Learning Bayesian Network Parameters with Domain Knowledge and Insufficient Data

Zhigao Guo BUCKLEYGUO@MAIL.NWPU.EDU.CN
Northwestern Polytechnical University
Xi'an (China)

Xiaoguang Gao CXG2012@NWPU.EDU.CN
Northwestern Polytechnical University
Xi'an (China)

Ruohai Di DRH@NWPU.EDU.CN
Northwestern Polytechnical University
Xi'an (China)

Abstract

To improve the learning accuracy of parameters in a Bayesian network (BN) from limited data, domain knowledge is often incorporated into the learning process as parameter constraints. Maximum a posteriori (MAP) based methods that use both data and constraints have been studied extensively. Among those methods, the qualitatively maximum a posteriori (QMAP) method exhibits high learning performance. In the QMAP method, when the data are limited, estimation from the data often fails to satisfy all the parameter constraints, which makes the overall QMAP estimation unreliable. To ensure that the QMAP estimation does not violate any given parameter constraint and further improve the learning accuracy, in this paper, we propose a qualitatively maximum a posteriori correction (QMAP-C) estimation algorithm, which regulates QMAP estimation by replacing the data estimation with a further constrained estimation. Experiments show that the proposed algorithm outperforms most of the existing parameter learning methods when the parameter constraints are correct.

Keywords: Bayesian network, parameter learning, insufficient data, domain knowledge

1 Introduction

A Bayesian network (BN) is a directed acyclic graph representing a model that combines probability theory and graphical model theory. The BN was systematically introduced by Judea Pearl (Pearl, 1988). Following approximately 30 years of global research, the BN has become a powerful tool for uncertainty analysis, and has been applied to various problems, such as gene analysis (Tamada et al., 2010), robot control (Infantes et al., 2011), fault diagnoses (Ibrahim and Beiu, 2011), target tracking (Mascaro et al., 2014), signal processing (Wachowski and Azimi-Sadjadi, 2014), ecosystem modeling (Landuyt et al., 2013), and educational measurement (Almond et al., 2015).

In general, learning a BN from data requires a data set of reasonable size, which is determined by the complexity of the network. With sufficient data, learning an accurate BN is tractable and can be accomplished by conventional methods, such as maximum likelihood (ML) (Redner and Walker, 1984). However, it is difficult to collect a large amount of data for some decision-making problems, such as rare disease diagnosis (Seixas et al., 2014), earthquake prediction (Hu et al., 2015), and parole assessment (Constantinou et al., 2015). In such cases, domain knowledge is often considered as supplementary information.

Domain experts find it convenient to provide qualitative parameter constraints in the form of $p_1 > 0.8$, $p_1 \approx p_2$, $p_1 > p_2$, $(p_1+p_2) > (p_3+p_4)$ (Helsper et al., 2004), etc., where p_1, p_2, p_3 and p_4 are parameters in a BN. Although such constraints look simple, they are very effective for improving BN modeling accuracy, especially when the given data is limited. In this paper, we focus on learning BNs in cases where the data is insufficient and domain experts have provided correct constraints on the parameters. By incorporating parameter constraints into the limited data, we present an improved MAP method.

The remainder of the paper is organized as follows: In Section 2, the related work on parameter learning with both sample data and parameter constraints is introduced. The studied problem is formalized and described in Section 3. An improved MAP method is presented in Section 4. In Section 5, experiments are presented to compare the proposed algorithm with other parameter learning algorithms on four benchmark BNs. In the last section, we summarize the presented work and note a few future research directions.

2 Related Work

BN parameter learning methods from insufficient data can be grouped into two types: MAP-based methods and non-MAP-based methods. In non-MAP-based methods, the BN parameters are computed by optimizing or regulating the constrained parameter estimation models. Among those methods, Wittig (Wittig and Jameson, 2000) proposed a constrained parameter learning algorithm. This algorithm can be applied to the cross-distribution parameter constraints, which define the relative relations between a pair of parameters over two different distributions. First, parameter constraints are constructed from qualitative expert knowledge. Then, an optimization model consisting of an entropy function and the parameter constraints is built. Finally, the built optimization model is optimized using the adaptive probabilistic networks method. Altendorf (Altendorf et al., 2005) also discussed a parameter learning method applicable to the cross-distribution constraints. What is interesting about the method is that it defines an objective function that integrates the parameter constraint model into the entropy function, and the function is then solved using the gradient-descent algorithm. Feelders (Feelders and Gaag, 2005) proposed an isotonic regression estimation (IRE) method for the cross-distribution constraints. Their algorithm employs the ML method to learn a set of initial parameters, and then elicits parameter orders from parameter constraints. The initial parameters are regulated by the algorithm so that the regulated parameters satisfy all the parameter orders. Isozaki (Takashi et al., 2009) suggested an minimum free energy (MFE) method that is suitable for axiomatic parameter constraints. Essentially, this method starts by constructing a free energy function, which consists of the Kullback-Leibler divergence and the entropy function, and is used as the objective function. Furthermore, the energy function and parameter constraints are integrated by the Lagrange multipliers, and the gradient-descent method is employed to solve the problem. The constrained maximum likelihood (CML) method was proposed by Campos (Campos et al., 2008). This method works to any convex parameter constraints. A convex optimization model is constructed with likelihood function and parameter constraints and the model is optimized using the convex optimization method. Campos discussed the constrained maximum entropy (CME) method applicable to any convex parameter constraints (Campos and Ji, 2008). In this method, an imprecise Dirichlet model (Walley, 1996) that combines the prior information and the data is created as a supplementary parameter constraint. Further, a convex optimization model containing the entropy function and the convex parameter constraints is constructed. The convex optimization method is applied to solve the formulated model. Zhou (Zhou et al., 2016) suggested a method for dealing with the cross-distribution constraints, named constrained optimization with flat prior (COFP). An objective function is considered in this method that combines the likelihood function and the penalty function derived from constraint violations. The objective function is solved using the sequential quadratic programming and the solutions are taken as the optimal parameters.

