4

Advanced algorithms of bayesian network learning and inference from inconsistent prior knowledge and sparse data with applications in computational biology and computer vision

Rui Chang
University of California, San Diego
USA

1. Introduction

Bayesian networks are a popular class of graphical probabilistic models for researches and applications in the field of Artificial Intelligence. Bayesian network are built on Bayes' theorem (16) and allow to represent a joint probability distribution over a set of variables in the network. In Bayesian probabilistic inference, the joint distribution over the set of variables in a Bayesian network can be used to calculate the probabilities of any configuration of these variables given fixed values of another set of variables, called observations or evidence. Bayesian networks have been widely used for efficient probabilistic inference and data mining in many fields, such as computational biology and computer vision (17; 18).

Before we can generate useful prediction and reasoning by Bayesian networks, it is required to construct these network models from any resources. Over decades, enormous algorithms have been proposed to construct (we use construct and model interchangeably in this chapter) these Bayesian networks. These methods can be roughly classified into two categories: i) top-down modeling methods and ii) reverse-engineering methods. Top-down modeling methods seek for direct solutions to Bayesian network structure and parameter assignments from any prior knowledge resources and domain experts. Currently, this class of methods usually recruits both probability elicitation procedures from domain experts (23) and quantitative knowledge engineering process to disclose the Bayesian network structure and parameters. The advantages of this type of methods are the direct assignment of the parameters and structures from domain knowledge and experts without computational complications. However, in most domains, these methods encounter practical obstacles due to the actual availability of quantitative information and to the limitation of an expert knowledge. In contrast, reverse-engineering approaches utilize machine learning algorithms to train (learn) Bayesian network structure and parameters from a collection of past observations. This process belongs to unsupervised learning in machine learning theory. The advantage of this class approaches is that, a training machine can automatically determine a best Bayesian network model with structure and parameters which optimally fits to the training data under the judgments of an object function or scoring function. (in stead of manually

evaluation in top-down methods). This score function is often the posterior probability function of a Bayesian network structure and parameters given the training data. The learned single best model is called Maximum-a-Posterior (MAP) estimation which is computed from data likelihood and prior distribution. In the last twenty to ten years, reverse-engineering approaches have become mainstream researches in the field of Bayesian network modeling. Fruitful results have been achieved, especially in the efficient learning of Bayesian network structure and parameters with (in-) complete data (4; 19–21).

However, a major problem of Bayesian network learning in most existing algorithms is the demands on a large amount of training samples to achieve good generalization performance. The generalization performance of a learned Bayesian network largely depends on the amount of training dataset and the quality of the prior provided to the learning process. Specially, if training data is scarce, it becomes crucial to use various forms of prior knowledge to improve the accuracy of learned models and avoid overfitting. Moreover, although the maximum a posteriori estimation, i.e., the selection of a single best Bayesian network model from the data by learning, is useful for the case of large data sets, independence assumptions among the network variables often make this single model vulnerable to overfitting. In realistic problems, the data basis is often very sparse and hardly sufficient to select one adequate model, i.e., there is considerable model uncertainty. In fact, selecting one single Bayesian model can then lead to strongly biased inference results. Therefore, it is preferable to adopt full Bayesian approaches, such as model averaging, to incorporate these model uncertainties.

2. Overview

2.1 Advanced Bayesian Network Modeling and Inference from Consistent and Inconsistent Prior Knowledge

As the first part of our methodology, we propose novel methods to make use of prior qualitative knowledge in a domain to construct Bayesian networks and generate quantitative probability predictions from these models. These algorithms stem from the observations that in many domains, enormous amounts of priori qualitative knowledge have been accumulated by original studies. This type of knowledge is often represented in terms of qualitative relational statements between two or more entities. For example, in biomedical domain, such a statement can be *smoking increases the risk of lung cancer*. In this statement, two entities are *smoking* and *lung cancer* and these two entities are connected to each other through a directed and functional relation: *increase*. The semantics encoded in this statement is: smoking positively influences lung cancer so that the probability and risk of lung cancer is increased under the condition of smoking. In genomics research, a common knowledge about biological molecular interactions would be a transcript factor binds to a gene and up-regulate this gene's expression level in a cell. In computer vision, qualitative statement can be among action units. For instance, "cheek raiser" tends to happen with "lip corner puller", when smiling. In this statement, cheek raiser increases the occurrence probability of lip corner puller. Similar qualitative statements can be found in many other domains, such as economy, politics, science and engineering indicating that our proposed methods have great promises in these fields. In fact, these inequality constraints have been proposed and used in qualitative probabilistic inference process, such as qualitative probabilistic network (25). However, due to the lacks of quantitative measurements in these qualitative knowledge and constraints, they have been ignored in the quantitative modeling of Bayesian networks.

In our top-down Bayesian inference method, we designed a knowledge model which captures the entities and their relationships in the statement. Various qualitative relations are mapped into mathematically meaningful constraints and inequalities over the Bayesian network structure and parameter space. Due to their qualitativeness, these constraints eventually define a prior distribution in the model space, i.e. model uncertainty. These constraints reduce the set of all possible Bayesian models to those which are consistent with the set of statements considered. This class of consistent models is used to perform full Bayesian inference which can be approximated by Monte Carlo methods, i.e. the quantitative inference and reasoning can be calculated in each of the Bayesian model and these quantitative results are averaged and weighted by the model posterior probability. This is even analytically tractable for smaller networks and statement sets.

Notably, qualitative knowledge are often inconsistent, i.e. there may exist contradicting qualitative statements on entities and/or their relations which eventually affect the model uncertainty in the constructed Bayesian network model space. Therefore, it is imperative to develop methods for reconciling inconsistent qualitative knowledge and for modeling Bayesian networks and performing quantitative prediction. To this end, we further propose a novel framework for performing quantitative Bayesian inference with model averaging based on the inconsistent qualitative statements as a coherent extension of framework of quantitative Bayesian inference based on a set of consistent hypotheses introduced above (33). Our method interprets the qualitative statements by a vector of knowledge features whose structure can be represented by a hierarchical Bayesian network. The prior probability for each qualitative knowledge component is calculated as the joint probability distribution over the features and can be decomposed into the production of the conditional probabilities of the knowledge features. These knowledge components define multiple Bayesian model classes in the hyperspace. Within each class, a set of constraints on the ground Bayesian model space can be generated. Therefore, the distribution of the ground model space can be decomposed into a set of weighted distributions determined by each model class. This framework is used to perform full Bayesian inference which can be approximated by Monte Carlo methods, but is analytically tractable for smaller networks and statement sets.

2.2 Related Works

In discrete model, qualitative causal knowledge have been utilized for abstract probabilistic graphical models, i.e. qualitative probabilistic network (QPN) (6) and reasoning algorithms in QPN have been proposed (5; 9). These algorithms perform qualitative inference with sign propagation in stead of quantitative predictions and neither inconsistent hypotheses could be dealt with.

2.3 Advanced Bayesian Network Learning with Integration of Prior Knowledge and Sparse data

As the second part of the methodology section, we introduce our latest algorithm developments in learning Bayesian network models. In this method, Bayesian network learning accuracy is drastically improved by integrating generic qualitative domain knowledge with training data. We use the knowledge model designed in section 3.1 to translate the causality in qualitative domain knowledge into a set of constraints over structural and parameter space. For parameter learning, we recruit a sampling approach to recover the prior belief distribu-

tion in parameter space out of the constraints. We then propose a novel Bayesian parameter score function which integrates this informative prior as soft regulation with the quantitative data statistics. In this way, the parameter posterior distribution is combinatorial regulated by both quantitative data and prior knowledge. In the conventional Bayesian network learning algorithm, MAP estimation usually employs Dirichlet priori to further regulate the statistical counts from the data. However, as discussed above, it is often impossible to determine the correct hyperparameters of this prior distribution which may result bias in the MAP estimation. Our algorithm resolves this issue by establishing an informative prior from domain qualitative knowledge. This informative prior provides the learning machine a correctly defined model subspace to seek for global maximum. By combining each possible prior pseudo counts in this subspace with data statistical counts, we can explore multiple local maximum estimates and determine the global maximum by model selection scheme. Thus, we avoid trapping in the local maximum. This method is particular useful in accurate learning of a Bayesian network under sparse training data. These algorithms can be naturally extended to BN structural learning which is under active developments.

2.4 Related Works

Researches have proposed a number of algorithms to learn Bayesian network parameters by utilizing various forms of prior knowledge, such as dirichlet function (28; 29). In (30–32), parameter learning schemes for various graphical models incorporating parameter sharing constraints are proposed. These algorithms provide efficient solutions for parameter learning with parameter sharing constraints, i.e. parameter equality in one multinomial conditional distribution. If a parameter satisfy the constraints, it obeys the dirichlet distribution with certain normalizer. Otherwise, the prior distribution is zero. A closed form normalization solution is derived in case of parameter sharing constraints. Moreover, some simple forms of inequalities within one conditional distribution are proposed (32). In this case, no closed-form solution is possible. Though, in (30–32), constrained parameter learning problem is treated as a constraint optimization problem and efficient algorithms are developed, the forms of the constraints are limited to either parameter sharing or inequality constraints within one conditional distribution, such as $P(A|B)>P(\overline{A}|B)$. More generic and important inequality constraints, such as $P(A|B)>P(A|\overline{B})$ is not addressed by their methods.

In (35) and (37), methods are proposed to deals with the inequality constraints in parameter learning. A penalty term is designed to regulate the likelihood which is derived from the monotonic influence with form of $P(A|B)>P(A|\overline{B})$. The violation term can only penalize the likelihood when the learned local maximum violates the constraints in the sign, but it can not distinguish a set of all possible local maximums obeying the constraints. So, final solution is not necessary a global maximum. (Eq.8 in (35) and Eq.9 in (37)). This is a serious problem in case of learning with very sparse data. In this case, although ML estimation may output an estimate obeying the sign of the constraints, this ML estimation is highly probable incorrect due to the amount of data. In this case, neither (35) nor (37) could use prior statistics to correct the estimation. As stated in (37), a soft Bayesian prior which regulates the ML term is desired. A similar iterative approach with penalty function was introduced in (36). The method in (42), however, includes constraints beyond the monotonicity constraints.

In (38), an averaging scheme is proposed. This method is only feasible up to 5/6 parents. (39) proposed a similar idea to ours independently. A method which uses a soft

Bayesian prior to regulate the ML score and introduce the concept of model uncertainty in the MAP estimation. The empirical Bayes and maximum posterior estimate in (39) and QMAP_{FBA},QMAP_{FMA} in my paper are comparable. However, (39) indirectly translates the prior knowledge into an intractable integration which has to be approximated. The dirichlet hyperparameters is replaced by another hyperparameter (Eq.14 in (39)). Their initial idea is to assign some confidence to constraints. (Eq.7 in (39)). But it may be easier and more efficient to handle this issue in the knowledge level than score level (34). Comparatively, we work directly on the parameter space through sampling and obtain the dirichlet hyperparameters directly. Thus, we believe our method can be more efficient and feasible than their method.

3. Methods

In this section, we formally propose our top-down Bayesian network modeling algorithm, i.e. Bayesian inference with only consistent and inconsistent qualitative prior knowledge. Next, we introduce our advanced Bayesian network learning algorithm by integrating both qualitative prior knowledge and data.

