 and ContChordalysis-MML outperforms all other methods. Over the increase of the number of variables with respect to same sample size, ContChordalysis-MML detects many less edges accurately and decreases the recall, precision FMeasure. Similar trends are also found in other methods, but not as good as ContChordalysis-MML. Hence, ContChordalysis-MML can work on any size of multivariate Gaussian distribution data efficiently.

Varying the Size of Cliques (|*C*|)

In this experiment, we vary maximal clique sizes from 2 to 5. Table 4 shows the recall, precision and FMeasure of the outputs of our method and other baselines. ContChordalysis-MML outperforms all of the competitive baselines.

While the maximal clique size is two, the degree of all vertices is one. All methods detect most of the true edges and their FMeasure are higher. Over the increment of the maximum size of cliques in the graph, FMeasure of all methods decreases. In this experiment, maximal size of cliques in the graphs inversely affects the FMeasure. Moreover, for the graphs with large cliques (clique size is more than 3), recall of existing strong baselines is significantly worse than smaller clique graphs, whereas our ContChordalysis-MML detects many true edges than other methods whatever the size of maximal cliques in the graph and, therefore, recall, precision and FMeasure of ContChordalysis-MML is higher than others.

Varying the Inverse Correlation Coefficient (*S*)

Correlation expresses the statistical association between random variables which strongly influences covariance matrices. According to Table 5, increase in the value of the inverse correlation coefficient S inversely impacts the covariance matrices and causes the decrease of FMeasure. Our ContChordalysis-MML can detect more than 44% true edges even when very small correlation exists between variables, whereas other methods cannot detect even 40% of true edges.

Table 4 Average recall (*Re*), precision (*Pr*), *FMeasure* (*FM*) and computational times (in seconds) of ContChordalysis-MML and its variants, TIGER, CLIME, GLasso, r-GLasso and FoBa-gdt for small world networks by varying the size of cliques $|C|$ while $V = 100$, $n = 1000$, and $S = 50$

Methods	$\|C\| = 2$				$\|C\| = 3$				$\|C\| = 4$				$\|C\| = 5$			
	Re	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)
ContChordalysis-MML	0.65	0.76	0.70	263.76	0.59	0.72	0.65	246.95	0.52 0	.69	0.59	214.27	0.46	0.65	0.54	219.24
ContChordalysis	0.37	0.64	0.47	220.56	0.34	0.61	0.44	211.32	0.26	0.55	0.35	182.76	0.21	0.50	0.30	172.20
ContChordalysis-Altmueller	0.30	0.67	0.41	168.6	0.28	0.63	0.39	165.12	0.19	0.60	0.29	123.96	0.15	0.47	0.23	123.84
ContChordalysis-BIC	0.21	0.61	0.31	116.28	0.17	0.55	0.26	109.68	0.10	0.50	0.17	69.36	0.09	0.37	0.14	84.24
TIGER	0.57	0.68	0.62	514.05	0.53	0.65	0.58	498.18	0.43	0.62	0.51	434.01	0.39	0.59	0.47	445.51
CLIME	0.53	0.66	0.59	676.25	0.48	0.63	0.54	653.74	0.36	0.58	0.44	540.62	0.36	0.55	0.44	565.81
GLasso	0.55	0.57	0.56	478.56	0.50	0.53	0.51	464.64	0.37	0.48	0.42	398.88	0.34	0.43	0.38	398.16
r-GLasso	0.56	0.60	0.58	260.64	0.53	0.58	0.55	258.48	0.41	0.55	0.47	221.28	0.34	0.46	0.39	211.08
FoBa-gdt	0.57	0.65	0.61	453.39	0.52	0.62	0.57	436.87	0.43	0.66	0.52	369.63	0.42	0.56	0.48	396.48

On account of discovering high-dimensional Gaussian graphical model, ContChordalysis-MML outperforms all methods in different experimental setups. Therefore, ContChordalysis-MML is a statistically efficient method to predict the statistical dependencies from high-dimensional Gaussian data.