In MAP-based methods, BN parameters are computed as linear interpolation values of the sample observations and the prior information. Among them, the qualitative maximum a posteriori (QMAP) method (Chang and Wang, 2010) was designed to tackle any convex parameter constraints. The method requires a certain amount of possible parameters sampled from the parameter constraints. In addition, hyper-parameters of the prior Dirichlet distribution are determined as the products of a equivalent sample size and the mean values of the sampled parameters. The optimal parameters are then computed as the interpolations of the sample observations and the hyper-parameters. A method, named as Beta distribution approximation-based Bayesian estimation (BABE) was presented to address the intra-distribution parameter constraints (Di et al., 2014). Assuming the prior distribution obeys a uniform distribution under parameter constraints, this method approximates the prior distribution using the beta distribution. The optimal parameters are further computed as the interpolation values of the sample observations and the prior parameters. Other methods dealing with the intra-distribution constraints include the multi-nominal parameter learning with constraints (MPL-C) method (Zhou et al., 2014). This method counts the frequency of the configuration states of certain child and parent nodes, and then, an auxiliary BN model is built by integrating both the sample data and the parameter constraints. Furthermore, the optimal parameters are computed as the mean values of the probability distribution.

To the best of our knowledge, the QMAP method is one of the best performing algorithms. Essentially, QMAP estimation can be expressed as $\frac{N_{ijk}+M_{ijk}}{N_{ij}+M_{ij}}$, where $\frac{N_{ijk}}{N_{ij}}$ and $\frac{M_{ijk}}{M_{ij}}$ are the estimates from the data and the parameter constraints, respectively. N_{ij}, M_{ij}, N_{ijk} and M_{ijk} are the number of observations from the data set and the equivalent data set. In this paper, we assume that the parameter constraints are all correct, which means the estimation $\frac{M_{ijk}}{M_{ij}}$ satisfies all the parameter constraints. Then, when the given data are limited, the data estimation $\frac{N_{ijk}}{N_{ij}}$ often violates the parameter constraints. In such cases, the estimation $\frac{M_{ijk}}{M_{ij}}$ will be negatively influenced by $\frac{N_{ijk}}{N_{ij}}$, which causes the overall QMAP estimation to fail to satisfy all the parameter constraints. To solve that problem, we propose a qualitatively maximum a posteriori correction (QMAP-C) estimation algorithm.

3 Preliminaries

3.1 Bayesian Network

A BN is characterized by its structure and parameters. Figure 1 shows a typical BN, i.e., the brain tumor BN (Cooper, 1984), whose nodes C, BT, SH, MC, CT and ISC denote coma, brain tumor, severe headaches, metastatic cancer, computed tomography scan, respectively.

In the brain tumor BN, nodes such as BT, MC, and ISC represent disease symptoms or diagnoses. Arrows from one node to another represent the influence of the top nodes on the bottom nodes. Parameters such as $P(C|BT, ISC)$ represent the strength of the joint influence exerted by the symptom nodes BT and ISC on the diagnosis node C. In this study, our objective is to learn the parameters of a discrete BN, whose structure is known beforehand.

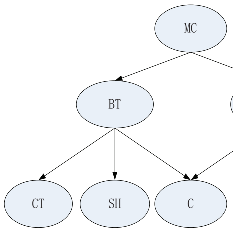

Figure 1: Brain tumor Bayesian network

3.2 Parameter Learning in a Bayesian Network

Parameter learning in a BN entails parameter estimation from a given sample data set. In this paper, samples with missing values are not considered. For a network with n node variables, parameter estimation can be expressed as a maximization problem of the log-likelihood function, which is

$$logP(D|\theta, G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} log\theta_{ijk}. \tag{1}$$

where θ denotes the parameters, G represents the network structure and r_i is the total state number of node i. Based on the BN decomposability property, the parameter estimation of a network can be decomposed into the product of the independent estimation of individual variable nodes and the ML estimation $\widehat{\theta}_{ijk}$ equals $\frac{N_{ijk}}{N_{ij}}$, where $N_{ij} = \sum_{k=1}^{r_i} N_{ijk}$.

3.3 Common Bayesian Network Parameter Constraints

Generally, eight types of parameter constraints (Wellman, 1990; Chang and Wang, 2010) can be formalized from qualitative domain knowledge, which are defined in Table 1.