3.1 Probabilistic Representation of a Qualitative Knowledge Model

Several qualitative models have been proposed in the context of Qualitative Probabilistic Networks (QPN). Qualitative knowledge models describe the process of transforming the qualitative statements into a set of probability constraints. The proposed Bayesian inference method outlined above is independent of the qualitative knowledge model, i.e. the model posterior probability is independent of the set of qualitative statements used, once the set of probabilistic inequality constraints which are translated from qualitative statements is given. Three existing qualitative models are the Wellman approach (25), the Neufeld approach (22) and the orders of magnitude approach (27). Here we follow the Wellman approach, where qualitative knowledge involves influential effects from parent nodes to child nodes which are classified according to the number of inputs from parents to child and their synergy. For the sake of simplicity, we restrict our discussion to binary-valued nodes. Logic "1" and "0" values of a node are defined as "present" and "absent" or "active" and "inactive", as synonyms, A and $\overline{A}$. For multinomial nodes, similar definitions can be applied.

3.1.1 Structural Qualitative Knowledge Model

The qualitative knowledge contained in the statements are describing two aspects of a belief network, i.e. structure and parameter. The structural knowledge of a simple network consisting node B and node A can be described with two first-order logic predicates:

$$\begin{aligned} Depend(A,B) &= 0/1 \\ Influence(A,B) &= 0/1 \end{aligned} \tag{1}$$

which describe whether A and B are dependent and whether the influence direction is from A to B; *Depend* and *Influence* are denoted by Dp and I as well as, the set of structural knowledge features is denoted by Π={Dp,I}.

3.1.2 Parameter Qualitative Knowledge Model

Under each structure feature, we extend the QPN model with two sets of dependent features, i.e. baseline qualitative knowledge features, Σ and extended qualitative knowledge features, Ψ. These two feature sets are used to describe the qualitative parameter knowledge.

3.1.2.1 Baseline Qualitative Knowledge Model

In QPN, a set of features define the basic properties of qualitative causal influences and their synergy classified by the number of inputs from parents to child which are refined in this paper and are referred to as *Baseline Qualitative Knowledge Model*. Baseline features transform qualitative statements into a primitive set of constraints on model parameter space. We discuss three cases of influences, namely single influence, joint influence and mixed joint influence. In addition, we discussed the qualitative influence derived from recurrent and/or conflicting statements. The definitions of the influences in our work are originated and refined based on the qualitative probabilistic network in (25) which enables us to translate the qualitative statements into a set of constraints in the parameter space which can be used to model the parameter distribution given the structure.

I. Single Influence

Definition 3.1 If a child node B has a parent node A and the parent imposes a isolated influence on the child, then qualitative influence between parent and child is referred to as *single influence*. Single influence can be further classified into single positive influence and single negative influence.

Definition 3.2 If presence of parent node A renders presence of child node B more likely, then the parent node is said to have a *single positive influence* on the child node. This can be represented by the inequality

$$Pr(B|A) \geq Pr(B|\overline{A}) \tag{2}$$

Definition 3.3 If presence of parent node A renders presence of child node B less likely, then parent node is said to have a *single negative influence* on child node. This can be represented by the inequality

$$Pr(B|A) \leq Pr(B|\overline{A}) \tag{3}$$

II. Joint Influence

Definition 3.4 If a child node B has more than one parent node and all parents influence the child in a joint way, then these influences between parents and child are referred to as *joint influence*. This joint influence can be either synergic (cooperative) or antagonistic (competitive) and the individual influences from the parents to the child can be either positive or negative.

Definition 3.5 If a joint influence from two or more parent nodes generates a combined influential effect larger than the single effect from each individual parent, then the joint influence is referred to as *plain synergic joint influence* or *plain synergy*.

Assume that parent nodes A and B impose positive individual influences on child node C, then the knowledge model can be defined as

$$Pr(C|A,B) \geq \left\{ \begin{array}{l} Pr(C|A,\overline{B}) \\ Pr(C|\overline{A},B) \end{array} \right\} \geq Pr(C|\overline{A},\overline{B}) \tag{4}$$

Definition 3.6 If joint influences from two or more parent nodes generate an combined influential effect larger than the sum of each single effect from an individual parent, then the joint influence is referred to as *additive synergic joint influence* or *additive synergy*.(24)

Assume in case that parent nodes A and B impose a positive individual influence on child node C, then we define

$$Pr(C|A,B) \geq Pr(C|A,\overline{B}) + Pr(C|\overline{A},B) \geq \left\{ \begin{array}{l} Pr(C|A,\overline{B}) \\ Pr(C|\overline{A},B) \end{array} \right\} \geq Pr(C|\overline{A},\overline{B}) \tag{5}$$

Similar rules can be applied to the case where A and B impose a negative individual influence on child node C. Comparing Eq. 5 with Eq. 4, we can conclude that *additive synergy* is a sufficient condition for *plain synergy* and *plain synergy* is a necessary but not sufficient condition for *additive synergy*. Therefore, if multiple parents demonstrate additive synergy, it is sufficient to judge that this influence is also plain synergy, but not vice-versa.
It is important to distinguish between plain synergy and additive synergy since they represent distinct semantic scenarios in a domain. For example, A is a protein and B is a kinase which phosphorylates protein A and produces the phosphorylated protein C. Because of the nature of this protein-protein interaction, neither B nor A alone can significantly increase the presence of C, but both together can drastically increase the presence of C which is greater than the sum of C in case of either A or B present. In this example A and B exhibit additive synergy and it is sufficiently to conclude that A and B has plain synergy as well.
Definition 3.7 If the joint influences from two or more parent nodes generate a combined influential effect less than the single effect from individual parent, then the joint influence is referred to as *antagonistic joint influence* or *antagonism*.
Assume that parent nodes A and B have independent positive single influences on child node C, the antagonistic influence of A and B can be represented by

$$Pr(C|\overline{A},\overline{B}) \leq Pr(C|A,B) \leq \left\{ \begin{array}{c} Pr(C|A,\overline{B}) \\ Pr(C|\overline{A},B) \end{array} \right\} \tag{6}$$

Similar rules can be applied to the case where A and B imposes a negative individual influence on child node C.
III. Mixed Joint Influence
In case that the joint effect on a child is formed by a mixture of positive and negative individual influences from its parents, the extraction of a probability model is not well-defined in general. Hence, we adopt the following scheme: If there are mixed influences from several parent nodes to a child node, and no additional information is given, then they are treated as independent and with equal influential strength. Assume that parent node A imposes positive single influence on child node C and parent node B imposes negative single influence on child node C, then the joint influence can be represented by

$$\begin{gathered} Pr(C|A,B) \geq Pr(C|\overline{A},B);\ Pr(C|A,\overline{B}) \geq Pr(C|\overline{A},\overline{B}); \\ Pr(C|A,\overline{B}) \geq Pr(C|A,B);\ Pr(C|\overline{A},\overline{B}) \geq Pr(C|\overline{A},B) \end{gathered} \tag{7}$$

Any additional structure can be brought into the CPT of the corresponding collider structure as soon as dependencies between influences are made explicit by further qualitative statements.

3.1.3 Extended Qualitative Knowledge Model

The extended qualitative knowledge model defines relative and absolute properties of probability configurations in qualitative causal influences and synergy from the baseline model. It includes the probabilistic ratio and relative difference between any number of configurations in a qualitative causal influence and the absolute probabilistic bound of any configuration in a causal influence. These extended features impose further restriction on the set of constraints generated by baseline model, therefore, restrain the uncertainty in Bayesian model space so that more accurate generalization can be achieved.

The extended qualitative knowledge features can be consistently represented by a linear inequality. In the case that node B impose single influence on node A, there are two probabilistic configurations. The linear constraints can then be written as

$$Pr(B|A) \geq, \leq R \times Pr(B|\overline{A}) + \Delta; Pr(B|A) \in [Bd_{min}, Bd_{max}]; Pr(B|\overline{A}) \in [Bd'_{min}, Bd'_{max}] \quad (8)$$

which R is *Influence Ratio*, Δ is *Influence Difference* and Bd, Bd' denote *bound*. In some cases, baseline and extended qualitative knowledge information are provided by the qualitative statements simultaneously. However, in most cases, extended knowledge features are not fully provided in the qualitative statements. In these cases, only baseline knowledge model will be used to generate constraints in model space to perform inference by model averaging. Once the qualitative knowledge is translated by the feature set $\{\Pi(Dp, I), \Lambda(\Sigma, \Psi(R, \Delta, Bd))\}$ according to Eq. 1 to Eq. 8, the distribution of ground models is defined by this knowledge. Once formulated, the Monte Carlo sampling procedure will make sure that all inequalities are satisfied for valid models.

3.1.4 Hierarchical Knowledge Model for Inconsistent Statements

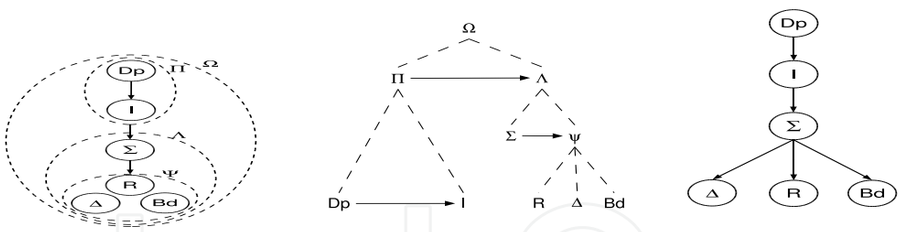
The dependent qualitative knowledge feature set can be represented by a hierarchical Bayesian network (HBN) (3). Within a knowledge HBN, the structural feature Π and parameter feature Λ are two first-level composite nodes. Π can be further decomposed into two leaf nodes Dp and I. The parameter feature Λ contains two second-level composite nodes, i.e. the baseline knowledge features Σ and extended knowledge features Ψ which consists of three leaf nodes R, Δ and Bd. Thus qualitative knowledge Ω can be described as $\Omega = \{\Pi(Dp, I), \Lambda(\Sigma, \Psi(R, \Delta, Bd))\}$, where $\Sigma = (SP, SN, PlSyn, AdSyn, Ant, MxSyn)$. The hierarchical knowledge model is shown in Figure 1(a) and a tree hierarchy in Figure 1(b). The equivalent Bayesian network is shown in Figure 1(c).