Acute Myeloid Leukemia Gene Expression Data

We also apply our methods to TCGA cancer gene expression datasets: AML (Acute Myeloid Leukemia) to discover the gene network. We download both gene expression datasets from cBioPortal[5]. In the experiments, we focus on the cancer-related transcription factors (TFs). The AML gene expression dataset contains 51 TFs and 173 samples. For the gold standard, we use the *regulatory potential scores*[6] between pairs of genes, i.e. TFs for AML cancer based on TF ChIP-seq binding data from the Cistrome Cancer Database.[7] Following the previous work [41], an edge is considered to exist between two TFs if their regulatory potential score is at least 0.5.

Table 6 presents the number of edges predicted by the baselines and the variants of our method. ContChordalysis-MML recovers not only more edges than other methods, but also more "unique true edges", i.e. those true edges which are not detected by other methods. Table 6 depicts the results, and shows that ContChordalysis-MML outperforms the other methods in terms of FMeasure.

We also compare the run time of the different methods to discover the graphical structures. As shown in Table 6, the speed trend is similar to those observed in the synthesis experiments, where ContChordalysis-BIC is the fastest method, followed by ContChordalysis-MML, which is in turn faster than the baselines.

The gold standard graph of the AML dataset is not chordal. Hence, we add some edges to the gold standard graph to make it chordal, which should give an upper bound on the performance of our methods along with the baselines. Therefore, we use the moralized AML dataset to find the upper bound on the performance of all of the comparing methods and to investigate the performance drop when the graph is not chordal. Columns 7–12 of Table 6 present the recovery of the number of moralized AML gold standard edges by the baselines and our method. Table 6 depicts the upper bound on FMeasure/precision/recall of our methods and the baselines. Original gold standard graph is nearly chordal, and we add only 19 edges to make it chordal. Therefore, there is not a significant difference between the upper bound FMeasure and the original FMeasure. Most important findings of this experiment is that TIGER, CLIME, GLasso, r-GLasso and FoBa-gdt do not perform well on chordal graphs, which is reflected in their FMeasures.

Breast Cancer Gene Expression Data

We also apply our methods to TCGA cancer gene expression datasets: BRCA (breast invasive carcinoma) to infer the gene network. We download BRCA gene expression datasets from cBioPortal. Similar to AML experiments, we focus on the cancer-related transcription factors (TFs). The BRCA gene expression dataset contains 729 TFs and 528 samples. For the gold standard, we use the regulatory potential scores between a pair of genes, i.e. TFs for BRCA cancer based on TF ChIP-seq binding data from the Cistrome Cancer Database. Following the previous work [41], an edge

[5] http://www.cbioportal.org.

[6] Regulatory potential scores are a computational tool to aid in the identification of putative regulatory sites of the human genome.

[7] http://cistrome.org/CistromeCancer.

Table 5 Average recall (*Re*), precision (*Pr*), *FMeasure* (*FM*) and computational times (in seconds) of ContChordalysis-MML and its variants, TIGER, CLIME, GLasso, r-GLasso and FoBa-gdt for small world networks by varying the inverse correlation coefficients S while $V = 100, |C| = 3$, and $n = 1000$

Methods	$S = 1$				$S = 5$				$S = 10$				$S = 50$				$S = 100$				$S = 500$			
	Re	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)	*Re*	*Pr*	*FM*	*T*(s)
ContChordalysis-MML	0.74	0.78	0.76	284.28	0.68	0.76	0.72	242.55	0.66	0.75	0.7	264.72	0.59	0.72	0.65	246.01	0.48	0.67	0.56	216.84	0.44	0.63	0.52	215.16
ContChordalysis	0.43	0.70	0.53	223.56	0.41	0.70	0.52	182.94	0.39	0.63	0.48	235.56	0.34	0.61	0.44	211.32	0.25	0.54	0.34	182.88	0.20	0.53	0.29	158.52
ContChordalysis-Altmueller	0.37	0.72	0.49	190.88	0.36	0.71	0.48	155.26	0.33	0.66	0.44	189.12	0.28	0.63	0.39	165.12	0.19	0.55	0.28	132.84	0.14	0.5	0.22	111.72
ContChordalysis-BIC	0.32	0.70	0.44	147.24	0.29	0.7	0.41	116.08	0.24	0.61	0.34	135.48	0.17	0.55	0.26	109.68	0.10	0.42	0.16	84.84	0.05	0.38	0.09	53.88
TIGER	0.65	0.72	0.68	563.04	0.63	0.71	0.67	509.69	0.60	0.68	0.64	548.09	0.53	0.65	0.58	498.18	0.41	0.63	0.5	423.43	0.36	0.58	0.44	407.33
CLIME	0.59	0.68	0.63	741.35	0.58	0.68	0.63	705.09	0.55	0.66	0.6	716.4	0.48	0.63	0.54	653.71	0.37	0.58	0.45	568.8	0.31	0.58	0.40	505.8
GLasso	0.61	0.61	0.61	542.88	0.59	0.61	0.60	444.21	0.59	0.57	0.58	514.08	0.50	0.53	0.51	464.64	0.37	0.48	0.42	401.04	0.29	0.47	0.36	347.28
r-GLasso	0.66	0.65	0.65	274.56	0.61	0.65	0.63	210.32	0.60	0.61	0.60	282.12	0.53	0.58	0.55	258.48	0.42	0.53	0.47	229.32	0.33	0.53	0.41	200.28
FoBa-gdt	0.67	0.71	0.69	492.87	0.64	0.71	0.67	439.29	0.59	0.64	0.61	477.96	0.52	0.62	0.57	436.82	0.40	0.57	0.47	378.84	0.32	0.56	0.41	332.43