Type	Form	Property	Type	Form	Property
1	$\theta_{ijk} \leq \theta_{ijk'}$	convex	5	$\theta_{ijk} \leq \theta_{ij'k}$	convex
2	$\theta_{ijk} \leq \theta_{i'j'k'}$	convex	6	$\theta_{ijk} \approx \theta_{i'j'k'}$	convex
3	$\sum_{k=1}^{r_i} \theta_{ijk} = 1$	convex	7	$\alpha_{ijk} \leq \theta_{ijk} \leq \beta_{ijk}$	convex
4	$\theta_{ij_1k} + \theta_{ij_2k} \leq \theta_{ij_3k} + \theta_{ij_4k}$	convex	8	$\theta_{ij_1k} * \theta_{ij_2k} \leq \theta_{ij_3k} * \theta_{ij_4k}$	concave

Table 1: Common Bayesian network parameter constraints.

In Table 1, parameter constraints of type (1-8) represent the intra-distribution constraint, inter-distribution constraint, axiomatic constraint, additive synergy constraint, cross-distribution constraint, approximate-equality constraint, range constraint, and product synergy constraint, respectively.

4 The Method

4.1 Qualitatively Maximum a Posteriori Estimation

The QMAP estimation is a posteriori estimation that incorporates both quantitative data and qualitative constraints. Its log-form score function is expressed by Eq. (2) and can be decomposed into a data likelihood and prior probability distribution:

$$logP(\theta|G,D,\Omega) = logP(D|\theta,G) + logP(\theta|\Omega,G) - P(D|\Omega,G), \tag{2}$$

where Ω is the set of parameter constraints. The data likelihood equals the conventional log-likelihood function expressed by Eq (1). The prior distribution is defined by the parameter constraints, from which independent prior parameter instances can be sampled. Thus, the log-likelihood prior probability distribution can be expressed as

$$logP(\theta|\Omega,G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} M_{ijk} log\theta_{ijk}, \tag{3}$$

where $M_{ijk} = \widehat{A} \cdot \alpha_{ijk}$, $\widehat{A}$ is the optimal equivalent sample size and α_{ijk} is the mean value of the parameters sampled from the parameter constraints. As such, the overall QMAP log-likelihood score function can be expressed as

$$logP(\theta|G,D,\Omega) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} (N_{ijk} + M_{ijk}) log\theta_{ijk} - P(D|\Omega,G). \tag{4}$$

Finally, the maximum estimation of the QMAP score function is computed as

$$\widehat{\theta}_{ijk} = \frac{N_{ijk} + M_{ijk}}{\sum\limits_{k=1}^{r_i}(N_{ijk} + M_{ijk})} = \frac{N_{ijk} + \widehat{A} \cdot \alpha_{ijk}}{N_{ij} + \widehat{A}}. \tag{5}$$

4.2 Qualitatively Maximum a Posteriori Correction Estimation

The qualitatively maximum a posteriori correction (QMAP-C) estimation is a improved QMAP estimation, whose score function can be expressed as

$$P(\theta|G,D,\Omega) = P(D|\theta,G,\Omega)P(\theta|G,\Omega)/P(D|\Omega,G). \tag{6}$$

The log-form score function of the QMAP-C estimation can be further expressed as

$$logP(\theta|G,D,\Omega) = logP(D|\theta,G,\Omega) + logP(\theta|G,\Omega) - logP(D|\Omega,G), \tag{7}$$

where $logP(D|\theta,G,\Omega)$ is not the conventional log-likelihood function but a constrained log-likelihood model, given by

$$logP(D|\theta,G) = \sum_{i=1}^{n}\sum_{j=1}^{q_i}\sum_{k=1}^{r_i} N_{ijk} log\theta_{ijk}, \tag{8}$$

$$s.t. \begin{array}{l} f(\theta) = 0 \\ h(\theta) \leq 0. \end{array} \tag{9}$$

Thus, the overall QMAP-C log-likelihood score function can be expressed as

$$logP(\theta|G, D, \Omega) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} (N'_{ijk} + M_{ijk}) log\theta_{ijk} - logP(D|\Omega, G). \tag{10}$$

Finally, the maximum estimation of the QMAP-C score function is given by

$$\widehat{\theta}_{ijk} = \frac{N'_{ijk} + M_{ijk}}{\sum\limits_{k=1}^{r_i} (N'_{ijk} + M_{ijk})} = \frac{N_{ij} \cdot \theta'_{ijk} + \widehat{A} \cdot \alpha_{ijk}}{N_{ij} + \widehat{A}}, \tag{11}$$

where θ'_{ijk} are the optimization solutions of the model expressed by Eqs. (8-9). In general, the QMAP-C estimation can be summarized as follows:

Algorithm 1 QMAP-C Estimation

```
Input:
    Data $D$;
    Network structure $G$;
    Parameter constraints $\Omega$;
    Equivalent sample size $A = \{1, ..., 20\}$;
Output:
    Estimated parameters $\theta$;
for $i = 1$ to $n$ do
    Sample $\alpha_i = \{\alpha_{ijk}\}$ from constraints;
    Determine $\widehat{A}$ via cross-validation;
    for $j = 1$ to $q_i$ do
        for $k = 1$ to $r_i$ do
            Compute the parameter $\theta'_{ijk}$;
            $\widehat{\theta}_{ijk} = \frac{N_{ij} \cdot \theta'_{ijk} + \widehat{A} \cdot \alpha_{ijk}}{N_{ij} + \widehat{A}}$;
        end for
    end for
end for
```