Hierarchical Bayesian Networks encode conditional probability dependencies in the same way as standard Bayesian Networks. The prior probability of a qualitative knowledge Ω can be written as a joint probability of $\{\Pi, \Lambda\}$ and can be decomposed according to the dependency between each component features as follows.

$$Pr(\Omega) \quad = \quad Pr(\Pi)Pr(\Sigma|\Pi)Pr(\Psi|\Sigma) \quad (9)$$

where $Pr(\Psi|\Sigma) = Pr(R|\Sigma)Pr(\Delta|\Sigma)Pr(Bd|\Sigma)$, $Pr(\Pi) = Pr(Dp)Pr(I|Dp)$ and $Pr(\Sigma|\Pi) = Pr(\Sigma|I)$. The conditional probabilities of qualitative knowledge features can be calculated by counting the weighted occurrences given a set of inconsistent statements. The weight of knowledge features equals to the credibility of their knowledge sources which may be evaluated by a domain expert or determined by the source *impact factor*. If no further information on the weights is available, they are set to 1. In this case, the conditional probability of features is computed only by occurrence count. For example, we assume a set of qualitative statements, $\widetilde{S} = \{S_1, S_2, S_3\}$, about *smoking* and *lung cancer* are observed: 1) *The risk is more than 10 times greater for smokers to get lung cancer than no-smokers.* 2) *Men who smoke two packs a day increase their risk more than 25 times compared with non-smokers.* 3) *There is not significant evidence to prove that smoking directly cause lung cancer, however, clinical data suggest that lung cancer is related to smoking*. The statements can be represented by a vector of features which is shown in Figure 2. The conditional probability of the features can be calculated straightforwardly by

$$\begin{array}{cc} Pr(I|Dp) = (w_1 + w_2)/w_a & Pr(\overline{I}|Dp) = (w_3)/w_a \\ Pr(r_1|\Sigma = SP) = w_1/w_b & Pr(r_2|\Sigma = SP) = (w_1 + w_2)/w_b \end{array}$$



(a) HBN (b) Tree (c) BN

Fig. 1. Hierarchical Bayesian Network on Qualitative Knowledge

Stat.	Dp	I	Σ	R	Δ	B	Weight
S_1	1	1	SP	[10,∞]	null	null	w_1
S_2	1	1	SP	[25,∞]	null	null	w_2
S_3	1	0	null	null	null	null	w_3

Fig. 2. Feature-vector of Statements

where $w_a = w_1 + w_2 + w_3$, $w_b = 2w_1 + w_2$, $Pr(Dp) = 1$, $Pr(SP|I) = 1$, $r_1 = [10, 25]$ and $r_2 = [25, \infty]$. One notion is that the knowledge features $\Psi = \{R, \Delta, Bd\}$ in Figure 1(a) are continuous-valued and therefore, can be transformed to discrete attributes by dynamically defining new discrete attributes that partition the continuous feature value into a discrete set of intervals. In the above example, the continuous feature R in S_1 has value range $[10, \infty]$ and a continuous value range $[25, \infty]$ in S_2. The continuous ranges can be partitioned into two discrete intervals: $r_1 = [10, 25]$ and $r_2 = [25, \infty]$, therefore, the qualitative knowledge $\widetilde{\Omega} = \{\Omega_1, \Omega_2, \Omega_3\}$ can be transformed from $\widetilde{S} = \{S_1, S_2, S_3\}$ with discrete-valued features.

3.1.4.1 Qualitative Knowledge Integration

Once we have calculated the conditional probabilities of knowledge features, the prior probability of qualitative knowledge can be computed according to Eq. 9. Thus the inconsistent knowledge components are ready to be reconciled. The qualitative knowledge transformed from the feature vector of statements in Figure 2 can be described by $\widetilde{\Omega}$:

$$\Omega_1 = \{1, 1, SP, [10, 25], \emptyset, \emptyset\} \quad \Omega_2 = \{1, 1, SP, [25, \infty], \emptyset, \emptyset\} \quad \Omega_3 = \{1, 0, \emptyset, \emptyset, \emptyset, \emptyset\} \tag{10}$$

where Ω_k=$\{Dp_k, I_k, \Sigma_k, R_k, \Delta_k, Bd_k\}$. If the weights of statements are set to 1, the knowledge prior probability is calculated, then we have $Pr(\Omega_1)$=2/9, $Pr(\Omega_2)$=4/9 and $Pr(\Omega_3)$=1/3.

$$\begin{aligned} Pr(\Omega_1) &= Pr(Dp)Pr(I|Dp)Pr(SP|I)Pr(r_1|SP) = 2/9 \\ Pr(\Omega_2) &= Pr(Dp)Pr(I|Dp)Pr(SP|I)Pr(r_2|SP) = 4/9 \\ Pr(\Omega_3) &= Pr(Dp)Pr(\bar{I}|Dp) = 1/3 \end{aligned} \tag{11}$$

The integrated qualitative knowledge thus preserved the uncertainty from each knowledge component. Each qualitative knowledge component Ω_k defines a model class with a set of constraints on the ground model space which is generated by its features. The model class and its constraints are used for modeling Bayesian networks and performing quantitative inference.

3.2 Bayesian Inference with Consistent Qualitative Knowledge

3.2.1 Bayesian Modeling and Inference

A Bayesian model m represents the joint probability distribution of a set of variables $\mathbf{X} = X_1, X_2, ..., X_D$ (19). The model is defined by a graph structure s, which defines the structures of the conditional probabilities between variables, and a parameter vector θ, the components of which define the entries of the corresponding conditional probability tables (CPTs). Hence, a Bayesian network can be written as $m = \{s, \theta\}$. If we believe that one single model m reflects the true underlying distribution, we can perform inference based on this model. Given some observations or "evidence" E, reflected by fixed measured values of a subset of variables, $\mathbf{X}_q = E$, we wish to derive the distribution of the remaining variables $X \in \mathbf{X} \backslash \mathbf{X}_q$. It is provided by their conditional probability given the evidence in light of the model, $Pr(X|E, m)$, which can be efficiently evaluated by known methods.(26)

In contrast, the full Bayesian framework does not attempt to approximate one true underlying distribution. Instead, all available information is used in an optimal way to perform inference, without taking one single model for granted. To formalize this statement for our purposes, let us classify the set of available information into an available set of data, D, and a body of non-numeric knowledge, Ω. The a posteriori distribution of models m is then given by

$$Pr(m|D, \Omega) = \frac{Pr(D|m)\ Pr(m|\Omega)}{Pr(D, \Omega)}. \tag{12}$$

The first term in the numerator of eq. (12) is the likelihood of the data given the model, which is not directly affected by non-numeric knowledge Ω, the second term denotes the model prior, whose task is to reflect the background knowledge. We obtain

$$Pr(m|D, \Omega) = \frac{1}{Z} Pr(D|m)\ Pr(m|\Omega), \tag{13}$$

where Z is a normalization factor which will be omitted from the equations for simplicity. The first term contains the constraints of the model space by the data, and the second term the constraints imposed by the background knowledge. In the full Bayesian approach, we can perform inference by model averaging. Now, given some observation or evidence E, the (averaged) conditional distribution of the remaining variable X is performed by integrating over the models:

$$Pr(X|E, D, \Omega) = \int Pr(X|E, m) Pr(m|D, \Omega) dm = \int Pr(X|E, m) Pr(D|m) Pr(m|\Omega) dm \tag{14}$$

3.2.2 Bayesian Network Inference with Qualitative Knowledge

In this paper we consider the extreme case of no available quantitative data, $D = \emptyset$. Even in this case, it is still possible to perform proper Bayesian inference,

$$Pr(X|E, \Omega) = \int Pr(X|E, m) Pr(m|\Omega) dm. \tag{15}$$

Now the inference is based on the general background information contained in Ω alone, and the specific information provided by the measurements E. This is reflected by the fact that inference results are conditioned on both quantities in eq. (15).

In order to determine $Pr(m|\Omega)$, we need a formalism to translate a body of qualitative knowledge into an a priori distribution over Bayesian models. For this we adopt the following notation for a Bayesian model class. A Bayesian model is determined by a graph structure s and

by the parameter vector θ needed to specify the conditional probability distributions given that structure. We refer to θ as one specific CPT configuration. A Bayesian model class $\widetilde{M}$ is then given by (i) a discrete set of model structures $\widetilde{S} = \{s_1, s_2, \ldots, s_K\}$, and (ii) for each structure s_k a (eventually continuous) set of CPT configurations Θ_k. The set of member Bayesian models $m \in \widetilde{M}$ of that class is then given by $m = \{(s_k, \theta)|k \in \{1, \ldots, K\}, \theta \in \Theta_k\}$. The model distribution now reads

$$Pr(m|\Omega) = Pr(s_k, \theta|\Omega) = \frac{Pr(\theta|s_k, \Omega)Pr(s_k|\Omega)}{\sum_{a=1}^{K} \int_{\Theta_a} Pr(\theta|s_a, \Omega)d\theta Pr(s_a|\Omega)}. \tag{16}$$

In eq. (16), first the set of allowed structures is determined by means of Ω, followed by the distributions of the corresponding CPT configurations. Then, we calculate the model's posterior probability $Pr(m|\Omega)$ in eq. 16. Inference is carried out by integrating over the structure space and the structure-dependent parameter space:

$$Pr(X|E, \Omega) = \sum_{k=1}^{K} \int_{\Theta_k} Pr(X|E, s_k, \theta)Pr(s_k, \theta|\Omega)d\theta. \tag{17}$$

It is very common to express non-numeric knowledge in terms of qualitative statements about a relationship between entities. Here we assume Ω to be represented as a list of such qualitative statements. In this form, the information can be used in a convenient way to determine the model prior, eq. (16): (i) Each entity which is referenced in at least one statement throughout the list is assigned to one variable X_i. (ii) Each relationship between a pair of variables constrains the likelihood of an edge between these variables being present. (iii) The quality of that statement (e.g., "activates", "inactivates") affects the distribution over CPT entries θ given the structures. In the most general case, the statement can be used to shape the joint distribution over the class of all possible Bayesian models over the set of variables obtained from Ω.

Here we propose a simplified but easy-to-handle way for constructing the prior model distribution. We use each statement to constrain the model space to that subspace which is consistent with that statement. In other words, if a statement describes a relationship between two variables, only structures s_k which contain the corresponding edge are assigned a nonzero probability $Pr(s_k|\Omega)$. Likewise, only parameter values on that structure, which are consistent with the contents of that statement, are assigned a nonzero probability $Pr(\theta|s_k, \Omega)$. If no further information is available, the distribution is constant in the space of consistent models.

3.3 Bayesian Inference with Inconsistent Qualitative Knowledge

In this section, we propose a novel approach to make use of a set of inconsistent qualitative statements and their prior belief distribution as background knowledge for Bayesian modeling and quantitative inference.

A Bayesian model m represents the joint probability distribution of a set of variables $X = \{x_1, x_2, ..., x_N\}$ (1). The model is defined by a graph structure s and a parameter vector θ, i.e. $m = \{s, \theta\}$. In full Bayesian framework, all available information is used in an optimal way to perform inference by taking model uncertainty into account. Let us classify the set of available information into an available set of training data D and a set of inconsistent qualitative background knowledge $\widetilde{\Omega} = \{\Omega_1, \ldots, \Omega_K\}$ on a constant set of variables. The posterior

distribution of models m is then given by

$$Pr(m|D,\widetilde{\Omega}) = \frac{Pr(D|m,\widetilde{\Omega})Pr(m|\widetilde{\Omega})Pr(\widetilde{\Omega})}{Pr(D,\widetilde{\Omega})} \tag{18}$$

The first term in the numerator of Eq. 18 is the likelihood of the data given the model. The second term denotes the model prior which reflects the inconsistent set of background knowledge and the last term is the prior belief of the knowledge set. Now, inference in the presence of evidence is performed by building the expectation across models:

$$Pr(X|D,E,\widetilde{\Omega}) = \int dm Pr(X|E,m)Pr(D|m,\widetilde{\Omega})Pr(m|\widetilde{\Omega})Pr(\widetilde{\Omega}) \tag{19}$$

In this paper we consider the extreme case of no available quantitative data, $D = \emptyset$.

$$Pr(X|E,\widetilde{\Omega}) = \int dm Pr(X|E,m)Pr(m|\widetilde{\Omega})Pr(\widetilde{\Omega}) \tag{20}$$