Table 6 The number of matched, predicted and unique prediction of edges by different methods in the AML gene expression data

Method	Number of predicted edges compared with gold standard data												Time
	Non-moralized						Moralized						
	The number of true edges 550						The number of true edges 569						
	Matched	Predicted	Unique	Recall	Precision	F-Measure	Matched	Predicted	Unique	Recall	Precision	F-Measure	
ContChordalysis-MML	237	277	45	0.43	0.86	0.57	254	277	47	0.45	0.92	0.6	385.25
ContChordalysis	59	125	5	0.11	0.47	0.18	71	125	4	0.12	0.57	0.2	182.36
ContChordalysis-Altmueller	26	69	0	0.05	0.38	0.09	37	69	0	0.07	0.54	0.12	86.78
ContChordalysis-BIC	14	50	0	0.03	0.28	0.05	22	50	0	0.04	0.44	0.07	122.58
TIGER	195	430	31	0.35	0.45	0.39	201	430	33	0.35	0.47	0.4	399.67
CLIME	130	255	4	0.24	0.51	0.33	137	255	3	0.24	0.54	0.33	1342.7
GLasso	115	253	0	0.21	0.45	0.29	116	253	0	0.2	0.46	0.28	391.2
r-GLasso	128	254	8	0.23	0.5	0.32	130	254	7	0.23	0.51	0.32	402.6
FoBa-gdt	62	152	2	0.11	0.41	0.17	64	152	2	0.11	0.42	0.17	214.9

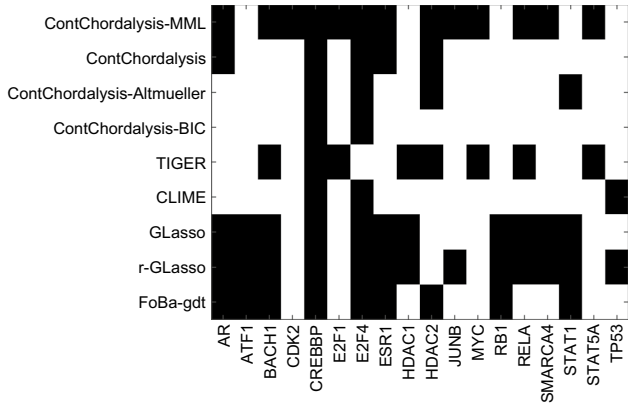

Fig. 1 Discovery of interacting gene with BRCA_1 which causes the appearance the breast cancer (BRCA) [37] by different methods. (Black spots in a row represent the discover of the interactions between the genes and BRCA_1 by a method labeled in *y*-axis.)

is considered to exist between two TFs if their regulatory potential score is at least 0.5.

Table 7 presents recall, precision, FMeasure and the running time of the variants of our method versus the baselines. We again see the same trend that ContChordalysis-MML outperforms other methods in terms of the performance measures. We list the number of edges predicted by ContChordalysis (including its variants) and the baselines in Table 7. We see that ContChordalysis-MML discovers more true edges than any other of compared methods. It also detects more than 3000 true gene pairs which are not detected by the baselines.