Algorithm 2 Determining $\widehat{A}$ via cross-validation method

```
Input:
    Node $i$;
    Data $D$;
    Network structure $G$;
    Sampled parameters $\alpha_i = \{\alpha_{ijk}\}$;
    Equivalent sample size $A = \{1, ..., 20\}$;
Output:
    Optimal equivalent sample size $\widehat{A}$;
Partition $D$ into $\{D_1, D_2, ..., D_{10}\}$;
for $w = 1$ to 20 do
    for $m = 1$ to 10 do
        $D_{testing} = D_m$;
        $D_{training} = \{D/D_m\}$;
        for $j = 1$ to $q_i$ do
            for $k = 1$ to $r_i$ do
                $D_{testing}$: $N'_{ij} = \{N'_{ijk}\}$;
                $D_{training}$: $N_{ij} = \{N_{ijk}\}$;
                $\theta_{ijk} = \frac{N_{ijk} + A_w \cdot \alpha_{ijk}}{N_{ij} + A_w}$;
            end for
        end for
        $L_m(G, D, \alpha_i, A_w) = \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N'_{ijk} log\theta_{ijk}$;
    end for
    $L(G, D, \alpha_i, A_w) = \sum_{m=1}^{10} L_m(G, D, \alpha_i, A_w)$;
end for
$\widehat{A} = arg \max_{A_w} L(G, D, \alpha_i, A_w)$;
```

5 Experiments

We carried out experiments on four benchmark BNs, Asia, Alarm, Win95pts, Andes, to compare the performance of different algorithms, which are evaluated by KL divergence (Kullback and Leibler, 1951) and running time, respectively. In the experiments, we considered eight algorithms: ML, ME, MAP, CO, CML, CME, QMAP, and QMAP-C.

5.1 Parameter Learning with Different Sample Sizes

Experiment settings: (1) The sample sizes considered were 50, 100, 150, and 200. (2) The parameter constraints were generated from true parameters of the networks, and the maximum number of constraints for each node was 30. The average KL divergence and running time for different networks are summarized in Tables 2 and 3, where the best results are highlighted in bold.

	ML	ME	MAP	CO	CML	CME	QMAP	QMAP-C
(a) Asia network								
50	0.559±0.391	0.165±0.025	0.125±0.043	0.096±0.054	0.048±0.019	0.025±0.009	0.009±0.002	**0.006±0.001**
100	0.386±0.246	0.132±0.017	0.093±0.026	0.089±0.037	0.049±0.015	0.021±0.008	0.008±0.002	**0.006±0.001**
150	0.325±0.237	0.114±0.018	0.078±0.025	0.089±0.039	0.047±0.013	0.019±0.007	0.007±0.002	**0.006±0.001**
200	0.264±0.212	0.101±0.018	0.067±0.022	0.085±0.040	0.044±0.012	0.018±0.008	0.007±0.002	**0.005±0.001**
(b) Alarm network								
50	0.541±0.074	0.236±0.007	0.220±0.014	0.323±0.032	0.211±0.031	0.130±0.005	0.125±0.002	**0.124±0.002**
100	0.494±0.063	0.202±0.009	0.187±0.015	0.218±0.029	0.194±0.029	0.125±0.005	0.117±0.002	**0.116±0.002**
150	0.459±0.075	0.180±0.007	0.166±0.013	0.198±0.031	0.176±0.028	0.115±0.004	0.111±0.002	**0.109±0.002**
200	0.434±0.068	0.166±0.007	0.152±0.012	0.186±0.028	0.165±0.025	0.108±0.005	0.106±0.002	**0.105±0.002**
(c) Win95pts network								
50	0.655±0.096	0.244±0.003	0.228±0.011	0.210±0.031	0.210±0.041	0.142±0.001	0.131±0.001	**0.130±0.001**
100	0.611±0.078	0.224±0.003	0.218±0.011	0.205±0.029	0.206±0.040	0.138±0.001	0.128±0.001	**0.127±0.001**
150	0.568±0.074	0.210±0.002	0.213±0.011	0.198±0.019	0.190±0.031	0.135±0.001	0.125±0.001	**0.125±0.001***
200	0.511±0.059	0.201±0.002	0.211±0.012	0.182±0.026	0.172±0.027	0.132±0.001	0.123±0.001	**0.123±0.001***
(d) Andes network								
50	1.021±0.064	0.179±0.002	0.188±0.010	0.361±0.027	0.202±0.022	0.079±0.001	0.050±0.001	**0.047±0.001**
100	0.827±0.067	0.134±0.002	0.147±0.009	0.326±0.024	0.193±0.017	0.064±0.002	0.045±0.001	**0.042±0.001**
150	0.744±0.065	0.111±0.001	0.129±0.010	0.314±0.024	0.179±0.021	0.056±0.002	0.042±0.001	**0.038±0.001**
200	0.671±0.059	0.096±0.002	0.117±0.009	0.294±0.027	0.168±0.021	0.049±0.002	0.039±0.001	**0.036±0.001**

* The QMAP-C estimation is slightly better than the QMAP estimation. And the detailed KL divergence are: QMAP (0.1252±0.0005, 0.1232±0.0005), QMAP-C (0.1246±0.0005, 0.1226±0.0005).