In this case, model prior distribution $Pr(m|\widetilde{\Omega})$ is determined soly by the inconsistent background knowledge set $\widetilde{\Omega}$. Each independent qualitative knowledge component, $\Omega_k \in \widetilde{\Omega}$, uniquely defines a model class, M_k, with a vector of features, i.e. $\widetilde{M} = \{M_1, \ldots, M_K\}$. The features are translated into a set of constraints which determine the distribution of the ground models within each model class.
First of all, the probability of a model class given the inconsistent knowledge set is written as

$$Pr(M_k|\widetilde{\Omega}) = \sum_{i=1}^{K} Pr(M_k|\Omega_i)Pr(\Omega_i|\widetilde{\Omega}) = Pr(\Omega_k) \tag{21}$$

where $\{Pr(M_k|\Omega_i) = 1, i = k\}$ and $\{Pr(M_k|\Omega_i) = 0, i \neq k\}$ since the *k-th* model class is uniquely defined by Ω_k and is independent to the other knowledge component. Secondly, the probability of a ground Bayesian model sample m in the *k-th* model class given the inconsistent knowledge set is

$$Pr(m \in M_k|\widetilde{\Omega}) = Pr(m|M_k)Pr(M_k|\widetilde{\Omega}) \tag{22}$$

Thus, the inference on X given evidence E and inconsistent knowledge set $\widetilde{\Omega}$ in Eq. 20 can be written as

$$Pr(X|E,\widetilde{\Omega}) = \sum_k \int_m dm Pr(X|m,E)Pr(m|M_k)Pr(\Omega_k)$$

where $Pr(m|\widetilde{\Omega}) = \sum_k Pr(m \in M_k|\widetilde{\Omega})$ and we assume the inconsistent knowledge set to be true, i.e. $Pr(\widetilde{\Omega}) = 1$. Therefore, the inference is calculated by firstly integrating over the structure space and the structure-dependent parameter space of a ground Bayesian model from a model class according to the constraints and performing such integration iteratively over all possible model classes with the prior distribution. The integration in Eq. 23 is non-trivial to compute, however, Monte Carlo methods can be used to approximate the inference.

3.3.1 ASIA Benchmark Model

The ASIA network (10) is a popular toy belief model for testing Bayesian algorithms. The structure and parameter of actual ASIA network is shown in Figure 3.
For demonstration, we consider the inconsistent qualitative statements with regarding to single edge between *Smoking* and *Lung Cancer*, as well as the collider structure of *Lung Cancer*,

Bronchitis and *Dyspnea*. The method applies to all of the entities and their relations in the ASIA network. *1. Although nonsmokers can get lung cancer, the risk is about 10 times greater for smokers. (http://www.netdoctor.co.uk);2. The lifetime risk of developing lung cancer in smokers is approximately 10%.(http://www.chestx-ray.com/Smoke/Smoke.html);3. Men who smoke two packs a day increase their risk more than 25 times compared with non-smokers.(http://www.quit-smoking-stop.com/lung-cancer.html)4. Lifetime smoker has a lung cancer risk 20 to 30 times that of a non-smoker(http://www.cdc.gov/genomics/hugenet/ejournal/OGGSmoke.htm)5. Only 15% of smokers ultimately develop lung cancer(http://www.cdc.gov/genomics/hugenet/ejournal/OGGSmoke.htm);6. The mechanisms of cancer are not known. It is NOT possible to conclusively attribute a cause to effects whose mechanisms are not fully understood.(http://www.forces.org/evidence/evid/lung.htm);7. It is estimated that 60% of lung cancer patients have some dyspnea at the time of diagnosis rising to 90% prior to death.(http://www.lungcancer.org/health_care/focus_on_ic/ symptom/dyspnea.htm)8. Muers et al. noted that breathlessness was a complaint at presentation in 60% of 289 patients with non-small-cell lung cancer. Just prior to death nearly 90% of these patients experienced dyspnea. (2);9. At least 60% of stage 4 lung cancer victims report dyspnea.(http://www.lungdiseasefocus.com/lung-cancer/palliative-care.php);10. Significantly more patients with CLD than LC experienced breathlessness in the final year (94% CLD vs 78% LC, $P < 0.001$) and final week (91% CLD vs 69% LC, $P < 0.001$) of life. (7);11. 95% of patients with chronic bronchitis and emphysema reported Dyspnea. (8)*
Each statement is analyzed by the hierarchical knowledge model in Figure 1(a) and the extracted features are summarized in Figure 3(c). In this statement set, the first six statements represent the relation between (tobacco)smoking and lung cancer. $\{S_1, \ldots, S_5\}$ describe a *single positive (SP)* influence from smoking to lung cancer with inconsistent knowledge features of the *ratio (R)* and *bound (Bd)*. However, statement S_6 declares a contradicting knowledge suggesting that smoking is not the cause of lung cancer. $\{S_7, \ldots, S_{11}\}$ describe the synergic influence from lung cancer and bronchitis to dyspnea. Without further information, it can be represented by *plain synergy with positive individual influence*. The knowledge on the extended features in Eq. 7 of the conditional probability distribution of this collider structure is not available, however, the knowledge on the extended features of the marginalized conditional probability space are provided in these statements. For simplicity, we assume the weight of every qualitative statement equals to 1, i.e. $\{w_i = 1, i = 1, \ldots, 11\}$. Due to the parameter independency (1), we can compute the conditional probability of each local structure independently. For each local structure, we calculate the conditional probability of knowledge features by counting its occurrence frequency. For the local structure of smoking and lung cancer in the ASIA network, the prior probability of the knowledge features can be calculated as $Pr(Dp)$=5/6, $Pr(I|Dp)$=1, $Pr(\overline{I}|\overline{Dp})$=1, $Pr(SP|I)$=1, $Pr(r_1|SP)$=1/5, $Pr(r_2|SP)$=1/5, $Pr(r_3|SP)$=2/5, $Pr(r_4|SP)$=1/5, $Pr(b_1|SP)$=1/2 and $Pr(b_2|SP)$=1/2 where $r_1 = [9, 11]$, $r_2 = [20, 25]$, $r_3 = [25, 30]$ and $r_4 = [30, \infty]$; $b_1 = [9\%, 11\%]$ and $b_2 = [14\%, 16\%]$. The continuous-valued feature R and Bd are discretized into $|R| = 4$ and $|Bd| = 2$ discrete-value intervals respectively. Based on the features and their prior belief, a set of qualitative knowledge $\widetilde{\Omega} = \{\Omega_1, \ldots, \Omega_{16}\}$ is formed in Figure 3(d).

3.3.1.1 ASIA Model Monte Carlo Sampling

Given the integrated qualitative knowledge set $\widetilde{\Omega}$ with prior probabilities, we now construct the Bayesian model class and the distribution on ground model space within each class. For demonstration purposes, we assume the partial structure and its parameters, i.e. $\{\alpha, \gamma, \lambda, f\}$, to be known as in Figure 3(b). Therefore the uncertainty of ASIA model space is restricted to the uncertainty of the local structure and parameter space on *Smoking* and *Lung Can-*

cer which can be described by $Pr(m|M_k)$ and $Pr(M_k)$ defined by $\{\Omega_k|k = 1,\ldots,9\}$, i.e. $\{M_k(\Omega_k)|k = 1,\ldots,9\}$, as well as the uncertainty of the local space on *Lung Cancer*, *Bronchitis* and *Dyspnea* which can be jointly determined by three types of model class, i.e. the root-dimension model class defined by Ω_{10}, the marginal-dimension model classes of lung cancer and dyspnea defined by $\{\Omega_i|i = 11,\ldots,14\}$ and the marginal-dimension model classes of bronchitis and dyspnea defined by $\{\Omega_j|j = 15,16\}$. Thus, there are total eight possible combination of these model classes, i.e. $\{M_k(\Omega_{10},\Omega_i,\Omega_j)|k = 10,\ldots,17; i = 11,\ldots,14; j = 15,16\}$ and each combination virtually forms a complete model class which defines the set of constraints on the structure and parameter space of ground Bayesian model for the local collider structure of lung cancer, bronchitis and dyspnea. The prior probability of each combination, $Pr(M_k)$ is the product of the prior probability of its independent components, i.e.

$$Pr(M_k) = Pr(\Omega_{10})Pr(\Omega_i)Pr(\Omega_j) \tag{23}$$

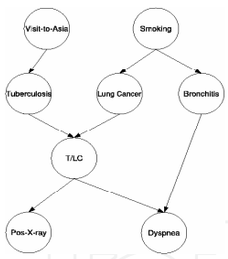
For each local structure, we perform 10,000 sampling iterations. In each iteration, we select a model class M_k randomly based on the prior probability of the model class, i.e $Pr(M_k)$. In each selected model class, we randomly choose 3 samples of ground Bayesian model m, whose structure and parameter space is consistent with the class constraints $Pr(m|M_k)$ as shown in Figure 1(a). In this way, for the local structure of smoking and lung cancer, the prior bability of the model class is equivalent to its knowledge component, i.e. $Pr(M_k)$=$Pr(\Omega_k)$. We generate total N=30,000 ground model samples from model classes $\{M_k(\Omega_k)|k = 1,\ldots,9\}$ defined by Ω_k in Figure 3(d). The ground model samples are shown in Figure 4(a). For the local collider structure of lung cancer, bronchitis and dyspnea, we generate N=30,000 ground model samples from the combination of model classes defined in Eq. 23 based on $\{\Omega_k|k = 10,\ldots,16\}$ in Figure 3(d). The marginal conditional probability samples are shown in Figure 4(b) and 4(c). Without further information on lung cancer, bronchitis and dyspnea, we can set their prior probabilities to be $1/2$. By taking average over the models in Figure 4(a) to 4(c), we can calculate the mean value for the conditional probability of lung cancer given smoking, i.e. $\overline{\beta_1}$=0.1255, $\overline{\beta_0}$=0.006, and of Dyspnea given lung cancer and Bronchitis, i.e. $\overline{\xi_0}$=0.2725, $\overline{\xi_1}$=0.9053, $\overline{\xi_2}$=0.5495 and $\overline{\xi_3}$=0.968. Note that since the *9th* model class defined by Ω_9 for the structure of lung cancer and smoking, i.e. $M_9(\Omega_9)$, contains no edge between the nodes, the parameter of this model class is null.

3.3.1.2 ASIA Model Inference

For each of the model sample, according to Eq. 23, we perform inferences *in silico* on the likelihood of a patient having lung cancer (Lc) given information about the patient's smoking status and clinical evidences including observation of X-ray, Dyspnea, and Bronchitis, i.e. $X_{obs} = \{Sm, Xr, Dy, Br\}$. The convergence of these prediction under a set of evidences $\widetilde{E} = \{E_1, E_2, E_3, E_4, E_5, E_6\}$ are shown in Figure 4(d). The true prediction values with parameters in Figure 3(b) under the evidence set $\widetilde{E}$ are listed below in Figure 5. The presence of bronchitis could explain away the probability of lung cancer and the presence of smoking increases the risk of getting lung cancer.