BRCA_1 is an important susceptible gene that may cause the appearance of breast and ovarian cancers in human body [8, 14, 32]. Pujana et al. [37] has identified the list of genes which have interacted with BRCA_1 to create cancer cells in breast and ovary. We are interested to assess the methods by comparing their discoveries on genes having interaction with BRCA_1. Fig. 1 shows that ContChordalysis-MML spots 13 true gene-pairs, whereas TIGER/CLIME/GLasso/r-GLasso/FoBa-gdt detect 11/3/7/12/9 true gene pairs, respectively.

It is known that many gene pairs can be responsible for the appearance of cancer cells in human body. We are also interested in knowing which important gene pairs have been detected by the different methods. According to [38], gene pairs with higher RTS (regulatory potential scores) are the important gene pairs for the appearance of breast cancer. We select 50 important gene pairs based on the RTS between the gene pairs. Fig. 2 shows that ContChordalysis-MML detected 37 gene pairs, while TIGER the strongest baseline method and r-GLasso detect just 27 gene pairs. CLIME, GLasso and FoBa-gdt discover fewer than 25 TF-pairs. Based on this evaluation, our method ContChordalysis-MML outperforms existing strong baselines by discovering more true important gene pairs.

We also analyse the computational time of the different methods for discovering associations among the variables. Table 7 shows that the variants of our method are faster than all the baselines.

Similar to the moralized experiment on AML gene expression data, we carry out an experiment to measure the upper bound on the performance of our methods and the baseline by making the gold standard graph chordal. Columns 5–7 of Table 7 present the recovery of the number of moralized BRCA gold standard edges by the baselines and our method. Table 7 depicts the upper bound on FMeasure/Precision/Recall of our methods and baselines on BRCA dataset. Similar to the AML dataset, there is not a significant performance drop when the network is not chordal. Our MML approach predicted more accurate edges than other methods and outperformed the baselines.

Fig. 2 Discovery of 50 prominent gene pairs, causing the breast cancer (BRCA), by different methods

Table 7 The number of matched, predicted and unique prediction of edges by different methods in the BRCA gene expression data

Method	Number of predicted edges compared with gold standard data											Time
	Non-moralized						Moralized					
	The number of true edges 25833						The number of true edges 26971					
	Matched	Predicted	Unique	Recall	Precision	F-Measure	Matched	Predicted	Unique	Recall	Precision	F-Measure	
ContChordalysis-MML	10824	14627	3077	0.42	0.74	0.54	11374	14627	3241	0.42	0.78	0.55	1201.48
ContChordalysis	694	1126	88	0.03	0.62	0.06	807	1126	97	0.03	0.72	0.06	446.78
ContChordalysis-Altmueller	228	409	0	0.01	0.56	0.02	273	409	1	0.01	0.67	0.02	158.93
ContChordalysis-BIC	201	346	0	0.01	0.58	0.02	236	346	0	0.01	0.68	0.02	201.36
TIGER	5475	7873	1584	0.21	0.7	0.32	5603	7873	1891	0.21	0.71	0.32	1553.51
CLIME	464	703	21	0.02	0.66	0.04	481	703	0	0.02	0.68	0.04	1544.11
GLasso	2593	3642	61	0.1	0.71	0.18	2264	3642	57	0.08	0.62	0.14	1497.6
r-GLasso	3129	4788	97	0.12	0.65	0.2	3257	4788	82	0.12	0.68	0.2	1521.6
FoBa-gdt	1780	3011	11	0.07	0.59	0.13	1859	3011	7	0.07	0.62	0.13	519.8

Table 8 The number of edges predicted by ContChordalysis, its MML variant and TIGER including the significant edges found by [7] and [33] from human gene expression data ancestry

Methods	The number of gold standard edges predicted by		
	Total	[7]	[33]
ContChordalysis-MML	618	45	74
ContChordalysis	108	13	19
TIGER [27]	306	40	70

Ancestry Gene Expression Data

We also perform experiment on another real dataset that was used by TIGER, one of the most competitive baselines. This dataset contains unrelated individuals of Northern and Western European ancestry from Utah (CEU), whose genotypes are available from the Sanger Institute website[8] [7]. The number of samples n is 60 and the dimension size d is 100. Bhadra and Mallick [7] have analyzed the data and found 55 significant interactions among the 100 chosen traits. Mohammadi and Wit [33] used a Bayesian method to infer the gene network with 281 edges which include all of the significant interaction discovered by [7]. Moreover, among the 281 edges, [33] identified 86 edges as significant interactions.