Table 2: KL divergence of different algorithms under different sample sizes.

	ML	ME	MAP	CO	CML	CME	QMAP	QMAP-C
(a) Asia network								
50	**0.000±0.000**	0.151±0.006	**0.000±0.000**	0.330±0.015	0.166±0.003	0.151±0.002	0.063±0.001	0.227±0.003
100	**0.001±0.000**	0.150±0.004	**0.001±0.000**	0.328±0.012	0.168±0.004	0.152±0.003	0.112±0.001	0.277±0.004
150	**0.001±0.000**	0.149±0.004	**0.001±0.000**	0.317±0.004	0.167±0.003	0.152±0.003	0.161±0.001	0.325±0.003
200	**0.001±0.000**	0.148±0.004	**0.001±0.000**	0.316±0.004	0.168±0.003	0.152±0.003	0.210±0.001	0.375±0.003
(b) Alarm network								
50	**0.000±0.000**	0.028±0.001	**0.000±0.000**	0.088±0.001	0.034±0.001	0.031±0.000	0.044±0.000	0.077±0.001
100	**0.001±0.000**	0.028±0.000	**0.001±0.000**	0.088±0.001	0.034±0.001	0.031±0.000	0.083±0.000	0.116±0.001
150	**0.001±0.000**	0.028±0.001	**0.001±0.000**	0.088±0.001	0.033±0.000	0.031±0.000	0.122±0.001	0.154±0.001
200	**0.001±0.000**	0.029±0.000	**0.001±0.000**	0.088±0.000	0.033±0.000	0.032±0.000	0.161±0.001	0.193±0.001
(c) Win95pts network								
50	**0.000±0.000**	0.049±0.001	**0.000±0.000**	0.123±0.002	0.052±0.001	0.050±0.000	0.059±0.000	0.112±0.001
100	**0.001±0.000**	0.049±0.001	**0.001±0.000**	0.127±0.002	0.054±0.001	0.050±0.000	0.113±0.000	0.166±0.001
150	**0.001±0.000**	0.049±0.001	**0.001±0.000**	0.126±0.002	0.054±0.001	0.051±0.000	0.165±0.001	0.219±0.001
200	**0.001±0.000**	0.049±0.001	**0.001±0.000**	0.125±0.001	0.054±0.001	0.051±0.000	0.219±0.001	0.271±0.001
(d) Andes network								
50	**0.000±0.000**	0.066±0.003	**0.000±0.000**	0.189±0.007	0.077±0.002	0.071±0.002	0.061±0.002	0.138±0.004
100	**0.001±0.000**	0.067±0.003	**0.001±0.000**	0.188±0.009	0.078±0.003	0.072±0.003	0.115±0.006	0.193±0.009
150	**0.001±0.000**	0.067±0.003	**0.001±0.000**	0.187±0.009	0.078±0.003	0.073±0.003	0.168±0.008	0.246±0.012
200	**0.001±0.000**	0.067±0.002	**0.001±0.000**	0.185±0.006	0.078±0.002	0.073±0.002	0.219±0.006	0.296±0.008

Table 3: Running time (seconds) of different algorithms under different sample sizes.

Experiment analysis: (1) In nearly all cases, QMAP-C performed better on learning accuracy than the other learning algorithms. In addition, with increasing data size, the QMAP estimation gradually approached the QMAP-C estimation, because as the amount of data increases, the purely data-driven estimation $\frac{N_{ijk}}{N_{ij}}$ becomes less likely to violate the constraints. (2) The QMAP-C method is more time-consuming than the QMAP method. The explanation is that it takes more computation complexity to optimize the equivalent sample size and regulate the data-driven estimation $\frac{N_{ijk}}{N_{ij}}$.

5.2 Parameter Learning with Different Constraint Sizes

Experiment settings: (1) The sample sizes for all the networks were set to 50. (2) For each node, at most, 30 parameter constraints were generated and an increasing number of constraints varying from 25% to 100% were used for parameter learning. The average KL divergence and running time for different networks are summarized in Tables 4 and 5.