3.3.2 Breast Cancer Bone Metastasis Prediction

We apply our framework to integrate a set of inconsistent qualitative hypotheses about the molecular interactions between Smad proteins of the TGFβ signaling pathway in breast cancer bone metastasis network. From recent studies (11–15), a set of qualitative statements on



(a) ASIA Network Structure

Node Para	CPT:β	
$\{Pa_i, Ch_i\}$	$Pr(Ch_i\|Pa_i)$	$Pr(Ch_i\|\overline{Pa_i})$
{V,T}	α_1=0.05	α_2=0.01
{S,LC}	β_1=0.3	β_2=0.01
{S,B}	γ_1=0.6	γ_2=0.3
{T/LC,X}	λ_1=0.9	λ_2=0.05
$\{\{Pa_1, Pa_2\}, Ch_i\}$	$Pr(Ch_i\|Pa_1, Pa_2)$	$Pr(Ch_i\|Pa_1, \overline{Pa_2})$	$Pr(Ch_i\|\overline{Pa_1}, Pa_2)$	$Pr(Ch_i\|\overline{Pa_1}, \overline{Pa_2})$
{{T/LC,S},D}	ζ_1=0.9	ζ_2=0.7	ζ_3=0.8	ζ_4=0.1
{{T/LC},T/LC}	f_1=1	f_1=1	f_1=1	f_2=0

(b) ASIA Network Parameter

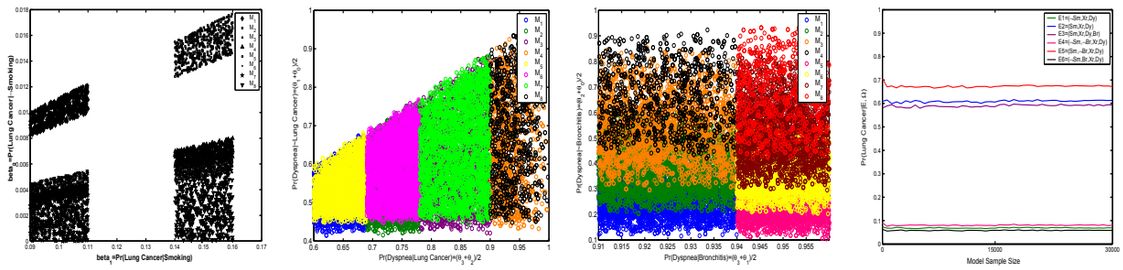
Stat.	Dp	I	Σ	R	Δ	Bd	Weight
S_1	1	1	SP	[9, 11]	null	null	w_1
S_2	1	1	SP	null	null	[9%, 11%]	w_2
S_3	1	1	SP	[25, ∞]	null	null	w_3
S_4	1	1	SP	[20, 30]	null	null	w_4
S_5	1	1	SP	null	null	[14%, 16%]	w_5
S_6	0	0	SP	null	null	null	w_6
S_7	1	1	SP	null	null	[60%, 90%]	w_7
S_8	1	1	SP	null	null	[60%, 90%]	w_8
S_9	1	1	SP	null	null	[60%, 100%]	w_9
$S_{10}(1)$	1	1	SP	null	null	[69%, 78%]	w_{10}
$S_{10}(2)$	1	1	SP	null	null	[91%, 94%]	w_{10}
S_{11}	1	1	SP	null	null	[94%, 96%]	w_{11}

(c) Feature-vector of Statements

Ω	Dp	I	Σ	R	Δ	Bd	$Pr(\Omega_i)$
Ω_1	1	1	SP	[9, 11]	null	[9%, 11%]	1/12
Ω_2	1	1	SP	[9, 11]	null	[14%, 16%]	1/12
Ω_3	1	1	SP	[20, 25]	null	[9%, 11%]	1/12
Ω_4	1	1	SP	[20, 25]	null	[14%, 16%]	1/12
Ω_5	1	1	SP	[25, 30]	null	[9%, 11%]	1/6
Ω_6	1	1	SP	[25, 30]	null	[14%, 16%]	1/6
Ω_7	1	1	SP	[30, ∞]	null	[9%, 11%]	1/12
Ω_8	1	1	SP	[30, ∞]	null	[14%, 16%]	1/12
Ω_9	0	0	null	null	null	null	1/6
Ω_{10}	1	1	PlSyn	null	null	null	1
Ω_{11}	1	1	SP	null	null	[60%, 69%]	3/11
Ω_{12}	1	1	SP	null	null	[69%, 78%]	4/11
Ω_{13}	1	1	SP	null	null	[78%, 90%]	3/11
Ω_{14}	1	1	SP	null	null	[90%, 100%]	1/11
Ω_{15}	1	1	SP	null	null	[91%, 94%]	1/2
Ω_{16}	1	1	SP	null	null	[94%, 96%]	1/2

(d) Integrated Qualitative Knowledge with Prior Probability

Fig. 3. ASIA Belief Network and Qualitative Statements and Knowledge in ASIA network



(a) Model Samples for Smoking and Lung Cancer

(b) Model Samples for Lung Cancer and Dyspnea

(c) Model Samples for Bronchitis and Dyspnea

(d) Quantitative Bayesian Inference

Fig. 4. ASIA Model Sampling and Inference

molecular interactions in the breast cancer bone metastasis network can be extracted. A Dynamic Bayesian model can be constructed based on this set of statements as shown in Fig. 6(a) and the quantitative prediction with forward belief propagation based on a set of consistent qualitative hypotheses has been introduced in (33).

In this section, we consider the inconsistent qualitative statements with regard to the mechanism of Smad7 in blockade of the TGFβ signals. In (14), the qualitative statements can be extracted as S_1*: Smad7 directly binds to the activated type I TGF-β receptor and inhibits phosphorylation of the R-Smads.;*S_2*: Smad6 acts in a different way as Smad7. It competes with the activated Smad1 for binding to Smad4.*; In (15), the qualitative statements can be extracted as S_3*: The inhibitory activity of Smad6 and Smad7 is thought to result from an ability to interfere with receptor interaction and phosphorylation of the receptor-regulated Smads.;*S_4*: However, their inhibitory activity might also result from their ability to form a complex with receptor-activated Smads.*;Similar statements can be extracted from (13) as S_5*: I-Smads (Smad6,7) interact with type I receptors activated by type II receptors.;*S_6*: I-Smads have also been reported to compete with Co-Smad (Smad4) for formation of complexes with R-Smads (Smad2/3).*

This set of statements represent the molecular interactions between I-Smad (Smad7), R-Smad (Smad2/3) and Co-Smad (Smad4). $\{S_1, S_3, S_5\}$ report the interaction between Smad7, type I TGFβ-receptor (TβRI) and Smad2/3. $\{S_4, S_6\}$ describe the interaction between Smad7 and Smad4 to form a complex whereas S_2 provides contradicting information. Each statement is analyzed by the hierarchical knowledge model in Figure 1(a) and the extracted features are

Exp.	E_1	E_2	E_3	E_4	E_5	E_6
True	0.17	0.87	0.84	0.21	0.91	0.11
Simulation	0.07	0.61	0.59	0.08	0.67	0.06

Fig. 5. Inference Results on ASIA Network

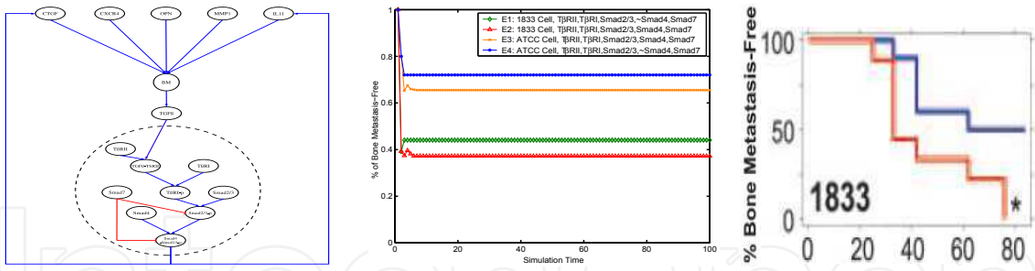
summarized in Figure 7(a). For simplicity, we assume the weight of every qualitative statement equals to 1, i.e. $\{w_i = 1, i = 1, \ldots, 6\}$. Due to the parameter independency (1), we can compute the conditional probability of each local structure by counting the occurrence frequency of the knowledge features independently. For the local structure of Smad7, TβRI and Smad2/3, the prior probability of the knowledge features can be calculated as $Pr(Dp)$=1, $Pr(I|Dp)$=1, $Pr(\bar{I}|\overline{Dp})$=1. For the local structure of Smad7, Smad4 and phosphorylated-Smad2/3 (Smad2/3-p), $Pr(Dp)$=2/3, $Pr(\overline{Dp})$=1/3, $Pr(I|Dp)$=1, $Pr(\bar{I}|\overline{Dp})$=1. Based on the features and their prior belief, a set of qualitative knowledge $\widetilde{\Omega}$ is formed in Figure 7(b). In this experiment, the extended features of the inconsistent knowledge are not available.

We now construct the Bayesian model class and the distribution on ground model space within each class. The uncertainty of the TGFβ-Smad BCBM model space is restricted to the uncertainty of the local structure and parameter space on Smad7, TβRI and Smad4 which is defined by {Ω_1,Ω_2} in Figure 7(b). The model classes can be expressed as $\{M_k(\Omega_k)|\text{k=1,2}\}$ and the prior probability of each model class equals to the prior probability of the knowledge, i.e. Pr(M_k)=Pr(Ω_k). We perform 10,000 sampling interactions. In each iteration, we select a model class M_k randomly based on the prior probability Pr(M_k). In each model class, we randomly generate 3 samples of the ground Bayesian model m by Monte Carlo method, whose structure and parameter space is consistent with the class constraints $Pr(m|M_k)$ as defined by Eq. 1 to Eq. 7. Therefore, we obtain N=30,000 ground models from the model classes. By taking average over the ground models, we can calculate the mean value for the conditional probability of the complex Smad4-Smad2/3-p given Smad7, Smad4 and Smad2/3-p. Note that since M_1 contains no edges between Smad7 and Smad4-Smad2/3-p, the parameter of this model class is null.

Each ground model is a Dynamic Bayesian network (DBN) which can be unrolled over time to form a series of 2TBNs (4). The prediction on the probability of bone metastasis given a set of evidences $E_i \in \{E_1, E_2, E_3\}$ in each model class, i.e. the integral in Eq. 23, can be calculated by integrating the predictions over all DBN models which is equivalent to compute firstly the mean DBN model with averaged parameters and then perform prediction on this mean DBN model (33). The simulation results and the observed bone metastasis probability in (11) are shown in Fig. 6(b) and Fig. 6(c).

3.3.3 Conclusion

In this paper, we proposed a hierarchical Bayesian model for modeling the semantics of the qualitative knowledge with a vector of features. The inconsistent knowledge components are integrated by calculating a prior distribution. The integrated qualitative knowledge set is used as prior background knowledge in modeling Bayesian networks and performing quantitative inference. We benchmarked our method with the ASIA network and applied our method to a real-world problem and simulation results suggest that our methods can reconcile the inconsistent qualitative uncertainty and produce reasonable quantitative prediction based on the inconsistent knowledge set.



(a) Signaling Pathway (b) Simulation Results (c) Observation

Fig. 6. Integrated TGFβ-Smad BCBM Network and Prediction

Stat.	Dp	I	Σ	R	Δ	Bd	Weight
S_1	1	1	MxSyn	null	null	null	w_1
S_2	0	0	null	null	null	null	w_2
S_3	1	1	MxSyn	null	null	null	w_3
S_4	1	1	MxSyn	null	null	null	w_4
S_5	1	1	MxSyn	null	null	null	w_5
S_6	1	1	MxSyn	null	null	null	w_6

(a) Feature-vector of Statements

$\widehat{\Omega}$	Dp	I	Σ	R	Δ	Bd	$Pr(\Omega_i)$
Ω_1	0	0	null	null	null	null	1/3
Ω_2	1	1	MxSyn	null	null	null	2/3
Ω_3	1	1	MxSyn	null	null	null	1

(b) Integrated Qualitative Knowledge with Prior Probability

Fig. 7. Qualitative Statements and Knowledge in TGFβ-Smad BCBM Network

3.4 Bayesian Network Learning with Informative Prior Qualitative Knowledge

We propose a framework for Bayes net parameter learning with generic prior knowledge. In this study, we use the knowledge model in section 3.1 to translate the qualitative domain knowledge into a set of inequality parameter constraints. We reconstruct the parameter priori distribution (i.e. priori pseudo counts) from these constraints. We then propose a novel Bayesian parameter score function which integrates this prior distribution with the quantitative data statistics. In this way, the parameter posterior distribution is combinatorially regulated by both quantitative data and prior knowledge.