Liu and Tiger [27] used this dataset to evaluate the performance of TIGER. We only test ContChordalysis and ContChordalysis-MML on this data. Table 8 presents the number of edges predicted by the TIGER[9] and the variants of our method. From Table 8, we can say that our ContChordalysis-MML discovered more accurate graphical structure than TIGER and outperformed TIGER and ContChordalysis.

Patient Classification Data

We carry out another experiments on the problem of predicting the breast cancer patient with their types. In this experiment, r-GLasso [3] predicts the cancer patients with pCR (pathological complete response) and residual disease (RD) from the graphical model structure. Patients having the same disease type will be connected with each other, otherwise they are not connected. In this experiment, we use the same dataset[10] that [3] used which contains 22,283 gene expression levels of 133 patients. There are 34 patients with pCR

[8] ftp://ftp.sanger.ac.uk/pub/genevar.

[9] This experiment is already carried out by [27] and reported in their paper.

[10] available at http://bioinformatics.mdanderson.org/pubdata.html.

Table 9 Average pCR and RD classification measurements

Method	Specificity	Sensitivity	*MCC*
ContChordalysis-MML	0.91	0.86	0.74
ContChordalysis	0.90	0.41	0.36
GLasso [16]	0.75	0.61	0.33
r-GLasso [3]	0.69	0.84	0.48
CLIME [10]	0.71	0.84	0.49

and 99 patients with RD. Similar to [3], we consider the results of [18] as the gold standard.

To measure the prediction accuracy, [3] used specificity, sensitivity and Matthews correlation coefficient[11] (*MCC*). Moreover, they consider *TP* and *TN* as the number of correctly predicted pCR and RD patients, respectively, and *FP* and *FN* as the number of erroneously predicted pCR and RD patients, respectively. Therefore, we use specificity, sensitivity and *MCC* instead of recall, precision and FMeasure to evaluate the performance. Avagyan et al. [3] also compared their method with GLasso and CLIME on this dataset.

Table 9 presents the pCR and RD patients classification results. In the table, we have reported the results of GLasso, r-GLasso and CLIME from [3]'s paper. Based on the results of Table 9, our method ContChordalysis-MML outperformed CLIME, GLasso and r-GLasso.

Real-Life Data Experiment on Finance Stock Performance of the Companies

We carry out further experiments on another real dataset: "Finance stock performance of the companies" used in [34], which contains 20 years financial performance of 490 companies. The number of samples in the dataset is 3450, where the financial footprints in individual days are considered as samples. Using this dataset, we identify the financial relationship between the companies. As we do not have any gold standard data for this dataset, we compute the log-likelihood of the held-out data to evaluate the performance of the methods.

We make use of fivefold cross validation, thereby dividing the dataset into five partitions of size 690 samples. We take any of these five partitions as the test set and use the other four partitions as the training sets to learn the structure of the graphical model.

Table 10 shows the average log-likelihood of the models recovered by ContChordalysis-MML is higher than that for the other methods. Furthermore, the average log-likelihood per edge for ContChordalysis-MML is higher than the other methods. Overall, these results indicate that ContChordalysis-MML is more accurate in predicting the association between variables compared to the baselines.

Table 10 Average log-likelihood of different methods on the dataset of finance stock performance of the companies

Method	Log-likelihood	Log-likelihood per edge
ContChordalysis-MML	**− 895.29**	**− 0.19**
ContChordalysis	− 989.72	− 1.44
ContChordalysis−Altmueller	− 922.53	− 1.82
ContChordalysis-BIC	− 995.61	− 2.09
TIGER	− 908.76	− 0.24
CLIME	− 969.63	− 1.25
GLasso	− 993.75	− 0.39
rGLasso	− 1011.17	−0.34
FoBa−gdt	− 973.6	− 1.17

Conclusion

We have proposed a scalable and statistically efficient approach for graphical model structure discovery involving continuous variables for exploratory data analysis. We introduce ContChordalysis and it variants, including a novel MML-based criterion, for structure discovery of Gaussian graphical models. Our methods are stepwise algorithms, where they add edges maximising a test statistics incrementally to the estimated graph. ContChordalysis makes use of log-likelihood ratio test, and ContChordalysis-MML uses an information theoretic criterion based on minimum message length principle. Our methods work with chordal graphs and decomposable models to make the computation of the test statistics efficient. We have presented extensive empirical results on synthetic and real-life datasets, and show that our ContChordalysis-MML method outperforms strong baselines in terms of both speed and the accuracy of the predicted associations from the data. More specifically, from the empirical results, it appears that ContChordalysis-MML is well suited for high-dimensional continuous data. Moreover, our MML-based objective function detects more true edges than other methods. However, due to predicting more edges, the precision and false discovery rate are also affected.

There are different avenues for future work. In ContChordalyis and its variants, we initially focus on 'true edge-detection'. So that they discover more false edges to maintain chordality. Our future direction is to extend our work by introducing the concept of backward elimination step after the forward selection to remove the undesired edges from

[11] The Matthews correlation coefficient (*MCC*) is used in machine learning as a measure of the quality of binary (two-class) classifications. $MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$.

the graphical model. It would certainly improve the precision and reduce the risk of over-fitting issues. Furthermore, we designed ContChordalysis and its variants based on the concept of chordality to compute the objective function efficiently. But in the real world, the true graph may not be necessarily chordal. Therefore, another possible improvement of our method is to develop the method without assuming the concept of chordality. We intend to further speed up our method by saving the computation for edges whose scores are not affected by the last added edge. Although multivariate Gaussian distributions are good approximations for many real-world phenomena, we believe that there are real-life data which may be better captured by other forms of distributions. Therefore, we are interested in extending our framework to capture a broader class of distributions governing the data.

Acknowledgements We are grateful to Prof. Ann E. Nicholson and Dr. Francois Petitjean for insightful discussion and comments.

Funding: This student/ research is not funded by any organizations.

Compliances with Ethical Standard

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Allisons L. Encoding General Graphs, (2017). http://www.allisons.org/ll/MML/Structured/Graph/, Accessed 16 May 2017
2. Altmueller S, Haralick RM. Approximating high dimensional probability distributions. Proceedings of 17th International Conference on Pattern Recognition , 2004 (ICPR'04) 2004;2:299–302
3. Avagyan V, et al. Improving the graphical lasso estimation for the precision matrix through roots of the sample covariance matrix. J Comput Graph Statist **Online Publication** (2017)
4. Banerjee O, et al. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. J Mach Learn Res. 2007;9:485–516.
5. Barabási AL, Albert R. Statistical mechanics of complex networks. Rev Mod Phys. 2002;74(1):47–97.
6. Beeri C, et al. On the desirability of acyclic database schemes. J ACM 1983;379–513
7. Bhadra A, Mallick B. Joint high-dimensional bayesian variable and covariance selection with an application to eqtl analysis. Biometrics. 2013;69(2):447–57.
8. Brose M, et al. Cancer risk estimates for BRCA1 mutation carriers identified in a risk evaluation program. J Natl Cancer Inst. 2002;94(18):1365–72.
9. Buhlmann P, van de Geer S. Statistics for high-dimensional data, methods, theory and applications. Berlin: Springer; 2011.
10. Cai T, et al. A constrained l_1 minimization approach to sparse precision matrix estimation. J Am Stat Ass. 2011;106:594–607.
11. Clauset A, et al. Power-law distributions in empirical data. SIAM Rev. 2007;51:661–703.
12. Deshpande A, et al. Efficient stepwise selection in decomposable models. Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence 2001;128–135
13. Dowe D, et al. MML estimation of the parameters of the spherical Fisher distribution. Algorithm Learn Theor. 1996;1160:213–27.
14. Finch A, et al. Salpingo-oophorectomy and the risk of ovarian, fallopian tube, and peritoneal cancers in women with a BRCA1 or BRCA2 mutation. J Am Med Assoc. 2006;296(2):185–92.
15. Foygel R, Drton M. Extended Bayesian Information Criteria for Gaussian Graphical Models. Proceedings of 24th Annual Conference on Neural Information Processing Systems. 2010;23:604–12.
16. Friedman J, et al. Sparse inverse covariance estimation with the graphical lasso. Biostatistics. 2008;9:432–41.
17. Giraud C. Introduction to high-dimensional statistics. London: Chapman and Hall/CRCs; 2014.
18. Hess L, et al. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. J Clin Oncol. 2006;24:4236–44.
19. Jackson M. Social and economic networks. Princeton: Princeton University Press; 2008.
20. Johnson C, et al. High-dimensional Sparse Inverse Covariance Estimation using Greedy Methods. Proceedings of the 15th International COnference on Artificial Intelligence and Statistics 2012
21. Kangas L, et al. Learning chordal markov networks by dynamic programming. Adv Neural Inf Process Syst. 2014;27:2357–65.
22. Kijima S, Kiyomi M, Okamoto Y, Uno T. On listing, sampling, and counting the chordal graphs with edge constraints. Theor Comput Sci. 2010;411(26):2591–601.
23. Koller D, Friedman N. Probabilistic graphical models: principles and techniques - adaptive computation and machine learning. Cambridge: The MIT Press; 2009.
24. Lauritzen S. Graphical models. Oxford: Oxford statistical science series; 1996.
25. Lauritzen S. Decomposition and decomposable graphs. CIMPA Summerschool, Hammamet 2011 2011
26. Ledoit O, Wolf M. A well-conditioned estimator for large-dimensional covariance matrices. J Multivar Anal. 2004;88(2):365–411.
27. Liu H. TIGER: a tuning-insensitive approach for optimally estimating Gaussian graphical models. Electr J Stat. 2017;11:241–94.
28. Liu J, et al. Forward-backward greedy algorithm for general convex smooth functions over a cardinality constraint. Proceedings of the 31st International Conference on International Conference on Machine Learning 2014;32:503–511
29. Liu W, Luo X. Fast and adaptive sparse precision matrix estimation in high dimensions. J Multivar Anal. 2015;135:153–62.
30. Magnus J, Neudecker H. Matrix differential calculus with applications in statistics and econometrics. New York: Willey; 1988.
31. Meinshausen N, Buhlmann P. Stability selection. J R Stat Soc. 2006;72:417–73.
32. Miki Y, et al. A strong candidate for the breast and overian cancer susceptinility gene `BRCA\_1`. Science. 1994;266:66–71.
33. Mohammadi A, Wit EC. Bayesian structure learning in sparse Gaussian graphical models. Bayesian Anal. 2015;10:109–38.
34. Petitjean F, Webb G. Scaling log-linear analysis to datasets with thousands of variables 2015;469–477
35. Petitjean F, et al. Scaling log-linear analysis to high-dimensional data. Proceedings of IEEE International Conference on Data Mining 2013;597–606
36. Petitjean F, et al. A Statistically Efficient and Scalable Method for Log-Linear Analysis of High-Dimensional Data. Proceedings of IEEE International Conference on Data Mining (ICDM) 2014;110–119
37. Pujana MA, et al. Network modeling links breast cancer susceptibility and centrosome dysfunction. Nature genetics. 2007;39:1338–49.

38. Qin Q, et al. ChiLin: a comprehensive ChIP-seq and DNase-seq quality control and analysis pipeline. BMC Bioinf. 2016;17:1274–86.
39. Waldrop L. Testing for graph differences using the desparsified lasso in high-dimensional data. Statistics Survey 2014
40. Wallace C, Boulton D. An information measure for classification. Comput J. 1968;11:185–94.
41. Wang C, et al. Solving log-determinant optimization problems by a newton-cg primal proximal point algorithm. SIAM J Optimiz. 2013;20:2994–3013.
42. Webb G. Layered critical values: a powerful direct-adjustment approach to discovering significant patterns. J Mach Learn. 2008;71:307–23.
43. Wormald N. Counting labeled chordal graphs. Graphs Comb. 1985;1:193–200.
44. Yuan M, Lin Y. Model selection and estimation in the gaussian graphical model. Biometrika. 2007;94(1):19–35.