	ML	ME	MAP	CO	CML	CME	QMAP	QMAP-C
(a) Asia network								
25%	0.603±0.342	0.168±0.027	0.130±0.034	0.256±0.204	0.224±0.208	0.101±0.033	0.099±0.028	**0.097±0.028**
50%	0.603±0.342	0.168±0.027	0.130±0.034	0.161±0.146	0.133±0.146	0.052±0.029	0.039±0.019	**0.035±0.019**
75%	0.603±0.342	0.168±0.027	0.130±0.034	0.084±0.079	0.042±0.068	0.033±0.021	0.014±0.009	**0.009±0.008**
100%	0.603±0.342	0.168±0.027	0.130±0.034	0.056±0.038	0.018±0.006	0.024±0.012	0.006±0.003	**0.002±0.002**
(b) Alarm network								
25%	0.529±0.071	0.318±0.007	0.219±0.014	0.372±0.045	0.359±0.046	0.271±0.008	0.260±0.008	**0.259±0.008**
50%	0.529±0.071	0.318±0.007	0.219±0.014	0.305±0.034	0.279±0.036	0.215±0.006	0.203±0.007	**0.202±0.007**
75%	0.529±0.071	0.318±0.007	0.219±0.014	0.261±0.041	0.224±0.039	0.173±0.005	0.165±0.005	**0.163±0.005**
100%	0.529±0.071	0.318±0.007	0.219±0.014	0.228±0.033	0.181±0.028	0.157±0.004	0.138±0.002	**0.136±0.002**
(c) Win95pts network								
25%	0.541±0.084	0.243±0.004	0.231±0.008	0.339±0.050	0.339±0.051	**0.201±0.004**	0.204±0.004	0.203±0.004
50%	0.541±0.084	0.243±0.004	0.231±0.008	0.261±0.035	0.257±0.032	0.171±0.004	0.164±0.003	**0.164±0.003***
75%	0.541±0.084	0.243±0.004	0.231±0.008	0.211±0.024	0.215±0.026	0.152±0.003	0.142±0.002	**0.141±0.002**
100%	0.541±0.084	0.243±0.004	0.231±0.008	0.181±0.017	0.187±0.021	0.144±0.002	0.132±0.001	**0.131±0.001**
(d) Andes network								
25%	1.041±0.091	0.179±0.003	0.189±0.012	0.619±0.047	0.573±0.053	**0.141±0.003**	0.146±0.002	0.147±0.002
50%	1.041±0.091	0.179±0.003	0.189±0.012	0.462±0.039	0.400±0.041	0.115±0.004	0.096±0.002	**0.093±0.002**
75%	1.041±0.091	0.179±0.003	0.189±0.012	0.352±0.029	0.285±0.029	0.095±0.004	0.065±0.002	**0.061±0.002**
100%	1.041±0.091	0.179±0.003	0.189±0.012	0.255±0.022	0.209±0.022	0.082±0.003	0.047±0.001	**0.042±0.001**

* The QMAP-C estimation is slightly better than the QMAP estimation. And the detailed KL divergence are: QMAP (0.1642±0.0029), QMAP-C (0.1637±0.0029).

Table 4: KL divergence of different algorithms under different constraint sizes.

	ML	ME	MAP	CO	CML	CME	QMAP	QMAP-C
(a) Asia network								
25%	**0.000±0.000**	0.196±0.014	**0.000±0.000**	0.257±0.018	0.205±0.009	0.195±0.008	0.094±0.008	0.297±0.012
50%	**0.001±0.000**	0.197±0.008	**0.001±0.000**	0.365±0.019	0.218±0.010	0.193±0.008	0.094±0.007	0.308±0.015
75%	**0.001±0.000**	0.193±0.008	**0.001±0.000**	0.515±0.026	0.223±0.009	0.199±0.008	0.093±0.006	0.318±0.013
100%	**0.001±0.000**	0.195±0.008	**0.001±0.000**	0.724±0.033	0.242±0.011	0.209±0.006	0.093±0.005	0.332±0.013
(b) Alarm network								
25%	**0.000±0.000**	0.034±0.002	**0.000±0.000**	0.053±0.003	0.035±0.002	0.035±0.002	0.054±0.003	0.088±0.004
50%	**0.001±0.000**	0.034±0.002	**0.001±0.000**	0.088±0.005	0.039±0.002	0.037±0.002	0.054±0.004	0.092±0.006
75%	**0.001±0.000**	0.034±0.002	**0.000±0.000**	0.139±0.008	0.043±0.003	0.039±0.002	0.054±0.004	0.097±0.005
100%	**0.001±0.000**	0.034±0.002	**0.000±0.000**	0.202±0.011	0.049±0.003	0.043±0.002	0.054±0.003	0.101±0.005
(c) Win95pts network								
25%	**0.000±0.000**	0.059±0.002	**0.000±0.000**	0.079±0.005	0.059±0.003	0.059±0.003	0.078±0.005	0.136±0.008
50%	**0.001±0.000**	0.059±0.003	**0.001±0.000**	0.125±0.007	0.062±0.003	0.060±0.003	0.078±0.005	0.140±0.007
75%	**0.001±0.000**	0.059±0.002	**0.001±0.000**	0.193±0.009	0.069±0.004	0.064±0.003	0.080±0.004	0.149±0.009
100%	**0.001±0.000**	0.059±0.002	**0.001±0.000**	0.281±0.011	0.075±0.003	0.068±0.002	0.079±0.004	0.153±0.007
(d) Andes network								
25%	**0.001±0.000**	0.072±0.006	**0.001±0.000**	0.112±0.010	0.077±0.007	0.074±0.006	0.068±0.005	0.145±0.012
50%	**0.001±0.000**	0.072±0.006	**0.001±0.000**	0.172±0.017	0.081±0.007	0.076±0.006	0.068±0.006	0.149±0.013
75%	**0.001±0.000**	0.073±0.007	**0.001±0.000**	0.257±0.028	0.088±0.008	0.081±0.007	0.069±0.007	0.156±0.007
100%	**0.001±0.000**	0.072±0.006	**0.001±0.000**	0.358±0.035	0.095±0.008	0.086±0.007	0.068±0.006	0.162±0.014

Table 5: Running time (seconds) of different algorithms under different constraint sizes.

Experiment analysis: (1) In most cases, QMAP-C performed better on learning accuracy than the other parameter learning algorithms, except for CME on the Andes and Win95pts networks when very limited constraints were available. Compared with the QMAP method, the QMAP-C method performed better on learning accuracy, especially when more constraints were available. The explanation is that, with increasing constraints, the purely data-driven estimation $\frac{N_{ijk}}{N_{ij}}$ in the QMAP estimation is more likely to violate the parameter constraints and that makes the overall QMAP estimation inaccurate. (2) The QMAP-C was less efficient than the QMAP method, especially when the number of constraints increased.

6 Conclusion

The advanced BN parameter learning algorithm, QMAP, fails to satisfy all of the parameter constraints, especially when insufficient data is available. In this paper, we present a modified QMAP algorithm, namely the QMAP-C algorithm. The main improvement of the proposed algorithm is that the learnt parameters satisfy all the convex parameter constraints under any cases. Because of that feature:

(1) When the provided parameter constraints are correct, the proposed QMAP-C algorithm outperforms the QMAP algorithm;

(2) When the provided parameter constraints are incorrect, the proposed QMAP-C algorithm would perform worse than the QMAP algorithm, especially when the number of incorrect constraints increases. The explanation is that, when the incorrect parameter constraints are incorporated, the regulation in the QMAP-C algorithm would change the data-driven estimation into a further inaccurate estimation and that would negatively influence the overall QMAP-C estimation.

Based on the above conclusions, before applying the QMAP-C algorithm to learn BN parameters, it is recommended to verify the correctness of the parameter constraints or domain knowledge. Future studies will focus on improving the constraint generality of the proposed method. Specifically, when non-convex parameter constraints are imposed, QMAP-C estimation could be further adjusted by methods such as isotonic regression.

Acknowledgments

We would like to sincerely thank the anonymous reviewers for the very helpful comments that significantly improve the manuscript. Also, we acknowledge the support for this study from the National Natural Science Foundation of China (61573285).

Appendix

Axiom *All the convex parameter constraints for a certain parameter θ_{ijk} can be finally transformed into an interval constraint $\theta_{ijk} \in [\theta^{L}_{ijk}, \theta^{U}_{ijk}]$.*

Theorem 1 *The QMAP estimation does not guarantee the satisfaction of all the convex parameter constraints.*

Proof. For the QMAP estimation (Eq.(6)) to satisfy the known constraint, it is expected to be

$$\theta^L_{ijk} \leq \frac{N_{ijk} + M_{ijk}}{N_{ij} + M_{ij}} \leq \theta^U_{ijk}. \tag{12}$$

Thus, it requires

$$(\theta^L_{ijk} N_{ij} + \theta^L_{ijk} M_{ij} - M_{ijk}) \leq N_{ijk} \leq (\theta^U_{ijk} N_{ij} + \theta^U_{ijk} M_{ij} - M_{ijk}). \tag{13}$$

However, for a data set of any size, the number of observations in the data set, N_{ijk}, could take any value larger or equal to zero, i.e., $N_{ijk} \geq 0$. Therefore, when $N_{ijk} < (\theta^L_{ijk} N_{ij} + \theta^L_{ijk} M_{ij} - M_{ijk})$ or $N_{ijk} > (\theta^U_{ijk} N_{ij} + \theta^U_{ijk} M_{ij} - M_{ijk})$, the QMAP estimation violates the parameter constraint. ■

Theorem 2 *The QMAP-C estimation guarantees the satisfaction of all the convex parameter constraints.*

Proof. The estimation θ'_{ijk} derived from the constrained likelihood model (by Eqs.(9-10)) certainly satisfies all convex the parameter constraints. Therefore,

$$\theta^L_{ijk} \leq \frac{N'_{ijk}}{N_{ij}} = \frac{N_{ij}\theta'_{ijk}}{N_{ij}} \leq \theta^U_{ijk}, \tag{14}$$

which means that $\theta^L_{ijk} N_{ij} \leq N'_{ijk} \leq \theta^U_{ijk} N_{ij}$. Furthermore, the mean value of the parameters sampled from the constraints, $P(X_i = k, \Pi_i = j|\Omega)$, also satisfies all the convex parameter constraints. Therefore,

$$\theta^L_{ijk} \leq \frac{M_{ijk}}{M_{ij}} = \frac{A \cdot P(X_i = k, \Pi_i = j|\Omega)}{A \sum\limits_{k=1}^{r_i} (P(X_i = k, \Pi_i = j|\Omega))} \leq \theta^U_{ijk}, \tag{15}$$

which implies that $\theta^L_{ijk} M_{ij} \leq M_{ijk} \leq \theta^U_{ijk} M_{ij}$. Finally, we can derive $\theta^L_{ijk}(N_{ij} + M_{ij}) \leq (N'_{ijk} + M_{ijk}) \leq \theta^U_{ijk}(N_{ij} + M_{ij})$, which is equivalent to

$$\theta^L_{ijk} \leq \frac{N'_{ijk} + M_{ijk}}{N_{ij} + M_{ij}} \leq \theta^U_{ijk}. \tag{16}$$

Hence, the QMAP-C estimation satisfies all the convex parameter constraints. ■

References

Russell G Almond, Duanli Yan, David M Williamson, Robert J Mislevy, and Linda S Steinberg. Bayesian networks in educational assessment. *Statistics for Social & Behavioral Sciences*, 473:95–103, 2015.

Eric E. Altendorf, Angelo C. Restificar, and Thomas G. Dietterich. Learning from sparse data by exploiting monotonicity constraints. In *Proceedings of the Twenty First Conference on Uncertainty in Artificial Intelligence*, pages 18–26, 2005.

Cassio P. De Campos and Qiang Ji. Improving bayesian network parameter learning using constraints. In *International Conference on Pattern Recognition*, pages 1–4, 2008.

Cassio P. De Campos, Yan Tong, and Qiang Ji. Constrained maximum likelihood learning of bayesian networks for facial action recognition. In *Proceeding of the 10th European Conference on Computer Vision*, pages 168–181, 2008.

Rui Chang and Wei Wang. Novel algorithm for bayesian network parameter learning with informative prior constraints. In *International Joint Conference on Neural Networks*, pages 1–8, 2010.

Anthony Costa Constantinou, Mark Freestone, William Marsh, Norman Fenton, and Jeremy Coid. Risk assessment and risk management of violent reoffending among prisoners. *Expert Systems with Applications*, 42(21):7511–7529, 2015.

G. F Cooper. Nestor: A computer-based medical diagnostic aid that integrates causal and probabilistic knowledge. 1984.

R. H. Di, X. G. Gao, and Z. G. Guo. Discrete bayesian network parameter learning based on monotonic constraint. *Systems Engineering & Electronics*, 36(2):272–277, 2014.

Ad Feelders and Linda C. Van Der Gaag. Learning bayesian network parameters under order constraints. *International Journal of Approximate Reasoning*, 42(1):37–53, 2005.

Eveline M. Helsper, Linda C. Van Der Gaag, and Floris Groenendaal. Designing a procedure for the acquisition of probability constraints for bayesian networks. In *Proceeding of the Fouteenth Conference on Engineering Knowledge in the Age of the Semantic Web*, pages 280–292, 2004.

Ji Lei Hu, Xiao Wei Tang, and Jiang Nan Qiu. A bayesian network approach for predicting seismic liquefaction based on interpretive structural modeling. *Georisk: Assessment and Management of Risk for Engineered Systems and Geohazards*, 9(3):200–217, 2015.

Walid Ibrahim and Valeriu Beiu. Using bayesian networks to accurately calculate the reliability of complementary metal oxide semiconductor gates. *IEEE Transactions on Reliability*, 60(3):538–549, 2011.

Guillaume Infantes, Malik Ghallab, and Flix Ingrand. Learning the behavior model of a robot. *Autonomous Robots*, 30(2):157–177, 2011.

S. Kullback and R. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.

Dries Landuyt, Steven Broekx, Rob D'Hondt, Guy Engelen, Joris Aertsens, and Peter L. M. Goethals. A review of bayesian belief networks in ecosystem service modelling. *Environmental Modelling & Software*, 46(7):1–11, 2013.

Steven Mascaro, Ann E. Nicholso, and Kevin B. Korb. Anomaly detection in vessel tracks using bayesian networks. *International Journal of Approximate Reasoning*, 55(1):84–98, 2014.

Judea Pearl. *Probabilistic Reasoning in intelligent systems : Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc, 1988.

Richard A Redner and Homer F Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26(2):195–239, 1984.

F. L. Seixas, B Zadrozny, J Laks, A Conci, and D. C. Muchaluat Saade. A bayesian network decision model for supporting the diagnosis of dementia, alzheimer?s disease and mild cognitive impairment. *Computers in Biology & Medicine*, 51(7):140–158, 2014.

Isozki Takashi, Kato Noriji, and Ueno Maomi. ”data temprature” in minimum free energies for parameter learning of bayesian networks. *International Journal on Artificial Intelligence Tools*, 18(5):653–671, 2009.

Yoshinori Tamada, Seiya Imoto, Hiromitsu Araki, Masao Nagasaki, Cristin Print, D. Stephen Charnock-Jones, and Satoru Miyano. Estimating genome-wide gene networks using nonparametric bayesian network models on massively parallel computers. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, 8(3):683–697, 2010.

Neil Wachowski and Mahmood R. Azimi-Sadjadi. Detection and classification of nonstationary transient signals using sparse approximations and bayesian networks. *IEEE/ACM Transactions on Audio Speech & Language Processing*, 22(12):1750–1764, 2014.

Peter Walley. Inference from multinomial data: Learning about a bag of marbles. *Journal of the Royal Statistical Society*, 58(1):3–57, 1996.

M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 44(3):257–303, 1990.

Frank Wittig and Anthony Jameson. Exploiting qualitative knowledge in the learning of conditional probabilities of bayesian networks. In *Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence*, pages 644–652, 2000.

Yun Zhou, Norman Fenton, and Martin Neil. Bayesian network approach to multinomial parameter learning using data and expert judgments. *International Journal of Approximate Reasoning*, 55(5):1252–1268, 2014.

Yun Zhou, Norman Fenton, and Cheng Zhu. An empirical study of bayesian network parameter learning with monotonic influence constraints. *Decision Support Systems*, 87: 69–79, 2016.