3.4.1 Qualitative Constraints and Sampling

In general, qualitative domain knowledge can define various constraints over conditional probabilities in a BN. As described in last section, most of these constraints can be represented by a linear regression function $f(\theta_{ijk}) \leq c, \forall i, j, k$ (c is a scaler), where θ_{ijk} is the conditional probability of the state of i-th node being k, given its j-th parent configuration. In particular, one type of constraints can be derived from this function. *Cross-distribution Constraints* defines the relative relation between a pair of parameters over different conditions. If two parameters in a constraint share the same node index i and value k, but different parent configuration j, the constraint is called cross-distribution constraint. This constraints can be usually derived from causality in the qualitative knowledge.

$$\theta_{ijk} \leq, \geq \theta_{ij'k} \forall j \neq j' \tag{24}$$

Given the constraints defined by f, we can withdraw samples of parameter which are consistent with the constraints, e.g. in Eq. 24, by accept-reject sampling. Since sampling can be done

at each node, it is relatively reasonable for demonstration. But node with more parent nodes, Gibbs sampling and simulated annealing can be used.

3.4.2 Qualitative Bayesian Parameter Score (QBPS)

In this study, we assume the data distribution is multinomial and prior is Dirichlet. The posterior probability of the parameter given the data in standard MAP estimation can be written as

$$logPr(\theta|G,D) = \log Pr(D|\theta,G) + \log Pr(\theta|G) - c = \log\{\alpha \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+N'_{ijk}-1}\} \quad (25)$$

where θ denotes the parameters in a Bayes net and G is the network's structure. i, j, k is defined as section 3.4.1. The first term in Eq. 25 represent the data statistics which is followed by the Dirichlet prior distribution with hyperparameter N'_{ijk} (1). α is a normalizer. In standard MAP method, N'_{ijk} is usually set to a very small and equal number which results in non-informative prior.

We propose a posterior probability which employs the informative prior constraints (f) in the last section. In previous methods (35–37), f is imposed into the posterior probability as an penalty term. The MAP estimation is transformed to constrained optimization problem. However, the violation term f in these cases can only penalize the likelihood when the learned local maximum violates the constraints in the sign, but it can not distinguish a set of all possible local maximums obeying the constraints. So, final solution is not necessary a global maximum (37). Therefore, it is desired to use prior constraints (such as Eq. 24) as soft regulations to the posterior probability in Eq. 25. We name this MAP-like score function as Qualitative Bayesian Parameter Score (QBPS).

$$\log Pr(\theta|G,D,\Omega) = \log Pr(D|\theta,G) + \log Pr(\theta|G,\Omega) - c \quad (26)$$

The difference between Eq. 26 and Eq. 25 is the addition of Ω to the posterior probability in Eq. 25. The first term in Eq. 26 is the data statistics as in the standard MAP estimation. The second term $Pr(\theta|G,\Omega)$ represent the parameter's prior distribution given prior knowledge Ω. Ω can represent any forms of generic prior constraints over the parameter space, such as Eq. 24. In conventional approaches, $Pr(\theta|G)$ can be any probability function, such as Gaussian or Dirichlet distribution function with pre-defined hyperparameters. In case of multinomial data, $Pr(\theta|G)$ oftenly take the form of beta distribution due to the conjugate distribution property. Thus, the problem is to fuse the prior knowledge Ω and its associated constraints (f) over parameter space with the beta distribution $Pr(\theta|G)$ which results in the constrained beta distribution $Pr(\theta|G,\Omega)$.

In general, we can either i) fit the beta distribution into the constrained parameter space by estimating the hyperparameters of Dirichlet distribution given a vector of constrained parameter samples θ^l_{ijk} (43). These samples can be obtained based on the accept-reject sampling. In this case, we only select one local maximum prior model (one instance of hyperparameter) to substitute the uncertainty in the (priori) parameter space (all possible instances of hyperparameter) or ii) admit the model uncertainty and utilize conjugate property of beta distribution to reconstruct the (priori) parameter space distribution based on all constrained parameter samples. In this case, we have

$$Pr(\theta|\Omega,G) = \alpha \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} {\theta^l_{ijk}}^{M^l_{ijk}} \quad \forall l = 1,...,L \quad (27)$$

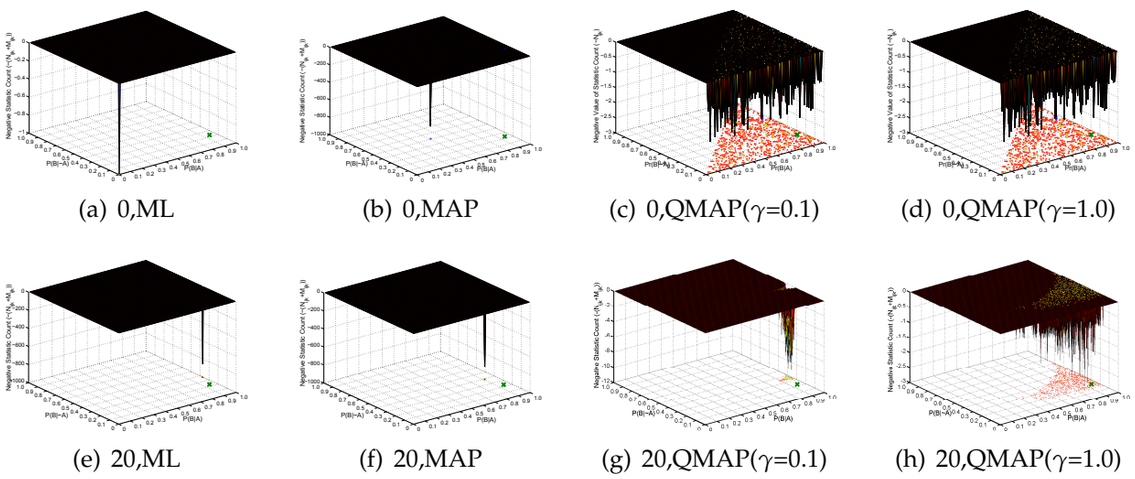
where θ^l_{ijk} is an instance of constrained prior parameter sample and M^l_{ijk} denotes the number of 'success' cases of this instance (X_i=k, Π_i=j$|\theta^l_{ijk}$) exists in the past A (A is an arbitrary number) samples. It is equal to

$$M^l_{ijk} = A \times Pr^l(X_i = k, \Pi_i = j|\Omega) \tag{28}$$

Together, the QBPS score can be written as

$$Pr(\theta|G, D, \Omega) = \alpha \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta^l_{ijk}{}^{N_{ijk}+M^l_{ijk}} \quad \forall l = 1, ..., L \tag{29}$$

where N_{ijk} is the number of occurrence in the training date for the ith node to have a value of k and for its parent to have a value of j and L is the total number of priori parameter samples from accept-reject sampling. (L is a large number) Thus, the local maximum estimation of a



(a) 0,ML (b) 0,MAP (c) 0,QMAP(γ=0.1) (d) 0,QMAP(γ=1.0)

(e) 20,ML (f) 20,MAP (g) 20,QMAP(γ=0.1) (h) 20,QMAP(γ=1.0)

Fig. 8. Parameter Learning in Toy Network: The network contains two binary nodes. A is an activator parent of B. X,Y-axles represent conditional probability P(B|A) and P(B|$\overline{A}$) respectively; Z-axis is equal to the negative value of posterior statistical counts [-(N_{ijk}+M^l_{ijk})] in Eq. 29.

QBPS score equals to

$$\widehat{\theta}^l_{ijk} = \frac{N_{ijk} + \gamma N_0 Pr^l(X_i = k, \Pi_i = j|\Omega)}{\sum_{k=1}^{K} N_{ijk} + \gamma N_0 Pr^l(X_i = k, \Pi_i = j|\Omega)} \tag{30}$$

where N_0 is equal to the number of total data samples. Now, we further assume that A and N_0 has a ratio γ, i.e. $A = \gamma \times N_0$. From Eq. 30, we can see that ratio γ actually specified the belief-ratio between data statistics and prior knowledge statistics. If γ=0, we neglect the statistics from the prior knowledge and only trust the statistics in the data, thus, our estimation in Eq. 30 converges to ML results; If γ=+∞, we neglect the statistics in the data and only trust the prior knowledge, the results converge to the previously mentioned constraint-based probabilistic inference in (Dynamic) Bayesian inference [9,10]. If 0<γ<+∞, the QBPS score is softly regulated by both data statistics and the prior knowledge and constraints in the domain.

Since the estimation in Eq.8 is a joint effect from both inequality constraints in qualitative prior knowledge and data observation, we name it as Qualitative Maximum a Posterior (QMAP) estimation.

3.4.3 QMAP Estimation

1. QMAP Estimation with Full Bayesian Approach

As we have shown, we can reconstruct the priori parameter distribution from prior constraints. Each priori parameter sample θ^l_{ijk} together with the given structure (G) define a prior network m^l. Each priori m^l can be mapped to a posteriori. Thus, the final posterior probability of all Bayesian network models is defined over this class of prior networks m^l in terms of a set QBPS scores (Eq. 29). Our final goal is to predict future observations on variable X from the training data (D) and priori constraints Ω. Given BN structure (G), this prediction can be calculated as integration over the parameter space weighted by its posterior probability.

$$Pr(X|G,D,\Omega) = \int_\theta Pr(X|\theta,G)Pr(\theta|G,\Omega,D)d\theta \tag{31}$$

The posterior probability of the parameter given data and qualitative prior knowledge, i.e. $Pr(\theta|G,\Omega,D)$, is in-turn an integration over all possible prior models (m) in the class defined by Ω, thus, we can extend Eq. 31 as

$$Pr(X|G,D,\Omega) = \int_\theta Pr(X|\theta,G)\int_m \frac{Pr(D|\theta,G)Pr(\theta|G,m)Pr(m|\Omega)}{Pr(D)}dmd\theta \tag{32}$$

$Pr(m|\Omega)$ in Eq 32 is equal to 1 since all the valid prior models (m) are consistent with the prior constraints Ω.

