A Mixture Copula Bayesian Network Model for Multimodal Genomic Data

Qingyang Zhang[1,3] and Xuan Shi[2]

[1]Department of Mathematical Sciences, [2]Department of Geosciences

University of Arkansas, Fayetteville, AR 72701

[3]To whom the correspondence should be addressed. Email: qz008@uark.edu

Abstract

Gaussian Bayesian networks have become a widely used framework to estimate directed associations between joint Gaussian variables, where the network structure encodes decomposition of multivariate normal density into local terms. However, the resulting estimates can be inaccurate when normality assumption is moderately or severely violated, making it unsuitable to deal with recent genomic data such as the Cancer Genome Atlas data. In the present paper, we propose a mixture copula Bayesian network model which provides great flexibility in modeling non-Gaussian and multimodal data for causal inference. The parameters in mixture copula functions can be efficiently estimated by a routine Expectation-Maximization algorithm. A heuristic search algorithm based on Bayesian information criterion is developed to estimate the network structure, and prediction can be further improved by the best-scoring network out of multiple predictions from random initial values. Our method outperforms Gaussian Bayesian networks and regular copula Bayesian networks in terms of modeling flexibility and prediction accuracy, as demonstrated using a cell signaling dataset. We apply the proposed methods to the Cancer Genome Atlas data to study the genetic and epigenetic pathways that underlie serous ovarian cancer.

Keywords: Bayesian network; Copula function; The Cancer Genome Atlas; Systems biology; Serous ovarian cancer

1 Introduction

In recent years, there has been considerable interest in estimating causal relationships between random variables in a graphical framework. Among several types of graphical models, Bayesian networks (BN) or equivalently, probability-weighted directed acyclic graphs (DAG) have received the most attention due to their simplicity and flexibility in modeling directed associations in the domain [1, 2, 3, 4]. The associations between d random variables can be summarized by a graph $\mathcal{G} = (V,E)$ in which $V = \{X_i | i = 1,2,...,d\}$ represents the set of variables and $E \subset V \times V$ represents the dependency between variables. Under the acyclicity and Markov assumptions, the joint likelihood function of $(X_1,...,X_d)$ in a BN has the following simple form based on the conditional densities:

$$f(X_1,...,X_d) = \prod_{i=1}^{d} f(X_i|\Pi_i), \tag{1}$$

where Π_i denotes the parent set of X_i, i.e., $\Pi_i = \{X_j | X_j \to X_i, X_j \in V \setminus \{X_i\}\}$ (Π_i can be empty).

The two most popular BN models are Gaussian Bayesian network (GBN) model [1] and multinomial Bayesian network (MBN) model [5], for continuous variables and discrete variables respectively. MBN models suffer from super-exponentially increasing number of parameters, therefore can only estimate small-scale networks in practice [5]. To deal with networks with relatively large number of nodes, GBN models have been commonly used due to their simple setup and efficient estimation. However, GBN models may fail to identify the true causalities when the joint distribution of interest is far from multivariate normal, for example, when the underlying distribution is bimodal or multimodal. To tackle the problem of non-normality, several new BN models have been developed, for instance, the logistic Bayesian network by Zhang et al. [4] which discretizes all the continuous variables to fit a multi-category logit model. Considerable work has also been done in nonparametric and semiparametric estimation of the BN structure. For instance, Voorman el al. [6] proposed the following nonparametric model to deal with non-normality issue:

$$X_i|\Pi_i = \sum_{X_k \in \Pi_i} f_{ik}(X_k) + \varepsilon_i,$$

where the $f_{ik}(\cdot)$ lies in some function space $\mathcal{F}$. The model by Voorman et al. focuses on estimating the

conditional mean $E(X_i|\Pi_i)$. It is essentially a generalized additive model without assuming the independence between ε_i and $f_{ik}(\cdot)$. However, this method relies on a known causal ordering of the true network which is unavailable in most cases.

In 2010, Elidan [7] introduced an innovative copula Bayesian network (CBN), a marriage between copula functions and graphical models, which extends conventional BN models to a more flexible framework. A CBN model constructs multivariate distribution with univariate marginals and a copula function C that links these marginals. In general, one can estimate marginals using parametric or non-parametric approach, and then use a small number of parameters to capture the dependence structure. However, as we shall see in a real data set (Section 4), the regular copula functions such as Gaussian copula may not be able to accurately depict multimodal joint distributions. In addition, the CBN model is subject to the choice of copula function for each local term. Motivated by Elidan's work, we extend the regular copula Bayesian networks to a mixture copula Bayesian network (MCBN) using finite mixture models, to better deal with non-normality, multimodality and heavy tails that are commonly seen in current massive genomic data. The parameters in a MCBN model can be efficiently estimated by a routine EM algorithm. As demonstrated by the real data, the performance of a two-component Gaussian MCBN is generally promising, and our model achieves reasonable accuracy in identifying the true edges in a sparse causal network.

The rest of this paper is organized as follows: In Section 2, we review Elidan's CBN model, and introduce the proposed MCBN model using a two-component Gaussian mixture for illustration. In Section 3, we present a heuristic local search approach combined with a routine EM algorithm for graph structure estimation, as well as the best-scoring network out of multiple predictions with random initial values. The comparison of three BN models is carried out over a cell signaling data set in Section 4. The new model is applied to the Cancer Genome Atlas (TCGA) data for serous ovarian cancer in Section 5. We discuss and conclude this paper in Sections 6 and 7.

2 Method

2.1 Copula and Elidan's Copula Bayesian network

Unless otherwise stated, we use $f(x_i) \equiv f_{X_i}(x_i)$, $F(x_i) \equiv F_{X_i}(x_i) \equiv P(X_i \leq x_i)$ as the marginals, and similarly for multivariate density $f(\mathbf{x}) \equiv f_{\mathbf{X}}(\mathbf{x})$. The formal definition of copula function is given below:

Definition 1. Let $(X_1, X_2, ..., X_d)$ be a vector of continuous random variables and $(F(x_1), F(x_2), ..., F(x_d))$ be the marginal distribution functions. The copula function of $(X_1, X_2, ..., X_d)$, $C : [0,1]^d \rightarrow [0,1]$, is defined as the cumulative distribution function of $(F(X_1), F(X_2), ..., F(X_d))$:

$$C(u_1, u_2, ..., u_d) = P(F(X_1) \leq u_1, F(X_2) \leq u_2, ..., F(X_d) \leq u_d). \tag{2}$$

By definition, a copula function is a multivariate distribution function where the marginals are uniform. By choosing an appropriate copula, one can generate multivariate distribution of any complex form. In practice, one can completely separate the choice of marginals and the choice of dependency patterns between random variables. Sklar's Theorem below guarantees that any multivariate distribution can be expressed with univariate marginals and a copula function which links these variables:

Theorem 1. Let $F(x_1, x_2, ..., x_d)$ be a multivariate distribution over real-valued d-dimension random vectors, then there exists a copula function that satisfies:

$$F(x_1, x_2, ..., x_d) = C(F(x_1), F(x_2), ..., F(x_d)). \tag{3}$$

Furthermore, the copula function C is unique when the marginal distribution $F(x_i)$ is continuous for $i \in \{1, 2, ..., d\}$.

By taking the first derivative for both sides of Equation (3), we can derive the copula density function defined as $c(F(x_1), F(x_2), ..., F(x_d)) = \frac{\partial^d C(F(x_1), ..., F(x_d))}{\partial F(x_1) ... \partial F(x_d)}$. The copula density is simply a ratio between the

joint density and the product of all the marginals:

$$c(F(x_1),F(x_2),...,F(x_d)) = \frac{f(x_1,...,x_d)}{\prod_i f(x_i)}. \tag{4}$$

An immediate consequence of Equation (4) is that $c(F(x_1),F(x_2),...,F(x_d)) = 1$ if and only if $X_1,...,X_d$ are independent. For a subset of variables $(Y,X_1,...,X_p)$, as $f(x_1,...,x_p) = \frac{\partial^p C(1,F(x_1),...,F(x_p))}{\partial x_1 ... \partial x_p}$, the conditional density $f(y|x_1,...,x_p)$ can be expressed as follows:

$$f(y|x_1,...,x_p) = \frac{c(F(y),F(x_1),...,F(x_p))f(y)\prod_{i=1}^{p} f(x_i)}{\frac{\partial C(1,F(x_1),...,F(x_p))}{\partial F(x_1)...\partial F(x_p)}\prod_{i=1}^{p} f(x_i)} = \frac{c(F(y),F(x_1),...,F(x_p))f(y)}{\int c(F(y),F(x_1),...,F(x_p))f(y)dy}. \tag{5}$$

Motivated by Equations (1) and (5), Elidan proposed a copula Bayesian network based on the following local density:

$$f(y|x_1,...,x_p) = f(y)G_c(y|x_1,...,x_p), \tag{6}$$

where $G_c(y|x_1,...,x_p) = \frac{c(F(y),F(x_1),...,F(x_p))}{\int c(F(y),F(x_1),...,F(x_p))f(y)dy} = \frac{c(F(y),F(x_1),...,F(x_p))}{E_Y(c(F(Y),F(x_1),...,F(x_p)))}$.

By Equation (6), we have the following decomposition for the joint density of variables in a Bayesian network:

Theorem 2. Let $(X_1,...,X_d)$ be d random variables (nodes) in a Bayesian network, and $\pi_i = \{x_j|X_j \in \Pi_i\}$. The joint density can be represented as follows:

$$f(x_1,...,x_d) = \prod_{i=1}^{d} G_c(x_i|\pi_i)\prod_{i=1}^{d} f(x_i). \tag{7}$$

Although the construction of local copulas can significantly reduce the complexity of the structure learning, choosing an appropriate copula for each local term $G_c(x_i|\pi_i)$ is essential. Elidan suggested a small set of pre-selected copula functions (or copula families) such as Gaussian copula, Frank's copula, Ali-Mikhail-Haq (AMH) copula and Gumbel-Barnett (GB) copula. However, as we will discuss in Section 4, these regular copula functions might be inadequate to model the complex dependence structure. To this end, we

extend the copula Bayesian network to a more flexible framework using finite mixture model.

2.2 A Mixture Copula Bayesian Network

For illustration purpose, we limit ourselves to Gaussian MCBN, but other mixture models such as Gamma mixture and Beta mixture models can be adapted similarly. The K-component Gaussian mixture copula for variables $(Y, X_1, ..., X_p)$ can be formulated as follows:

$$C(F(y), F(x_1), ..., F(x_p)) = \sum_{k=1}^{K} \alpha^{(k)} \Phi_{\Sigma_k}^{(k)}(\Phi^{-1}(F(y)), \Phi^{-1}(F(x_1)), ..., \Phi^{-1}(F(x_p))),$$

where $\alpha^{(k)}$ and $\Phi_{\Sigma_k}^{(k)}$ denote the weight and cumulative distribution function (CDF) of the kth Gaussian component respectively, and $\Phi^{-1}(\cdot)$ represents the quantile function of $N(0,1)$. The corresponding copula density can be obtained immediately:

$$c(F(y), F(x_1), ..., F(x_p)) = \frac{\partial C(F(y), F(x_1), ..., F(x_p))}{\partial F(y) \partial F(x_1) ... \partial F(x_p)} = \frac{\sum_{k=1}^{K} \alpha^{(k)} \phi_{\Sigma_k}^{(k)}(\Phi^{-1}(F(y)), \Phi^{-1}(F(x_1)), ..., \Phi^{-1}(F(x_p)))}{\phi(\Phi^{-1}(F(y)))\phi(\Phi^{-1}(F(x_1)))...\phi(\Phi^{-1}(F(x_p)))},$$

where $\phi(\cdot)$ represents the standard normal density function.

The Gaussian MCBN model above takes advantage of finite mixture model to better fit the bimodal and multimodal distributions. Similar as in the Elidan's copula Bayesian network, the marginals should be estimated prior to fitting the mixture copula, with either parametric or nonparametric method. We can, for example, fit the marginals using parametric or nonparametric method, then transform $(y, x_1, ..., x_p)$ to $(F(y), F(x_1), ..., F(x_p))$ using the fitted CDF functions. The transformed values will be used for estimating the copula function. Based on the estimated mixture copula for each local term in BN, we can calculate the joint likelihood by Equation (7).

3 Graph estimation using EM and local search algorithms

3.1 EM algorithm for finite Gaussian mixture

In this part, we introduce the EM algorithm to estimate the mixture copula for each local term $G_c(x_i|\pi_i)$.

For a given variable X_i and its parent set Π_i, the regular k-means algorithm can provide warm starts for the mean vector $\boldsymbol{\mu}_k$ (of dimension $|\Pi_i|+1$) and the covariance matrix $\boldsymbol{\Sigma}_k$ (of dimension $(|\Pi_i|+1)\times(|\Pi_i|+1)$) for each mixture component, as well as the mixing rate $\alpha^{(k)}$. Let $u_{hj}=\Phi^{-1}(F_{X_h}(x_{hj}))$ and $\boldsymbol{u}_j=\{u_{hj}\}$, where x_{hj} is the observed value for variable X_h and sample j, $X_h\in\{X_i,\Pi_i\}, j=1,2,...,N$. Let $\boldsymbol{z}=(z_1,...,z_N)$ be the vector of indicators for the membership of each sample (mutually exclusive and exhaustive), i.e., $\alpha^{(k)}=P(z_j=k), j=1,...,N$ and $\sum_{k=1}^{K}\alpha^{(k)}=1$. Denote $\Theta_k=(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)$ and $\boldsymbol{\Theta}=\{\Theta_k\}$, the EM algorithm with missing information $\boldsymbol{z}$ can be implemented as follows:

- **E Step**: Given current estimate of all the parameters ($\alpha^{(k)}$, $\boldsymbol{\Theta}$), we compute the weighted membership as follows:

$$\omega_{jk}\leftarrow P(z_j=k|\boldsymbol{u}_j,\boldsymbol{\Theta})=\frac{\phi_k(\boldsymbol{u}_j|z_j=k,\Theta_k)\alpha^{(k)}}{\sum_{m=1}^{K}\phi_m(\boldsymbol{u}_j|z_j=m,\Theta_m)\alpha^{(m)}}, 1\le j\le N, 1\le k\le K.$$

- **M Step:** Use data $\boldsymbol{u}_j$ and membership weights to update all the parameters:

$$\alpha^{(k)}\leftarrow\frac{\sum_{j=1}^{N}\omega_{jk}}{N},$$

$$\boldsymbol{\mu}_k\leftarrow\frac{1}{\sum_{j=1}^{N}\omega_{jk}}\sum_{j=1}^{N}\omega_{jk}\boldsymbol{u}_j,$$

$$\boldsymbol{\Sigma}_k\leftarrow\frac{1}{\sum_{j=1}^{N}\omega_{jk}}\sum_{j=1}^{N}\omega_{jk}(\boldsymbol{u}_j-\boldsymbol{\mu}_k)(\boldsymbol{u}_j-\boldsymbol{\mu}_k)^T.$$

Given an estimate of the graph structure $\mathcal{G}$ and the parameters $\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})$, the log-likelihood can be written as:

$$\ell(\hat{\mathcal{G}},\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})) = \log f(x_1,...,x_d|\hat{\mathcal{G}},\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})) = \sum_{i=1}^{d} G_c(x_i|\pi_i) + \sum_{i=1}^{d} f(x_i),$$

where the denominator of $G_c(x_i|\pi_i)$, i.e., $E_{X_i}(c(F(X_i),F(\pi_{i1}),...,F(\pi_{ip_i})))$ must be evaluated. Here we use notation p_i as the number of parents of X_i, i.e., $p_i = |\Pi_i|$. A simple idea for estimating $G_c(x_i|\pi_i)$ is to generate a list of Monte Carlo samples $(x^*_{i1},x^*_{i2},...,x^*_{iM})$ from $f(x_i)$, and by law of large numbers:

$$\frac{1}{M}\sum_{j=1}^{M} c(F(x^*_{ij}),F(\pi_{i1}),...,F(\pi_{ip_i})) \xrightarrow{a.s.} E_{X_i}(c(F(X_i),F(\pi_{i1}),...,F(\pi_{ip_i}))) \text{ as } M \to \infty,$$

where $x^*_{ij} \sim f(x_i)$. However, it is noteworthy that drawing samples from $f(x_i)$ might be complicated and time-consuming when marginals were estimated with nonparametric method. Further, the likelihood $\ell(\hat{\mathcal{G}},\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}}))$ may fail to converge due to the randomness of $G_c(x_i|\pi_i)$ estimation. Therefore for practical consideration, one can directly use all the observations as samples so that the convergence is guaranteed.

3.2 Score-based local search for learning MCBN

In this part, we introduce an efficient heuristic search algorithm based on Bayesian information criterion (BIC) to learn the structure of underlying network $\mathcal{G}$. The BIC score can be evaluated by the following formula:

$$BIC(\hat{\mathcal{G}},\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})) = -\ell(\hat{\mathcal{G}},\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})) + \frac{1}{2}\log(N)|\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})|,$$

where $\ell(\hat{\mathcal{G}},\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}}))$ represents log-likelihood function, $\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})$ is the set of all the parameters including the mixing rates, mean vectors and covariance matrices of Gaussian components, and $|\hat{\boldsymbol{\Theta}}(\hat{\mathcal{G}})|$ denotes the total number of free parameters in $\hat{\mathcal{G}}$. We start from a randomly generated network or empty network, and greedily advances through basic edge operation including addition, deletion and reversal, until BIC score reaches the minimum [7]. Unfortunately, this local search algorithm may easily get trapped in local maximum due to the high dimensionality and non-convexity of the likelihood function, making it impractical to find the global maximum. Enlightened by one of the reviewers, we conducted the heuristic search algorithm for

multiple times, each with a random initial value, and the best-scoring network (with minimum BIC score) was returned as the best predicted network.

4 Comparison with existing models

In this section, we compare the proposed MCBN model with two existing BN models, including the GBN model and Elidan's CBN model. We tested the three models using a flow cytometry dataset generated by Sachs et al. [8]. Sachs' data contains simultaneous measurement on 11 protein and phospholipid components, which was used for elucidating the signaling pathway structure in the cells of human immune system. The known network shown in Figure 1a is a Bayesian Network containing 11 nodes and 20 causal relations. Each causal edge in the network was well validated by experimental intervention, therefore this network structure is often used as the benchmark to assess the accuracy of different directed or undirected graphical models.

[Figure 1 about here]

Sachs' data has both continuous and discrete versions. In our analysis, we used the continuous data which was log-transformed and normalized by subtracting the mean and dividing by standard deviation. Three BN models were then applied to the preprocessed data for network structure learning, with detailed implementation as follows:

- **GBN**: We considered the linear regression setting, $X_i = \sum_{X_j \in \Pi_i} \beta_j X_j + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i^2)$, where the graph structure and parameters were estimated by a Blockwise Coordinate Descent (BCD) algorithm proposed by Fu and Zhou [1]. It has been shown that the BCD algorithm outperforms the popular PC algorithm [9] under regular settings. The intervention information was also incorporated in the modeling and a geometric sequence of 100 candidate tuning parameters $(\lambda_1, ..., \lambda_{50})$ were predefined $(\lambda_1 = 0.001, \lambda_{100} = 1)$. All the calculations were done using the source code provided by the authors (personal communication).

- **MCBN**: For simplicity of calculation, we considered a two-component Gaussian MCBN. The two-component Gaussian mixture model were also applied to the univariate marginals. Figure 2 shows two examples of fitted marginals for proteins *Art* and *Erk*.

[Figure 2 about here]

We set the maximum number of parental nodes at 5, i.e., $\max_i |\Pi_i| \leq 5$. The local search algorithm with BIC criterion was applied to BN structure learning, starting from an empty network. In the EM estimation of the copula function, we used k-means ($K = 2$) to obtain initial values for all the parameters, and used threshold $|\alpha_{i+1}^{(1)} - \alpha_i^{(1)}| \leq 10^{-4}$ for convergence, where $\alpha_{i+1}^{(1)}$ and $\alpha_i^{(1)}$ represent the resulting mixing rates in two consecutive EM runs.

- **CBN**: Elidan's CBN model can be treated as a special case of MCBN model when the copula density function has only one component (Gaussian copula). For the sake of comparison, all the marginals were also fitted using two-component Gaussian mixture. Same threshold as in MCBN was used as convergence criterion of the EM algorithm.

The estimated graphs by three different models are shown in Figure 1b-d. Table 1 summarizes true positive rate (TPR), false discovery rate (FDR) as well as running times by the three models (all timing were carried out on a Intel Xeon 3.2GH processor). In this comparison, a predicted edge is considered correct if both connection and direction are correct. It can be seen that the proposed MCBN model achieves significantly higher accuracy than the two existing models in terms of TPR and FDR, but it is more computationally expensive than the two simpler BNs. To further improve prediction, we conducted 100 predictions using random initial networks and obtained the best-scoring network, which contained 25 predicted edges. Out of 20 true edges, 13 were correctly identified in the best-scoring network. Furthermore, we compared different models in capturing the dependency pattern between variables. Figure 3 shows the scatterplot of *Art* and *Erk*, and the plots of simulated samples from three generative models. Compare to other models, the two-component Gaussian MCBN better depicted the multimodal dependency between *Akt* and *Erk*.

[Table 1 about here]

[Figure 3 about here]

To select the most confident edges, we calculated the log-likelihood decrease by removing one edge from the network. We found that an edge giving more likelihood increase has higher probability to be a true edge in the network. For instance, we selected the 10 most confident edges based on the likelihood change, and seven of them turned out to be true edges including *Akt→Erk*, *PKC→P38*, *PIP3→PIP2*, *PKA→Raf*, *PKC→JNK*, *PKC→Raf* and *PLCg→PIP2*. In addition, we evaluated the performance of our model in predicting the network skeleton (undirected edges). The proposed MCBN was compared with two simple alternatives including Pearson's correlation and Spearman's correlation. In this comparison, a predicted edge is considered correct as long as the connection is correct. Figure 4 shows the undirected networks by three approaches, and the TPR/FDR are summarized in Table 2.

[Figure 4 about here]

[Table 2 about here]

5 Application to TCGA ovarian cancer data

In this section, we applied the proposed MCBN to the Cancer Genome Atlas (TCGA) data [10], to study the interactions between oncomarkers that are associated with serous ovarian cancer. The TCGA data is one of the most comprehensive cancer genomic data sets, with more than 30 cancer types and subtypes which include but not limited to ovarian cancer, breast cancer, lung cancer, brain cancer and liver cancer. The sample sizes range from 50 to 1200 for different cancer types, and each sample is represented by both the molecular profile and clinical information. The molecular profile contains measurements for various types of (epi)genetic factors including gene expression quantification (both microarray and RNAseq), DNA methylation, single nucleotide polymorphism (SNP), copy number variation (CNV), somatic mutation, and microRNA etc. The clinical data provide information such as race, gender, tumor stage, outcome of surgery and resistance to chemotherapy.

The TCGA ovarian cancer data collected 567 tumor samples and 8 organ-specific normal controls. We incorporated three data types into our model including gene expression level, DNA methylation level (on gene promoter region) and CNV. The data were normalized using a quantile normalization method by Balstad et al. [11, 12] to correct the bias due to non-biological causes. In addition, we applied an effective method by Hsu et al. [13] to remove age and batch effects (three age groups are defined as < 40 y.o., $[40, 70]$ y.o., and > 70 y.o.). Hsu's method is essentially a median-matching and variance-matching strategy. For example, the batch-effect-adjusted gene expression value can be obtained as follows:

$$g^*_{ijk} = M_i + (g_{ijk} - M_{ij}) \frac{\hat{\sigma}_{g_i}}{\hat{\sigma}_{g_{ij}}},$$

where g_{ijk} represents the expression level of gene i from batch j and sample k, M_{ij} denotes the median of $g_{ij} = (g_{ij1}, ..., g_{ijn})$, M_i denotes the median of $g_i = (g_{i1}, ..., g_{iJ})$, $\hat{\sigma}_{g_i}$ and $\hat{\sigma}_{g_{ij}}$ are the standard deviation of g_i and g_{ij}, respectively.

The set of biomarkers was identified by a stepwise correlation-based feature selector (SCBS) by Zhang et al. [4], which mimics the hierarchy of underlying causal network. The SCBS algorithm starts from selecting the nodes that are strongly associated with the phenotype node and progressively select the nodes that are associated with the selected nodes in previous step. This algorithm is more effective in identifying phenotype-associated nodes, especially those nodes that are indirectly associated with the phenotype. By 3 runs of SCBS, we identified 73 oncomarkers including the expression level of 50 genes, CNV at 15 sites and methylation level at 8 sites. Among the 73 oncomarkers, many were previously reported in the literature including *BRCA1* [10], *BRCA2* [10], *RB1* [14], *PTEN* [15], and *OPCML* [16].

We then fit a MCBN model to study the regulatory relationships between these oncomarkers. The marginals were fitted by a two-component Gaussian mixture (other mixture models can also be used, e.g., Beta-mixture for DNA methylation). Figure 5 and 6 show several examples of the fitted marginals for *TP53* (expression level), *SPARC* (expression level), *BRCA1* (methylation level) and *NOTCH3* (methylation level).

[Figure 5 about here]

[Figure 6 about here]

In the biological network, we assumed that the genetic or epigenetic change (CNV and DNA methylation) cannot be induced by gene expression, and imposed this constraint into our modeling (Note: this assumption is completely from biological point of view and it can be dropped without affecting our modeling and computing). The predicted graph (in Figure 7, the best-scoring network from 100 predictions) contains 73 nodes connected by 124 directed edges. Many of the edges in the graph can be confirmed in the literature. To name a few, the edge between *AURKA* and *BRCA2* may be due to the fact that a negative regulatory loop exists between *AURKA* and *BRCA2* expression in the ovarian cancer[17]. The connection between *STAT3* and *ETV6* was suggested previously that *ETV6* is a negative regulator of *STAT3* activity [18]. The edges between *RAB25* (methylation) and *RAB25* (expression) and between *CSNK2A1* (CNV) and *CSNK2A1* (expression) had been reported in several studies [10, 19, 20]. Other highly ranked edges (based on likelihood increase) include but not limited to: *STAT3*→*DLEC1*, *PTEN*→*EGFR*, *RIMBP2*→*BRCA2* and *ARID1A*→*ERD* which can be confirmed in the literature of cancer biology [10, 21, 22, 23, 24, 25]. These findings demonstrate the effectiveness of the MCBN model. In addition, as illustrated in Figure 8, the two-component Gaussian MCBN is accurate in depicting the dependency between the gene expression level and methylation level.

[Figure 7 about here]

[Figure 8 about here]

6 Discussion

In this paper, we proposed a novel Bayesian network model to analyze recent cancer genomic data at the system level. The major innovation of our model is explicitly modeling the multimodal dependency structure between variables through copula function and more accurately estimating the causal network structure. The parameters in mixture copula were efficiently estimated by a routine EM algorithm, and the directed network structure was estimated by minimizing the BIC score.

The proposed Bayesian network model allows strict probabilistic inference of biological pathways, however, it also has several limitations. First, it lacks flexibility to model the cyclic mechanism due to the

acyclicity constraint, for instance, A→B→A, which though may exist in gene regulatory network. Second, the parameter estimation assumes sparsity of network for computational feasibility. If the true network is dense or locally dense, the weak causations may fail to be detected. Third, due to the model complexity, the implementation of MCBN is more computationally expensive than simpler BN models such as Gaussian BN model and regular copula BN model. For large data sets, one need reduce the number of variables by filtering out irrelevant and redundant variables, and then feed the selected variables into network model for causal inference.

It is noteworthy that the Gaussian MCBN used in the two illustrative examples can be generally adapted to other mixture models such as Gamma mixture and Beta mixture. The number of mixture components can be further increased depending on the complexity of the underlying dependency structure. For relatively small data set, it is also possible to conduct statistical testing to select the best number of mixture components for each local term, however, this will significantly increase the computational complexity.

7 Conclusions

Understanding the biological mechanism of cancers has significant practical importance for clinical diagnosis and treatment. In this paper, we developed a mixture copula Bayesian network model for causal inference using complex cancer genomic data. The proposed model is based on finite mixture models and copula functions, and it explicitly models multimodality in the data. The graph structure and model parameters can be efficiently estimated by a routine EM approach, embedded in a heuristic search algorithm based on Bayesian information criterion. The prediction could be further improved by selecting the best-scoring model from multiple predictions with random initial values. In addition, we proposed a likelihood-based approach to select the most confident edges. The proposed MCBN model was applied to a flow cytometry data and the TCGA ovarian cancer data for inferring the causal relationships between different biological features. Compare to existing Bayesian network models, MCBN better depicts the complex dependency structure between variables, therefore may better predict the underlying causal network.

Acknowledgement

Support has been provided in part by the Arkansas Biosciences Institute, the major research component of the Arkansas Tobacco Settlement Proceeds Act of 2000.

Authors' contributions

QZ conceived the study. QZ and XS analyzed the data. QZ wrote the manuscript. Both authors read and approved the final manuscript.

Competing Interests

The authors have declared that no competing interests exist.

Data Availability

The flow cytometry data by Sachs et al can be downloaded from `http://science.sciencemag.org/content/suppl/2005/04/21/308.5721.523.DC1`. TCGA ovarian cancer data can be downloaded via TCGA data portal `https://tcga-data.nci.nih.gov`.

Abbreviations

TCGA: The Cancer Genome Atlas; BN: Bayesian network; GBN: Gaussian Bayesian network; CBN: Copula Bayesian network; MCBN: Mixture copula Bayesian network; EM: Expectation-Maximization; BIC: Bayesian information criterion.

References

[1] F. Fu and Q. Zhou. Learning sparse causal gaussian networks with experimental intervention: Regularization and coordinate descent. *Journal of American Statistical Association*, 108(501):288–300, 2013.

[2] N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3):601–20, 2000.

[3] Y. Xu, J. Zhang, Y. Yuan, R. Mitra, P. Muller, and Y. Ji. A bayesian graphical model for integrative analysis of tcga data. *2012 IEEE International Workshop on Genomic Signal Processing and Statistics*, 2012(31), 2012.

[4] Q. Zhang, J. Burdette, and J.-P. Wang. Integrative network analysis of tcga data for ovarian cancer. *BMC Systems Biology*, 8(1338):1–18, 2014.

[5] B. Ellis and W. H. Wong. Learning causal bayesian network structures from experimental data. *Journal of American Statistical Association*, 103(482):778–789, 2008.

[6] A. Voorman, A. Shojaie, and D. Witten. Graph estimation with joint additive models. *Biometrika*, 101(1):85–101, 2014.

[7] G. Elidan. Copula bayesian networks. *Advances in neural information processing systems*, 23:559–567, 2010.

[8] K. Sachs, O. Perez, D. Pe'er, D.A. Lauffenburger, and G.P. Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–9, 2005.

[9] M. Kalisch and P. Buhlmann. Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research*, 8:613–636, 2007.

[10] The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474:609–15, 2011.

[11] B. Bolstad, R. Irizarry, M. Astrand, and T. Speed. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19(2), 2002.

[12] K. Mai and Q. Zhang. Identification of biomarkers for predicting the overall survival of ovarian cancer patients: a sparse group lasso approach. *International Journal of Statistics and Probability*, 5(6), 2016.

[13] F. Hsu, E. Serpedin, T. Hsiao, A. Bishop, E. Dougherty, and Y. Chen. Reducing confounding and suppression effects in tcga data: an integrated analysis of chemotherapy response in ovarian cancer. *BMC Genomics*, 13, 2012.

[14] H. Song, S. Ramus, D. Shadforth, L. Quaye, S. Kjaer, S. Gayther, and P. Pharoah. Common variants in rb1 gene and risk of invasive ovarian cancer. *Cancer Research*, 66(20):10220–6, 2006.

[15] Y. Takei, Y. Saga, H. Mizukami, T. Takayama, M. Ohwada, K. Ozawa, and M. Suzuki. Overexpression of pten in ovarian cancer cells suppresses i.p. dissemination and extends survival in mice. *Molecular Cancer Theraputics*, 7(3):704–11, 2008.

[16] A. Mckie, S. Vaughan, E. Zanini, I. Okon, L. Louis, C. DeSousa, M. Greene, N. Chayen, and H. Gabra. The opcml tumor suppressor functions as a cell surface repressor-adaptor, negatively regulating receptor tyrosine kinases in epithelial ovarian cancer. *Cancer Discovery*, 2(2):156–71, 2012.

[17] F. Yang, X. Guo, G. Yang, D. Rosen, and J. Liu. Aurka and brca2 expression highly correlate with prognosis of endometrioid ovarian carcinoma. *Modern Pathology*, 24(6), 2011.

[18] N. Schick, E. Oakeley, N. Hynes, and A. Badache. Tel/etv6 is a signal transducer and activator of transcription 3 (stat3)-induced repressor of stat3 activity. *Journal of Biological Chemistry*, 279(37):38787–96, 2004.

[19] K. Wrzeszczynski, V. Varadan, J. Byrnes, E. Lum, S. Kamalakaran, D. Levine, N. Dimitrova, M. Zhang, and R. Lucito. Identification of tumor suppressors and oncogenes from genomic and epigenetic features in ovarian cancer. *PLoS One*, 6(12), 2011.

[20] Y. Liu, X. Tao, L. Jia, K. Cheng, Y. Lu, Y. Yu, and Y. Feng. Knockdown of rab25 promotes autophagy and inhibits cell growth in ovarian cancer cells. *Molecular Medicine Reports*, 6(5):1006–12, 2012.

[21] J. Yuan, F. Zhang, and R. Niu. Multiple regulation pathways and pivotal biological functions of stat3 in cancer. *Scientific Reports*, 5(17663), 2015.

[22] A. Carracedo and P. Pandolfi. The pten-pi3k pathway: of feedbacks and cross-talks. *Oncogene*, 27:5527–41, 2008.

[23] J. Wu and C. Roberts. Arid1a mutation in cancer: Another epigenetic tumor suppressor. *Cancer Discovery*, 3(1):35–43, 2013.

[24] L. Xi, K. Brogaard, Q. Zhang, B. Lindsay, J. Widom, and J.-P. Wang. A locally convoluted cluster model for nucleosome positioning signals in chemical maps. *Journal of the American Statistical Association*, 109(505):48–62, 2014.

[25] E. Matveeva, J. Maiorano, Q. Zhang, A. Eteleeb, P. Converting, J. Chen, V. Infantino, S. Stamm, E. Rochka, J.-P. Wang, and Y. Fondufe-Mittendorf. Involvement of parp1 in the regulation of alternative splicing. *Cell Discovery*, 2(15046), 2016.

Tables and Figures

Table 1: Comparison of three different BN models

Model	P	TPR	FDR	Time(seconds)
Gaussian BN	27	0.40	0.704	5.60
Copula BN	24	0.40	0.667	1.39
Mixture Copula BN	25	0.650	0.480	22.67

Presented in the table are number of predicted edges (P), true positive rate (TPR), false discovery rate (FDR), as well as the CPU time (in seconds) by three different BN models.

Table 2: Comparison with Pearson's and Spearman's methods

Model	P	TPR	FDR
Pearson's correlation	25	0.55	0.56
Spearman's correlation	25	0.50	0.60
Mixture Copula BN	25	0.75	0.40

Presented in the table are number of undirected edges (P), true positive rate (TPR), false discovery rate (FDR) by three different approaches. For Pearson's and Spearman's methods, we selected top 25 edges with strongest correlation coefficients.

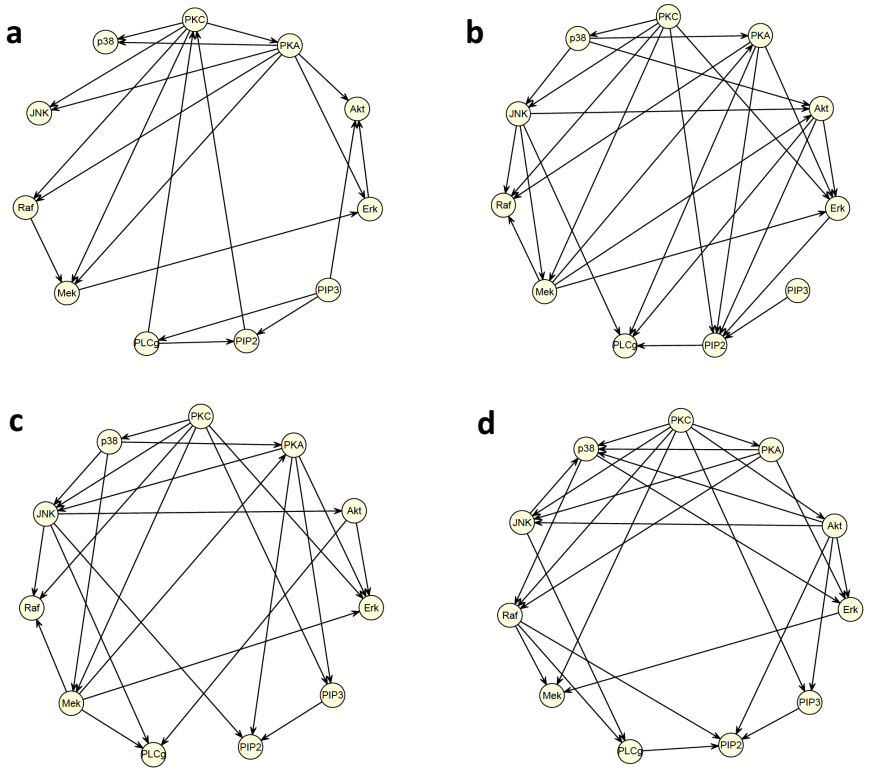

Figure 1: Comparison of three Bayesian network models on Sach's data: (a) The benchmark network; (b) Network predicted by GBN model; (c) Network predicted by Gaussian CBN model; (d) Network predicted by two-component Gaussian MCBN model.

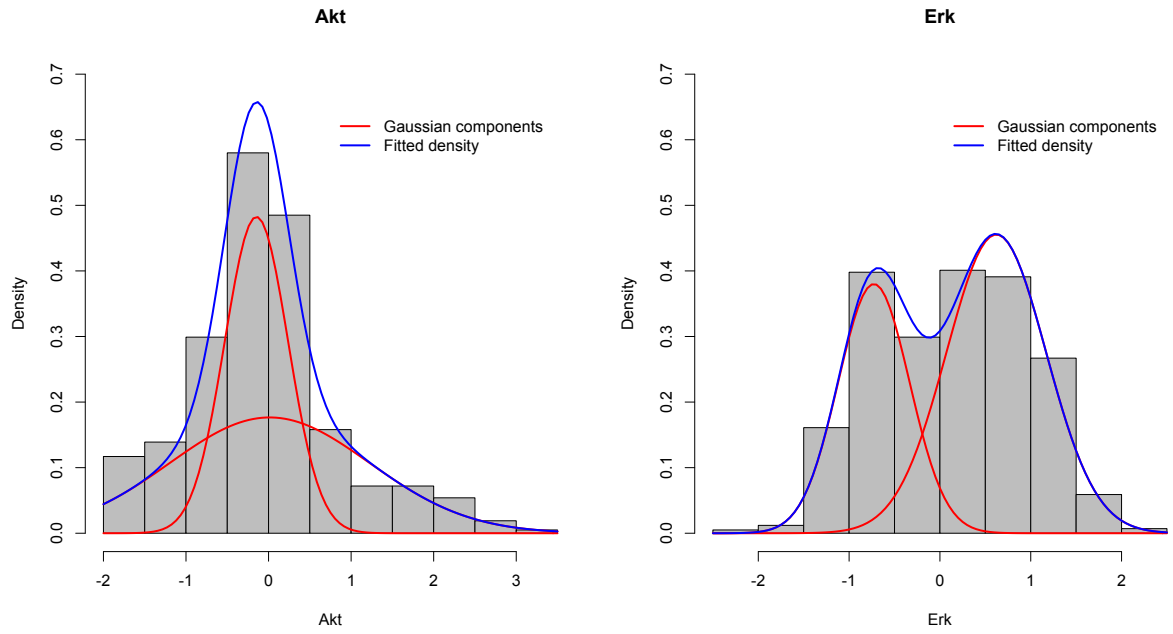

Figure 2: Fitted marginals by a two-component Gaussian mixture for the abundance of proteins *Akt* (left) and *Erk* (right).

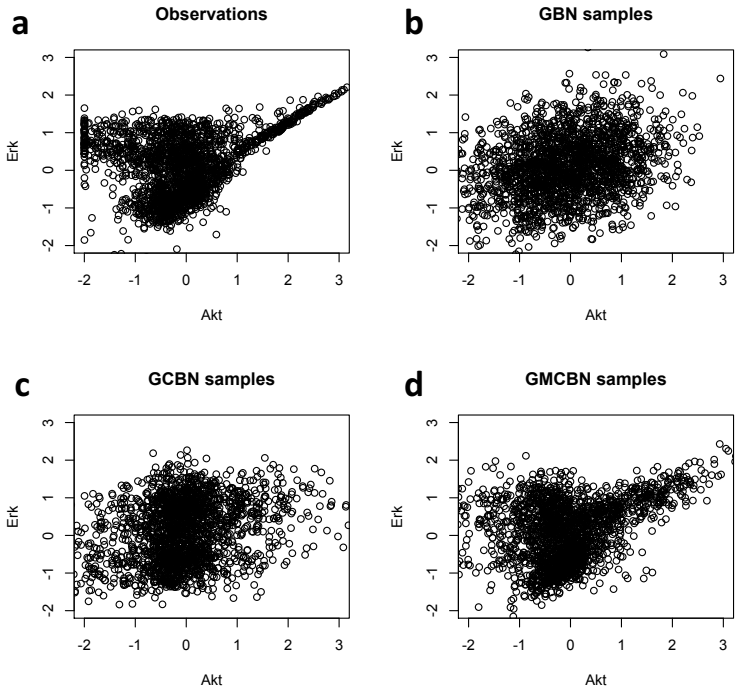

Figure 3: Dependence between proteins *Art* and *Erk*: (a) Observations; (b) Simulated samples from GBN; (c) Simulated samples from Gaussian CBN; (d) Simulated samples from two-component Gaussian MCBN.

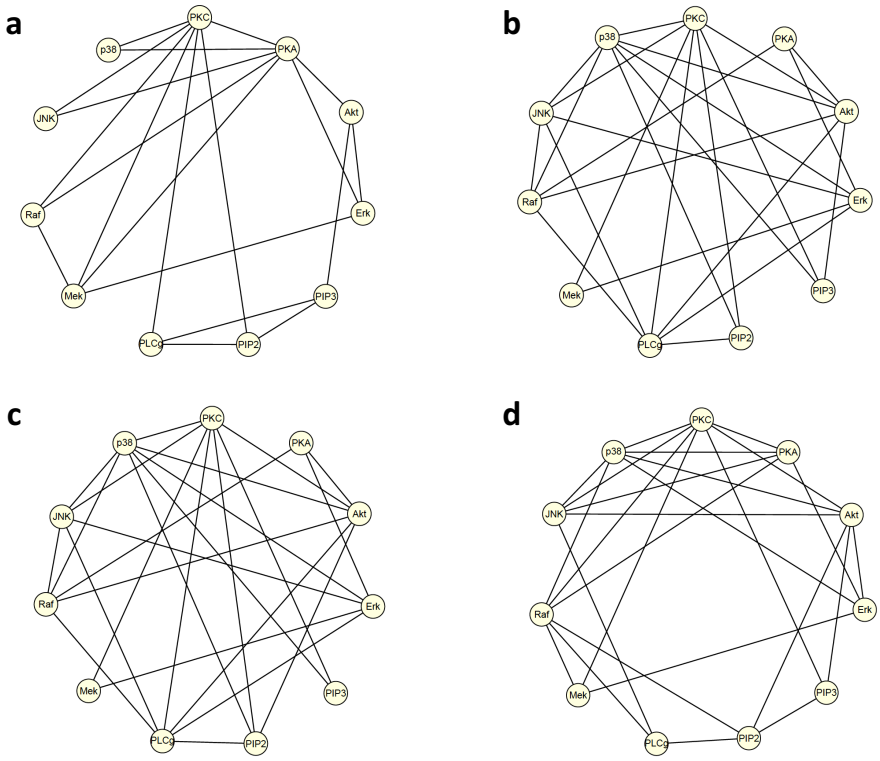

Figure 4: Comparison of three undirected networks: (a) Skeleton of the known network presented in Figure 1a; (b) Network consisted of top 25 edges based on Pearson's correlation coefficient; (c) Network consisted of top 25 edges based on Spearman's correlation coefficient; (d) Skeleton of network predicted by MCBN model presented in Figure 1d.

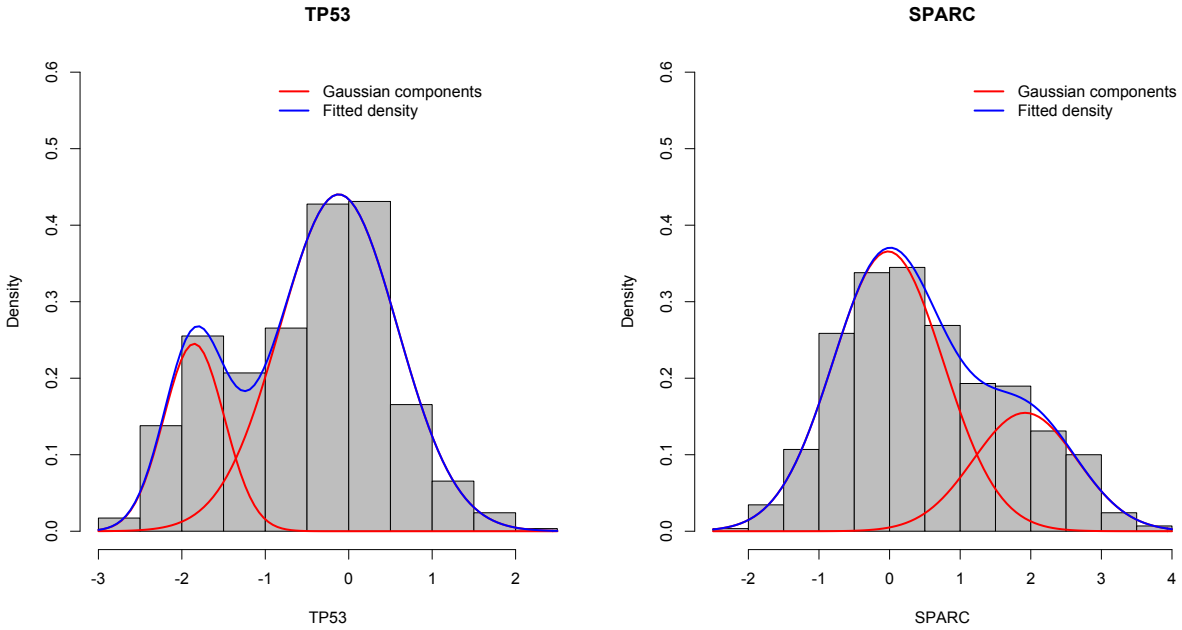

Figure 5: Fitted marginals by a two-component Gaussian mixture for the expression level of gene *TP53* (left) and *SPARC* (right).

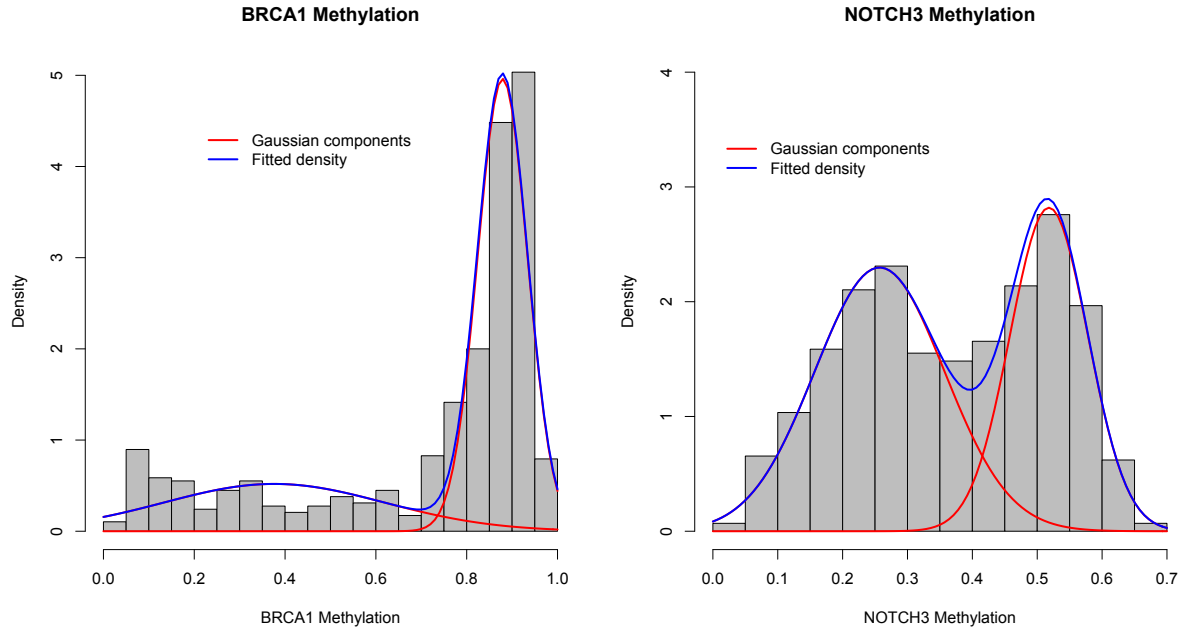

Figure 6: Fitted marginals by a two-component Gaussian mixture for the promoter methylation level of gene *BRCA1* (left) and *NOTCH3* (right).

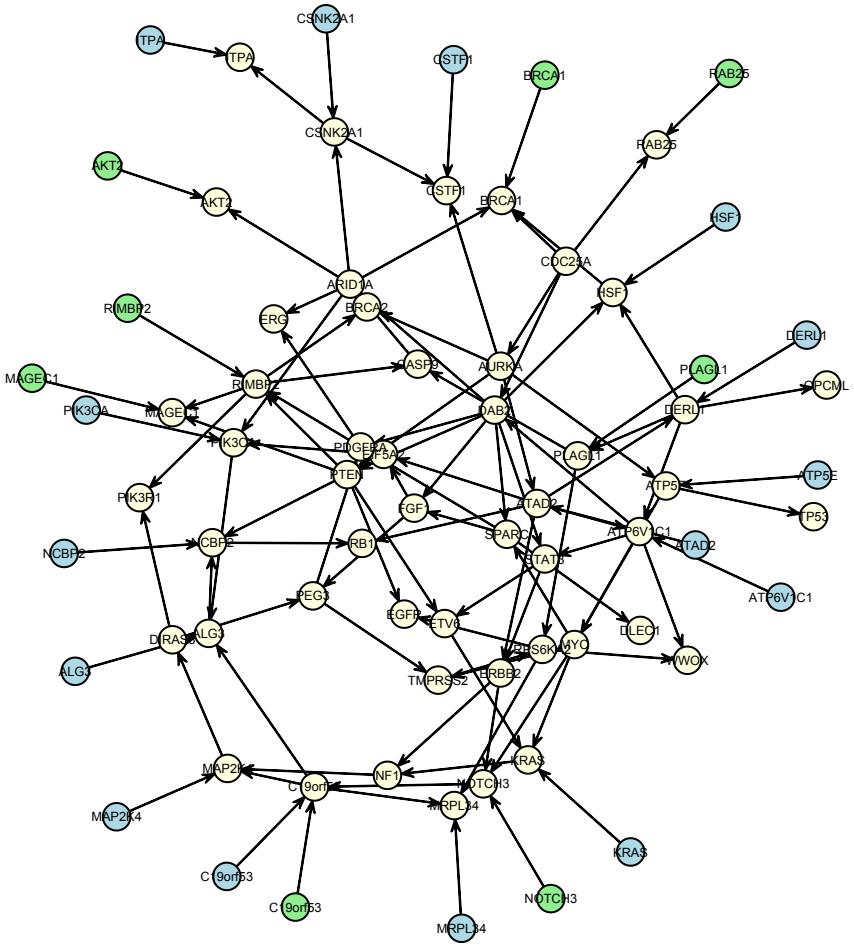

Figure 7: Predicted network by a two-component Gaussian MCBN model, containing the expression level of 50 genes (in light yellow), methylation level at 8 sites (in light green) and CNV at 15 sites (in light blue).

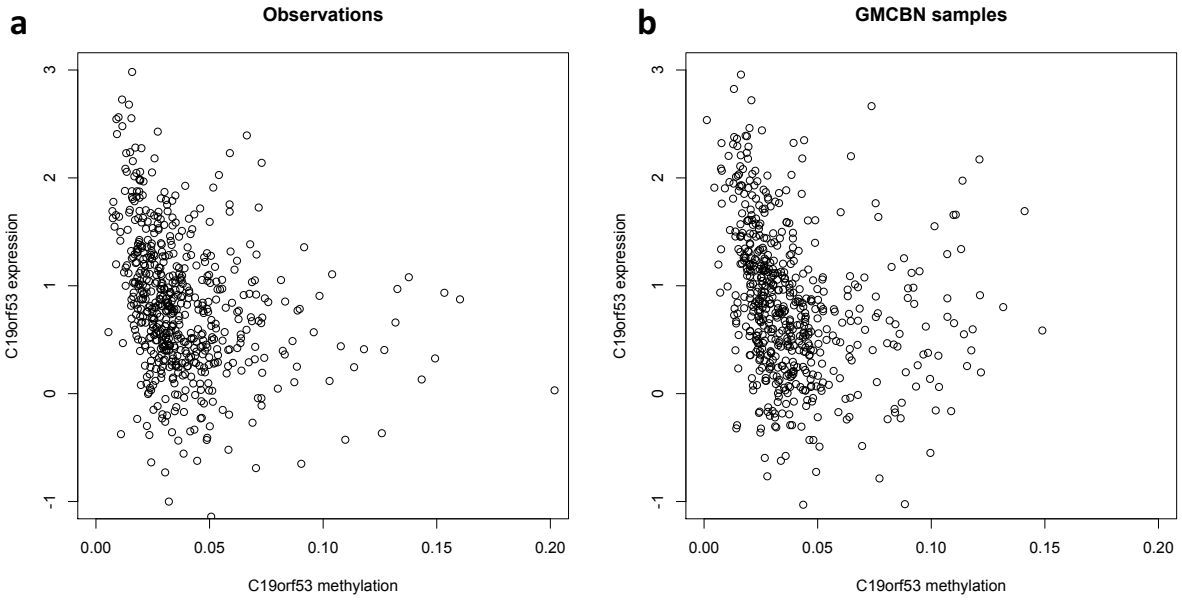

Figure 8: Dependence between the methylation level and expression level of gene *C19orf53*: (a) Observations; (b) Simulated samples from the two-component Gaussian MCBN.