The outer integration can be approximated by its local maximum if we assume the QBPS curve for each model is peaky, then we can write the inference as $Pr(X|\widehat{\theta},G)$. With full Bayesian approach, final QMAP estimation of the parameter can be optimized by integrating the set of local QBPS maximums over the prior network space, i.e. selecting the QMAP estimation which maximize the integrated QBPS score.

$$\widehat{\theta} = argmax_\theta \left\{ \int_m \frac{Pr(D|\theta,G)Pr(\theta|G,m)Pr(m|\Omega)}{Pr(D)}dm \right\} = argmax_\theta \left\{ \frac{1}{L}\sum_{l=1}^{L} \alpha \prod_{ijk} \theta_{ijk}^{N_{ijk}+M^l_{ijk}} \right\} \tag{33}$$

Note that each prior network m^l uniquely associate with a pseudo prior statistical count M^l_{ijk}. The prior network space is discrete. By taking the derivative of Eq. 33 wrt θ_{ijk}, we obtain the constrained QMAP estimation with full Bayesian approach as

$$\widehat{\theta}_{QMAP,FBA} = \frac{1}{L}\left\{ \sum_{l=1}^{L} \frac{N_{ijk}+M^l_{ijk}}{\sum_k N_{ijk}+M^l_{ijk}} \right\} \tag{34}$$

2. QMAP with Frequentist Maximization Approach

On the other hand, the final QMAP estimation can be obtained by frequentist maximum approach to select one single best estimate from the parameter posteriori space. In this way, we could pick up the maximum from a set of local maximums.

$$\widehat{\theta}_{QMAP,FMA} = argmax_{\{l\}} \left\{ \frac{N_{ijk}+M^l_{ijk}}{\sum_k N_{ijk}+M^l_{ijk}} \right\} \tag{35}$$

An example plot of posterior statistical counts in Eq. 29 is shown in Fig. 8. In case of ML learning, the M^l_{ijk} is equal to zero for all i,j,k. In case of MAP learning, we simulated a typical scenario, where the dirichlet parameters are set equally to a scalar. In this case, the dirchlet parameters tends to smooth the posterior score by adding equal amount of pseudo counts for all i,j,k. The smoothed posterior favors to the uniformly distribution in this case. By setting these prior pseudo counts to 1, conventional MAP methods try to minimize this biased smooth effect. However, the bias remains significant when the training data is relative small. In Fig. 8(g) and 8(h), we show that our proposed QMAP methods augment the posterior distribution by reconstructing the prior from the qualitative knowledge and each prior distribution sample M^l_{ijk} is combined with the data statistics to regulates posterior counts on equal opportunities. In this way, we can explore the multiple local maximums sit in the posterior space so that we ensure to select the global maximum.

3.5 Experiments

3.5.1 Experiment Design

We evaluate our proposed parameter learning methods using a realistic AU recognition data. We test our algorithm in following learning conditions: a) In extreme case, we assume there are no available training data and we use only generic qualitative domain knowledge which are derived from causality in a BN to estimate the parameter. b) In standard case, we do not employ any domain knowledge which is eventually equivalent to ML estimation. c) In an fusion case, we use both training data and generic qualitative domain knowledge to learn the parameter. We compare our results to standard ML and MAP estimation results.

3.5.2 Facial Action Unit Recognition

In this section, we apply our method to facial action unit (AU) recognition. The Facial Action Coding System (FACS) (40) is the most commonly used system for facial behavior analysis. Based on FACS, facial behaviors can be decomposed into a set of AUs, each of which is related to the contraction of a specific set of facial muscles. An automatic AU recognition system has many applications. Current AU recognition methods tend to perform AU recognition individually, ignoring their relationships with other AUs. Due to the underlying physiology and the facial anatomy, AUs often move in a coordinated and synchronized manner in order to produce a meaningful expression. To represent the dependencies among AUs, Tong et al (41) proposed to use Bayesian Network to capture the relationships among AUs. Following their work, we propose to use the same BN model to capture the relationships among the 14 most common AUs as shown in Figure 9(a), where the larger circular nodes in the model represent AUs while the smaller nodes represent their image measurements. They have demonstrated that the BN model is superior to the state of the arts AU recognition method. But to use the model, they need a large amount of training data, which is often hard to acquire. We will show that we can achieve comparable results using only a fraction of their training data. Using the model, we extract constraints based on the following rules provided by domain experts: 1. *Marginal Constraint*: In spontaneous cases, some AUs rarely occur. One example for this case is AU27, and the rule is P(AU27 = 1)$\leq$P(AU27 = 0), where 1 means presence and 0 means absence. 2. *Causality-derived Cross-distribution Constraint*: As shown in Figure 4, every link between two AU nodes has a sign provided by the domain expert. The + sign denotes positive influence,which means two AU nodes have co-occurrence relationship, while a negative sign denotes negative influence, which means the two AU nodes have mutual exclusive relationship. Considering an AU node AU_i has only one parent node AU_j, if the

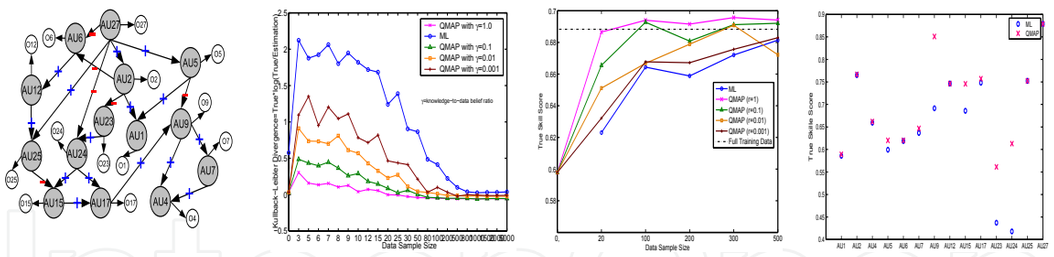
sign of the link is positive, we have $P(AU_i = 1|AU_j = 0) \leq P(AU_i = 1|AU_j = 1)$, e.g. $P(AU1 = 1|AU2 = 0) \leq P(AU1 = 1|AU2 = 1)$; if the sign of the link is negative, then we can get $P(AU_i = 1|AU_j = 1) \leq P(AU_i = 1|AU_j = 0)$, e.g. $P(AU6 = 1|AU27 = 1) \leq P(AU6 = 1|AU27 = 0)$. If an AU node AU_i has more than one AU parent nodes, AU^P denote all the parent nodes with positive links, and AU^N denote all the parent nodes with negative links. Then we get $P(AU_i = 1|AU^P = 0, AU^N = 1) \leq P(AU_i = 1|AU^P = 1, AU^N = 0)$, e.g. $P(AU15 = 1|AU24 = 0, AU25 = 1) \leq P(AU15 = 1|AU24 = 1, AU25 = 0)$. 3. *Range Constraint*: If an AU node AU_i has more than one parent nodes AUP , and all of them with positive influence, then $P(AU_i = 1|AU^P = 1) \geq 0.8$. If an AU node AU_i has more than one parent nodes AU^N, and all of them with negative influence, then $P(AU_i = 1|AU^N = 1) \leq 0.2$.

Please note the above constraints are due to either facial anatomy or due to certain facial patterns. They are generic enough to be applied to different databases and to different individuals.

3.5.3 Integrative Learning with domain knowledge and data

The 8000 images used in experiments are collected from Cohn and Kanades DFAT-504. In each simulation run, we randomly select 0 to 5000 samples out of 8000 samples for training and we repeat learning task for 20 times. Training data are used for learning the parameters in the AU BN (Figure 9(a)). After the learning, we select 1000 untouched samples for testing. Testing data are used to perform AU recognition through inference given learned BN. We assume the training data is complete. In the first part, we show the learning results in K-L divergence on the AU subnetwork in Figure 9(a). In the second part, we show the real classification results. We apply ML and QMAP estimation with qualitative domain knowledge defined above to learning the parameters in the AU subnetwork. The K-L divergence is shown in Figure 9(b). The x-axis and the y-axis denote training sample size and K-L divergence respectively. The K-L result is actually the mean K-L divergence which is calculated by averaging the parameter learning results over all randomly selected training samples under each specific sample size. We can see that: i) QMAP with γ=1 performs significantly better than ML estimation under every training data size. More specifically, the K-L divergence for ML estimation with 3 training sample is decreased from 2.21 to 0.24 for QMAP with γ=1. Even at 5000 training samples, the K-L divergence for ML estimation is decreased from 0.04 to close to 0 for QMAP estimation; On the other hand, we can evaluate the results by counting how many training samples are required to achieve specific desired K-L divergence level for ML, MAP and QMAP method respectively. At 3 training sample, K-L divergence for QMAP estimation is 0.24. In order to obtain equivalent or better K-L divergence level, ML estimation needs 200 samples. At 5000 training sample, K-L divergence for ML estimation is 0.04 which can be achieved by QMAP with 10 samples. These results are extremely encouraging, as using our methods with domain-specific yet generic qualitative constraints, and with a small number of manually labeled data (10), we can achieve similar learning accuracy to the ML estimation with full training dataset (5000).

The encouraging learning results of our QMAP method shed light over the usage of generic qualitative domain knowledge in learning task. Therefore, in this section, we explore an extreme case of parameter learning by ignoring all training data sample but only employing the set of qualitative constraints (same set of constraints defined above) to learn the AU subnetwork parameters. In this case, the data statistics counts in Eq. 30 is zero due to lack of training data. The parameter estimation is only determined by priori pseudo counts given the qualitative knowledge. The K-L divergence in this case is 0.0308 which is lower than K-L



(a) AU Recognition Network (b) KL Divergence (c) Averaged AU Recognition Skill (d) AU Node Skill with 200 samples

Fig. 9. Comparison of AU recognition network parameter learning results from ML and QMAP respectively. a) AU Recognition Network with AU nodes and measurement nodes; b) K-L divergence measurement of parameter learning in AU network based on training dataset with various sample size. Comparison of AU recognition skill using the BN learned from ML and QMAP respectively. We compare QMAP to standard ML skills. c) AU Recognition Network; d) AU Recognition skill score at 200 training samples on AU nodes;

divergence of ML learning with full training data (5000 training samples). Meanwhile, this K-L divergence level corresponds to that of QMAP learning with γ=1 at 25 data samples.

3.5.4 Classification

In this section, we want to study the performance of the proposed learning methods by using such learned BN model for AU classification. For AU classification, we need feed the BN model with AU measurements computed from Gabor Wavelet jets. Given the AU measurements, we want to infer the true states of each AU using the model parameters learnt with our method. Specifically, we want to study the AU recognition performance under different amount of training data including the extreme case of using no training data at all, and compare the classification results with those in (36). We perform classification based on the learned AU network from ML and our proposed QMAP approach in section 3.5.3). For demonstration, we select the learned AU network parameter under training dataset with representative sample size: 0, 20, 100, 200, 300 and 500. After learning, we randomly select 1000 untouched data samples for classification test. Figure 9(c) shows the AU recognition results. The x-axis represent the training data size for learning AU network parameters (in case of 0 training size, no training data but only qualitative prior knowledge is used for AU network parameter estimation) and y-axis denotes the true skill score (the difference between true positive rate and false positive) respectively. The true skill is calculated by averaging all AU nodes' skill score. We can see from Figure 9(c), the true skill score for QMAP with various belief-ratio (γ) is significantly better than the skill score for ML estimation under nearly all training data sample size except for QMAP with γ=0.01. In particular, even at sparse training data (20 samples), the average true skill score for all AU nodes increases from 0.6229 for ML estimation to 0.6866 for QMAP with γ=1, to 0.6655 for QMAP with γ=0.1, to 0.6512 for QMAP with γ=0.01 and to 0.6322 for QMAP with γ=0.001; At 100 training samples, true skill score further enhances from 0.6644 for ML estimation to 0.6940 for QMAP with γ=1, to 0.6928 for QMAP with γ=0.1, to 0.6668 for QMAP with γ=0.01 and 0.6677 for QMAP with γ=0.001. While training sample size grows to 200, 300, and 500 samples, the true skill score from QMAP with γ=1.0 is equal to 0.6916, 0.6957 and 0.6942 respectively and tends to converge. In the

same case, ML estimation shows consistently lower classification ability than QMAP. Please note that, using full training dataset (7000 samples for training and 1000 samples for testing), true skill score for ML estimation converge at 0.6883 (shown as the black dashed line in Figure. 9(c)). From the above results, we can conclude that i) our proposed QMAP estimation by integrating domain-specific yet very generic qualitative prior constraints with quantitative training data significantly improves the AU recognition results comparing to ML estimation at all sample size spanning from sparse data to rich data. This observation is particularly true with γ=1; ii) Our proposed QMAP estimations (with different γ) needs much fewer training samples for AU network to achieve equivalent and even better AU recognition results than ML estimation. iii) Comparing the true skill score of QMAP estimation to the score of ML estimation with full training dataset, we can see that, with a much smaller number of manually labeled data (around 35 samples) ,QMAP with γ=1 can already achieve much better AU recognition results than ML estimation with full training dataset (7000 samples). While decreasing the weight on prior knowledge to γ=0.1, QMAP requires from 80 to 250 training samples to achieve better AU classification results than ML estimation with full training dataset. When γ reduces to 0.01, QMAP needs around 300 samples to outperform ML estimation with full training dataset. This number keeps increasing while γ reduces. When γ=0.001, the true skill score of QMAP tends to converge with ML estimation. Therefore, in practice, we shall put a larger weight on qualitative prior knowledge as long as our knowledge are valid in a domain. The above observation is also consistent with our K-L measurements in Figure 9(b). In summary, we demonstrate that by our approach, qualitative prior constraints can be integrated into standard BN parameter learning to achieve significantly improved prediction results. Next, we want to compare our results with a well developed method in AU recognition (36). To this end, we compare the true skill score of our QMAP at 200 training samples to the skill score of Constrained-ML (CML) estimation (Figure4(b) in (36)) at 300 training samples. The true skill of each AU node of our QMAP is plot with optimized γ is shown in 9(d). Firstly, we can see that our QMAP approach significantly improves the true skill on AU node number 5, 9, 15, 23 and 24. Slightly improve the skill on AU node 1, 7, 17. The rest skill is equivalent to ML estimation. Comparatively, our method boost the skills on those AU nodes (6, 23, 12, 25, 17, 24, 9, 4) whose skill score is worse than ML estimation in (36).

4. References

[1] D. Heckerman.(1996). A tutorial on learning with bayesian networks, Microsoft Research, USA, Tech. Rep.

[2] D. J. Dudgeon and M. Lertzman.(1998). Dyspnea in the advanced cancer patient, J. of Pain and Symptom Management, vol. 16(4), pp. 212-219.

[3] E. Gyftodimos and P. Flach.(2002). Hierarchical bayesian networks: A probabilistic reasoning model for structured domains, Proceedings of the ICML-2002 Workshop on Development of Representations, pp. 23-30.

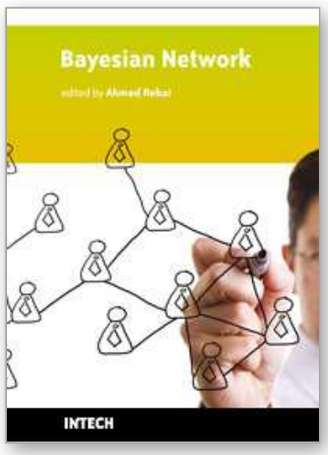
[4] K. Murphy.(2002). Dynamic bayesian networks:representation, inference and learning, Ph.D. dissertation, University of California, Berkeley, USA.

[5] M. J. Druzdzel and M. Henrion.(1993). Efficient reasoning in qualitative probabilistic networks, Proceedings of the Eleventh National Conference on Artificial Intelligence. Washington DC: AAAI Press, 1993, pp. 548-553.

[6] M. P. Wellman.(1990). Fundamental concepts of qualitative probabilistic networks, Artificial Intelligence, vol. 44, pp. 257-303.

[7] P. Edmonds, S. Karlsen, S. Khan, and J. Addington-Hall.(2001). A comparison of the palliative care needs of patients dying from chronic respiratory diseases and lung cancer, Palliative Medicine, vol. 15(4), pp. 287-295.
[8] R. Kinsman, R. Yaroush, E. Fernandez, J. Dirks, M. Schocket, and J. Fukuhara.(1983). Symptoms and experiences in chronic bronchitis and emphysema, Chest, vol. 83, pp. 755-761.
[9] S. Renooij, L. C. van der Gaag, and S. Parsons.(2002). Context-specific sign-propagation in qualitative probabilistic networks, Artificial Intelligence, vol. 140, pp. 207-230.
[10] S. L. Lauritzen and D. J. Spiegelhalter.(1998). Local computations with probabilities on graphical structures and their application to expert systems, J. Royal Statistics Society B, vol. 50(2), pp. 157-194.
[11] Y. Kang, W. He, S. Tulley, G. P. Gupta, I. Serganova, C. R. Chen, K. Manova-Todorova, R. Blasberg, W. L. Gerald, and J. Massague.(2005). Breast cancer bone metastasis mediated by the smad tumor suppressor pathway, Proceedings of the National Academy of Sciences of the USA.
[12] Y. Kang, P. M. Siegel, W. Shu, M. Drobnjak, S. M. Kakonen, C. Cordón-Cardo, T. A. Guise, and J. Massagué.(2003). A multigenic program mediating breast cancer metastasis to bone, Cell, vol. 3, no. 6, pp. 537-549.
[13] K. Miyazono.(2000). Positive and negative regulation of TGF-β signaling, Journal of Cell Science, vol. 113, no. 7, pp. 1101-1109.
[14] Y. Shi and J. Massagué.(2003). Mechanisms of TGF-β signaling from cell membrane to the nucleus, Cell, vol. 113.
[15] Y. Zhang and R. Derynck.(1999). Regulaiton of smad signalling by protein associations and signalling crosstalk, Trends in Cell Biology, vol. 9, no. 7, pp. 274-279.
[16] Thomas Bayes, *An essay towards solving a Problem in the Doctrine of Chances*, Philosophical Transactions of the Royal Society of London,1763.
[17] S.L. Lauritzen and D.J. Spiegelhalter, *Local Computations with Probabilities on Graphical Structures and Their Application to Expert Systems*, Journal of the Royal Statistical Society, 1988
[18] Judea Pearl,*Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*,Morgan Kaufmann Publishers, Inc., San Mateo, CA, USA,1988
[19] David Heckerman, *A Tutorial on Learning with Bayesian Networks*,1996
[20] David Heckerman, *Learning Bayesian Networks: The Combination of Knowledge and Statistical Data*, KDD Workshop, 1994
[21] Nir Friedman and Moises Goldszmidt, *Learning Bayesian networks with local structure*, Learning in graphical models,1999
[22] Eric Neufeld, *A probabilistic commonsense reasoner*, International Journal of Intelligent Systems, 1990
[23] M.J. Druzdzel and L.C. van der Gaag, *Elicitation of probabilities for belief networks: combining qualitative and quantitative information*, Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, 1995
[24] Silja Retive Approaches to Quantifying Probabilistic Networks,Ph.D Thesis Universiteit Utrecht,2001
[25] Michael P. Wellman, *Fundamental Concepts of Qualitative Probabilistic Networks*, Artificial Intelligence, 1990
[26] M. Dejori and M. Stetter, *Identifying interventional and pathogenic mechanisms by generative inverse modeling of gene expression profiles*, J. Comput. Biology,1135-1148,2004

[27] Jesús Cerquides and Ramon López de Màntaras, *Knowledge Discovery With Qualitative Influences and Synergies*,Proceedings of the Second European Symposium on Principles of Data Mining and Knowledge Discovery,1998
[28] D. Heckerman. Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Proc. KDD Workshop*,
[29] D. Geiger and D. Heckerman. A characterization of the dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, 25:1344Ű1369, 1997.
[30] R. S. Niculescu. Exploiting parameter domain knowledge for learning in bayesian networks. *Technical Report CMU-TR-05-147*, Carnegie Mellon University, 2005.
[31] R. S. Niculescu, T. Mitchell, and R. B. Rao. Parameter related domain knowledge for learning in graphical models. *In Proceedings of SIAM Data Mining conference*, 2005.
[32] R. S. Niculescu, T. Mitchell, and R. B. Rao. Bayesian Network Learning with Parameter Constraints. *Journal of Machine Learning Research*, 7:1357Ű1383, 2006.
[33] R. Chang, M. Stetter, and W. Brauer. Quantitative Inference by Qualitative Semantic Knowledge Mining with Bayesian Model Averaging. *IEEE Transaction on Knowledge and Data Engineering*, Vol. 20, No. 12, December, 2008.
[34] R. Chang, W. Brauer, and M. Stetter. Modeling semantics of inconsistent qualitative knowledge for quantitative Bayesian network inference. *Neural Networks*, 21(2-3): 182-192, 2008.
[35] F. Wittig and A. Jameson, Exploiting Qualitative Knowledge in the Learning of Conditional Probabilities of Bayesian Networks. *The 16th Conference on Uncertainty in Artificial Intelligence*, USA, 2000.
[36] Yan Tong and Qiang Ji, Learning Bayesian Networks with Qualitative Constraints, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008.
[37] E. Altendorf, A. C. Restificar and T. G. Dietterich: Learning from Sparse Data by Exploiting Monotonicity Constraints. *The 21st Conference on Uncertainty in Artificial Intelligence*, USA, 2005: 18-26.
[38] Linda van der Gaag, B. Hans and Ad Feelders, Monotonicity in Bayesian Networks. *The 20th Conference on Uncertainty in Artificial Intelligence*, USA, 2004.
[39] Y. Mao and G. Lebanon, Domain Knowledge Uncertainty and Probabilistic Parameter Constraints. *The 25th Conference on Uncertainty in Artificial Intelligence*, USA, 2009.
[40] P. Ekman and W. Friesen. Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, 1978.
[41] Yan Tong, Wenhui Liao, Zheng Xue and Qiang Ji, A Unified Probabilistic Framework for Spontaneous Facial Activity Modeling and Understanding, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.
[42] Wenhui Liao and Qiang Ji, Learning Bayesian Network Parameters Under Incomplete Data with Qualitative Domain Knowledge, Pattern Recognition, Volume 42 , Issue 11, Pages 3046-3056, 2009.
[43] T. P. Minka, Estimating a dirichlet distribution, 2003. [Online].



Bayesian Network

Edited by Ahmed Rebai

ISBN 978-953-307-124-4

Hard cover, 432 pages

Publisher Sciyo

Published online 18, August, 2010

Published in print edition August, 2010

Bayesian networks are a very general and powerful tool that can be used for a large number of problems involving uncertainty: reasoning, learning, planning and perception. They provide a language that supports efficient algorithms for the automatic construction of expert systems in several different contexts. The range of applications of Bayesian networks currently extends over almost all fields including engineering, biology and medicine, information and communication technologies and finance. This book is a collection of original contributions to the methodology and applications of Bayesian networks. It contains recent developments in the field and illustrates, on a sample of applications, the power of Bayesian networks in dealing the modeling of complex systems. Readers that are not familiar with this tool, but have some technical background, will find in this book all necessary theoretical and practical information on how to use and implement Bayesian networks in their own work. There is no doubt that this book constitutes a valuable resource for engineers, researchers, students and all those who are interested in discovering and experiencing the potential of this major tool of the century.

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen