Institut für Informatik
der Technischen Universität München

Advanced Probabilistic Network Modeling Framework with Qualitative Prior Knowledge

Rui Chang

Vollständiger Abdruck der von der Fakultät für Informatik der Technischen Universität München zur Erlangung des Akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:	Univ.-Prof. Dr. H. J. Schmidhuber
Prüfer der Dissertation:	
1.	Univ.-Prof. Dr. Dr.h.c.mult. W. Brauer, em.
2.	Univ.-Prof. B. Brügge, Ph.D.

Die Dissertation wurde am 07.01.2008 bei der Fakultät der Technischen Universität München eingereicht und durch die Fakultät für Informatik am 28.01.2008 angenommen.

To my parents

Abstract

The ever increasing amount of information in every scientific and industrial domain have been an exciting challenge for computer scientist to handle vast amount of data and to represent human understandings of a domain in a systematic and mathematic way. Over decades, probabilistic modeling with probability theory and statistical learning algorithms has been popular for accomplishing this task due to the stochastic characteristics of the nature. Quantitative measurements are generated from various kinds of ”sensors” in all types of science and industry and we need to make sense of these data, i.e. to extract important patterns and trends, and understand ”what the data says”. This is often called learning from data, reverse-engineering or bottom-up modeling. Among these learning algorithms, Bayesian network computational framework has become particular popular due to the ability of Bayesian network to model cause-effect interactions between the variables in a domain. For example, in bioinformatics, vast amount of ”-omics” data are generated by high-throughput screening techniques. Learning method with Bayesian networks has been used to construct gene regulatory networks from transcriptomic data and to predict protein-protein interactions based on proteomic data.

In practice, the data basis in reverse-engineering approach can be very sparse. Therefore, it is hardly sufficient to select one adequate model, i.e. there is considerable model uncertainty. Selecting one single Bayesian model can then lead to strongly biased inference results. In this case, full Bayesian approach with model averaging can be used to alleviate the bias. In this approach, one major difficulty is to specify prior distribution function on the Bayesian network structure space and parameter space in order to compute a posterior probability. One important information resources that could provide solutions to this problem is qualitative prior distribution which largely exists in every science and industry domain. In addition, human have a deep intuition that causality is a central and cohesive aspect of their perceptions, therefore, one subtype of these qualitative prior knowledge, i.e. qualitative causal knowledge which describes the cause-effect relations between multiple entities with any form of uncertainties, are particularly well-suited to represent human understandings and to get approximated characterizations of the behavior of the interested domain. For example, in a qualitative causal statement: *”smoking increases the risk of lung cancer”*, two entities: smoking and lung cancer are related to each other. Moreover, smoking positively influences lung cancer since lung cancer risk is increased in case of smoking. It is therefore desirable to make use of this body of evidence in probabilistic modeling with Bayesian network.

This thesis is concerned with developing a powerful probabilistic modeling framework to represent human understandings of a domain based on qualita-

tive prior knowledge. More precisely, to construct a Bayesian network structure with cause-effect relationships between the entities in a domain and parameterize these interactions according to the semantics of qualitative knowledge. One problem here is that qualitative knowledge provides no quantitative information to parameterize edges in Bayesian network and parameters need to be configured based on soley qualitative information. We attack this problem by proposing a qualitative knowledge model which is responsible for constructing mathematical constraints to define parameter distribution based on the qualitative knowledge. This approach incorporates the concept of model uncertainty due to the qualitative nature of the statements and automatically select a class of possible Bayesian models which are consistent with the semantics of the statements. Quantitative Bayesian network inference is performed by averaging inferences of each Bayesian network in this class with full Bayesian approach.

However, knowledge is well-known to be inconsistent and incomplete. Knowledge has spatial and temporal properties like other physical systems, i.e. knowledge exist in space-time dimension. The spatial property describes that knowledge represents information on a specific sub-structure of a domain and the temporal property states that knowledge represents human understandings at a particular time point. Thus, these knowledge are incomplete and may be updated by complementary discovery. Moreover, another significant drawback of knowledge is inconsistency. In the same domain, there may exist contradicting qualitative statements on dependency, causality and parameters over a set of entities. In this thesis, we propose several successful methods to deal with knowledge incompleteness and inconsistency, and integrate the Bayesian networks based on the set of knowledge to form an complete and coherent representation of the underlying system.

Acknowledgements

This dissertation is developed based on my work within a corporative Ph.D program between the Informatics Institute of Technical University of Munich and the Learning System Department (CT IC4) of Siemens AG. In the past 3 years, it has been my extreme pleasure and fortune to work within such a friendly and professional atmosphere provided by both parties of my Ph.D program. This thesis would not have been possible without the kindness and generosity of all the members of our group. I feel grateful and indebted to have received their helps and suggestions.

First of all, I would like to thank my academic supervisor, Prof. Wilfried Brauer from the Institute of Informatics of Technical University of Munich who has been constantly supporting me on my researches by providing me his remarkable insights and constructive suggestions. Prof. Wilfried Brauer is such kind person so that whenever I need his help, he is always there for me. I am very grateful to his kindness and I am deeply impressed by his open-mindedness and astonishing academic achievements.

Secondly, I would like to thank Siemens AG, especially Siemens Corporate Technology, for endowing me such a great opportunity to carry out the cutting-edge researches in machine learning and Bioinformatics field and for providing me the Siemens Doctoral scholarship. I feel grateful to Prof. Bernd Schürmann, the leader of the Learning Systems Department at Siemens AG, for his constant support to my research. I appreciate his emphasis on both scientific research and real-world applications.

Specially, I am indebted to my co-supervisor at Siemens AG, Dr. Martin Stetter, the principle investigator of my team, who has brought me into the world of statistical learning and Bioinformatics. He has the greatest influence in my intellectual, knowledge development and my commitment to my future career. I am deeply impressed by his enthusiasm, intelligence, open-mindedness which constantly encourages me to overcome difficulties and to pursue successes in my research and make my time at Siemens AG memorable.

I appreciate the friendship and fellowship to my colleagues, Dr. Mathäus Dejori, who often enlighten me with his smart ideas. His solid knowledge in machine learning and sharp way of thinking often bring us surprises. Also, I enjoyed the scientific discussions with Andreas Nägele who often provide me useful advices from another point of view. His logic way of thinking impressed me a lot. Meanwhile, I would like to thank Holger Arndt who answered my questions on programming and computer hardware. Specially, I would like to thank Dr. Jakub Pijewski with whom I enjoyed to talk about biology with his talented minds. I would like to thank to the secretary Mrs. Christina Singer at Siemens AG and Mrs. Erika Leber from Technical University of Munich, who kindly helped me to deal with administration documents. Moreover, I would like to thank to Dr. Volke Tresp, Dr. Kai Yu, Dr. Shipeng Yu, Dr. Xu Zhao, Yi Huang and Huaien Gao, who often answer my questions and give pleasant chats which make my research joyful. Finally, I would like to thank to my parents and Ms. Wenbo He who constantly give me the most love, support and care in my personal life without which I could not have ever accomplished my Ph.D.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The rapid growth of information in every scientific and industrial domain raises exciting challenge in handling vast amount of data and modeling underpinnings of a domain in a systematic and mathematic manner. In recent years, probabilistic network has become popular as practical representations of knowledge for reasoning under uncertainty. The probabilistic network computational framework uses a graphical model to capture random variables in a domain and relations between them, with probabilities that represent the uncertainties in the domain. The framework offers powerful algorithms of quantitative reasoning, such as predictive inference and diagnostic reasoning. Among these and other probabilistic graphical models, directed graphical models (also called Bayesian Networks or Belief Networks) [82, 111] are particular attractive for researches with the Artificial Intelligence (AI) and statistics communities. The most prominent advantage of Bayesian network is that an edge from one node to another can be viewed as "causality" and the uncertainty of this causal relation is quantified by conditional probability distribution (CPD).

Bayesian network computational framework has been widely utilized in various nature and industry-inspired domains, such as medical diagnosis, prognosis and nature language processing. The most widely used Bayesian network are undoubtedly the ones embedded in Microsoft's products, including the Answer Wizard of Office 95, the Office Assistant of Office 97, and over 30 Technical Support Troubleshooters [78]. BNs originally arose out of an attempt to add probabilities to expert systems, and this is still the most common use for BNs. Another famous example is QMR-DT network [72] is a two-level graphical model. The top level of the graph contains nodes for the diseases, and the bottom level contains nodes for the symptoms. The goal is to infer the posterior probability of each disease given all the symptoms. An interesting fielded application is the Vista system which is a decision-theoretic system that has been used at NASA Mission Control Center in Houston for several years. The system uses Bayesian networks to interpret live telemetry and provides advice on the likelihood of alternative failures of the space shuttle's propulsion systems.

Perhaps, the most popular application of Bayesian network framework in now days is computational biology and bioinformatics. A common thread in projects in bioinformatics and computational biology is the use of mathematical tools to extract useful information from data produced by high-throughput biological techniques. Biologists are working since decades to untangle the complex and diverse principles and mechanisms underlying living cellular process and the pathogenesis of diseases. Consequently, enormous amount of data about genes, proteins, signaling transduction pathways, disease markers, metabolites and their relations to phenotype concepts are generated to provide a comprehensive knowledge base which favors the elucidation of yet unknown underpinnings of cellular process. With rapid accumulation of these data, existing data and knowledge bases are at the edge to turn into an inextricable jungle. Therefore, scientists share the opinion that discernible principles are imperative to be mined from the overwhelming collection of data.

All processes in a living cell are directly and indirectly related to each other through complex, recurrent and mutually interacting signaling pathway. Proteins are synthesized from genes, interact with each other and with smaller molecules, and bind to RNA and DNA where they regulate the production of other proteins. It is not sufficient to take individual components of cellular mechanism for granted. Global understandings of the landscape cellular processes with irresolvable coupled and concerted interaction components in a cell require analysis at a systems level. In light of this view, Bayesian networks are well suited towards a complete understanding of the cellular mechanisms and related inter- and intracellular processes. Bayesian networks are useful for extracting meaningful biological insights from the resulting data sets and provide a concise representation of complex cellular networks by composing simpler components. Computational framework based on well-understood principles for inferring Bayesian networks from data facilitate a model-based methodology for analysis and discovery. This methodology and its capabilities are illustrated by several recent applications to gene expression data.

There are major two types of data mining practice in inferring Bayesian network of cellular interaction network from "-omics" data (reverse-engineering approach). The first class of methods automatically identifies patterns and trends in the data. For example, the patterns can be cluster of genes which are co-expressed of a specific organism under certain conditions. Clustering studies may disclose many extended cluster of molecules, e.g. cluster of genes, which collectively change their expression levels when a cell or organism transit their status from one mode to another. [2, 35]. In fact, it has been shown that many coexpressed genes and proteins are known to interact to each other in a signaling transduction pathway which indicates that these global expression patterns reflect the execution of specific cellular programs. However, clustering analysis can not provide the structure of molecular interaction networks, i.e. to answer questions like: which gene(s) are dominant components in the genetic regulation network underlying a specific global expression pattern? Which protein functions fail and drive the organism to global disorders. Therefore, the second class of reverse-engineering approaches have concentrated on inferring the structure of molecular interaction network from "-omics" data. (It is also known as structure learning of graphical models in machine learning community [53]) For example, learning a structure of genetic regulatory network from microarray data [36, 47, 54, 96]. In this approach, the set of expression data are assumed to be drawn from a high-dimensional multivariate probability density function which is modeled with Bayesian network with adaptive structure. Each node in the network is assigned to a gene and each edge between genes hints towards a regulatory relationship between them. The edge can be presumptively interpreted as a causal relationship. Bayesian network inference algorithms can be applied on the top of the inferred genetic regulatory network to answer "what-if" scenarios.

1.1 Motivation

These and other practices of computational modeling with Bayesian network, especially, the computations of inferring molecular interaction network from "-omics" data in bioinformatics, reveal a number of built-in problems of the data driven reverse-engineering approach. There are mainly three concerns with these approaches, i.e. sparse data which may induce overfitting, computational complexity due to high-dimensionality of the data and integrative analysis of data sets from multiple resources.

In machine learning, overfitting is likely to occur when inducing a probabilistic network structure in the presence of sparse data set with thousands of attributes and relative large noise level [73]. The learner is expected to reach a state where it will be able to predict the correct output for examples not in the training data set, thus generalizing to situations not presented during training. Instead, when overfitting occurs, the learner may adjust to very specific random features of the training data, that have no causal relation to the target function where the performance of the learner on the training examples still increases while the performance on unseen data becomes worse. For example, in studies of inducing genetic regulation network from microarray data, high-throughput screening techniques enable biologists to measure the expression levels of thousands of genes in one experiment [26, 68, 110]. The data generated from these experiments consists of instances, each one of which has thousands of attributes. However, the largest datasets available today contain only few hundreds of instances. Moreover, the genuine signals, i.e. correlations and causations between genes, are overwhelmed by the random noise. These data sets hardly provide sufficient entire "statistics" of an underlying system. We cannot expect to learn robust and detailed model from such a sparse data set and it is not too surprising that our results overfit to the sparse data. Certainly, overfitting can be avoided to some extent by additional techniques in both statistics and machine learning. The statistical re-sampling based methods, i.e. cross-validation and bootstrap, are often used to estimate generalization error (bias) and to explore the robustness (variance) of the Bayesian network structures learned from sparse data [24, 34, 38]. In machine learning framework, full Bayesian approach with Bayesian model averaging (BMA) can be used to compute a posterior probability distribution over all possible models to reflect the true model uncertainty [76]. This approach computes an average of the quantity of interest under each of the models considered, weighted by their posterior model probability. Thus, Bayesian model averaging provides a coherent mechanism for accounting uncertainty in model selection and avoids over-confident inferences and decisions. In fact, it has been demonstrated that averaging over all the models in this fashion provides better average predictive power, as measured by a logarithmic scoring rule, than using any single model [50, 69].

While Bayesian model averaging is an intuitively attractive solution to the problem of overfitting by incorporating model uncertainty, it is not yet part of the standard data analysis toolkit. One of the major difficulties in the implementation of Bayesian model averaging methods is how to specify the prior distribution over model structure space and model parameter space [50, 76]. In the domain where little prior information on parameters is available, non-informative prior distribution, e.g. uniform distribution, is a reasonable choice [38,50]. However, others propose to employ Dirichlet distribution, to compute the marginal likelihood of data given model structure [49, 57]. On the contrary, if there are sufficient amount of prior information available in a domain, the diverse prior information can be used. For example, in cell biology, we often have prior knowledge about the molecular interactions and the "signs" of the interaction, i.e. whether an interaction is a positive or negative regulation. Qualitative probabilistic network [108] are proposed to model this prior knowledge and to translate the knowledge into constraints on the entries of multinomial conditional probability table(See Section 1.3.1). Another solution is to model the "signs" of qualitative prior knowledge with constrained Dirichlet distributions which is equivalent to imposing penalties on the scoring function in structural learning [47]. Exponential function with arguments derived from qualitative prior knowledge can be also utilized to penalize the scoring function [112]. A prototype method is proposed for using qualitative probabilistic information to construct a probability distribution function in the parameter hyperspace. This approach automatically generate parameters of interest by solving a complex system of (in)equalities [29]. Even though, this method is useful in automatic generating parameters from prior probabilistic information, however, the restrictive set of constraint complex with a large number of arguments (joint probabilities) often introduce infeasible computation complexity in practice.

In learning of Bayesian network structures from data, there are two major approaches for inducing the structures. The first approach puts the learning task as a constraint satisfaction problem. In that approach, we try to estimate properties of conditional independence among the attributes in the data and find a graph which satisfies all these independence constraints in the data. Usually this is done using a statistical hypothesis test, e.g. [83, 101, 103]. Although the constraint satisfaction approach is efficient, it is sensitive to failures in independence tests. Thus, the common opinion is the second approach which is a better tool for learning structure from data. The second approach poses learning as an optimization problem. We start by defining a statistically motivated score function that describes how fit each possible structure is to the data. These scores include Bayesian scores (e.g. BDe scores) [49] and BIC, MDL scores [20, 64]. The task of learner is then to find a structure that maximizes the score. In general, this is an NP-hard problem [18], and thus we need to resort to heuristic methods. The commonly used maximum-likelihood based learning methods often result in a complete graph since it has the largest number of parameters to fit to the data best. Thus, it is necessary to specify a prior distribution over the discrete structure model space in the objective function to prevent the overfitting of learned structures to the data. For structure priors, when there is little prior knowledge, a well-principled way to avoid this kind of overfitting is to impose a prior on models which is a penalty function of the number of edges and the size of families to discourage networks that are globally or locally dense [5,48,77]. This principle is compatible to the proposition stated in Occam's Razor which assume that the simplest consistent hypothesis about the target function is actually the best (Consistent means that the hypothesis of the learner yields correct outputs for all of the training examples). In domains where large amount of prior information is available, we can construct a prior

Bayesian network and penalize the objective function by deviation of candidate models from this prior network. The deviation is computed by counting the number of mismatched edges [49]. In [76], a graph weight matrix is introduced to model the prior ”fuzzy” knowledge on the presence/absence of edges and their orientations in a structure.

The second built-in problem in data-driven reverse-engineering approach is caused by the large number of variables in data, i.e. high-dimensionality of the data. In particular, when learning structure, with the number of variables increases the space of possible graph structures grows superexponentially and makes the learning problem NP-hard. Thus, one has to resort to heuristic search strategies (See section 1.2.3). Most applied heuristic search techniques, such as greedy hill-climbing, search all possible local changes in each step and apply the one that leads to the biggest improvement in score. The computational complexity of these evaluations becomes inextricable when we learn from high-dimensional data. Therefore, as preventing the overfitting in the learning, we can use prior knowledge on the structures space to reduce the size of the search space and improve both the speed of induction and more importantly, the quality of the learned network. In addition, the acyclic directed structure of Bayesian network can not capture the temporal characteristics of a system. Thus, Markov networks or more precisely dynamic Bayesian network which allows cyclic connections has been proposed

Regarding to the third built-in problem, as more sources of high-dimensional data have become available, many efforts have been made to automatically integrate both homogeneous and heterogeneous types of data for the prediction of the mechanisms and features of a underlying system. These multi-source data sets are usually independent measures of entities at different scales and/or facets of the system based on a variety of techniques and platforms. These scenarios are particular true in biology domain where various technologies can be used to produce genome-scale data sets (”-omics” data sets) that provide systems-level measurements for all levels of cellular components in a model organism. In addition, multiple measurements are available from implementations of different platforms within each level of the cellular parties. For example, oligonucleotide microarrays and cDNA arrays are both microarray platforms to measure the gene expression level, and protein microarrays provide measurements over the expression levels of proteins in a cellular system as well as ChIP-chip technique is used to investigate interactions between proteins and DNA in vivo. These multi-scale and multi-origin data yield unprecedented prospective of the cellular internal networking as well as raise remarkable challenges in analysis of these data to recover the central workings which trace the biological information flow from the genome to the ultimate cellular phenotype. Some of these data are prone to introducing technical artifacts. This can bias the data, which can falsely expose sample differences in the absence of a biological cause. In addition, uniform, standardized data representations are seldom adopted, which complicates cross-experiment comparisons. Data quality, context and cross-lab variations represent another important hurdle. Therefore, integrative computational frameworks becomes imperative in which researchers are rising to the challenges by using omics data integration to address fundamental biological questions that would increase our understanding of systems as a whole. Despite these challenges, however, investigators are making progress in identifying, extracting and interpreting biological insights from omics data sets by data integration. Statistical tests are employed in homogeneous data integration which combine single-level measurements from different platforms [1, 80] as well as many efforts have been made to automatically integrate heterogeneous microarray data sets with the prior knowledge on genome-scale, e.g. protein-protein interaction database, ChIP-chip data and promoter motifs for the prediction of protein interactions and gene regulations [52, 54, 104].

These and other discussions on the built-in problems and their solutions in data driven reverse-engineering approaches have invariably revealed one fact, i.e. the remarkable importance of prior information in reverse-engineering methods to prevent overfitting by providing structure and parameter prior distributions and to optimize computational complexity by reducing the heuristic search space, as well as to better recover the underlying network of a system by integrating homo- and heterogeneous multiple-scale and multiple-origin data sets. In this dissertation, we extensively study the precise effects of (qualitative) prior information in probabilistic modeling practices with Bayesian networks. To this end, we will only consider the statistics and uncertainty presented by prior information, i.e. we utilize solely qualitative prior information in our study and therefore, no quantitative data information is available to shield our insights in the function and effects of prior knowledge in probabilistic modeling with Bayesian networks. Due to the fact that human's intuitions and perceptions focus on causation which represent a more fine-tuned relations than simple yes-no binary qualitative relations and fact that the directed connections in the Bayesian networks can be interpreted as causality under proper assumptions, we concentrated on this type of qualitative prior knowledge in our study, i.e. qualitative causal knowledge, which describes the cause-effect relations between multiple entities with any form of uncertainties that can be naturally utilized to represent human understandings and to get approximated characterizations of the behavior of the interested system. The causal knowledge is usually accommodated by textual statements in scientific publications and open access knowledgebase. The basic assumption of our study on probabilistic modeling with prior knowledge is that, in each of the qualitative statement, a group of entities on both ends of the directed connection can be identified and the causality between these entities indicates that the event(s) of the entities at the down-stream of the relation (effects) are regulated by the event(s) of the entities at the up-stream (causes) of the relation. For example, in a qualitative causal statement: *"smoking increases the risk of lung cancer"*, two entities: smoking and lung cancer are related to each other and, smoking positively regulates lung cancer since the risk of lung cancer events is increased by the events of smoking. If the entities are discrete variables with a set of possible values, they are consistent with the discrete Bayesian networks in which directed connections can be quantified by a multinomial table of conditional probabilities and thus, regulations in the prior causal knowledge can be used to specify some properties of this conditional probability table (See section 1.3.1 and 2.1). If the entities are continuous variables, the properties of directed connections can be modeled by certain density function, e.g. Gaussian distribution. Throughout this dissertation, we assume these entities are discrete multinomial variables. Recall the fact that all major built-in problems in data driven reverse engineering approaches can be alleviated and/or resolved by specifying structure and parameter prior distributions. Therefore, our study on probabilistic modeling with qualitative prior knowledge is aimed to investigate the methods in which qualitative prior information can be used to define the distribution functions over the discrete structure space and to specify the density functions over the (eventually) continuous parameter space so that the (in)dependence among the entities of interest can be induced and the directed regulations between these entities can be properly quantified according to the semantics of qualitative prior information, thus, prior network(s) can be construct based on only qualitative information. Following this line, the ultimate goal of our study is to generate well-generalized quantitative inference and reasoning results based on qualitative prior knowledge which then, will be incorporated into the data driven reverse engineering approaches.

From the prospective of engineering, the works in this dissertation can be deemed as an artificial intelligence approach to construct a knowledge-based expert system which supports quantitative conclusions based on qualitative inputs. An expert system is a program composed by a set of rules that analyze information (input by users) about a specific class of problems, as well as providing mathematical analysis of the problems and recommend a course of user action in order to implement corrections. It is a system that utilizes reasoning capabilities to reach conclusions. There are various expert systems in which a "rulebase" and an "inference engine" cooperate to simulate the reasoning process that a human expert pursues in analyzing a problem and arriving at a conclusion. In our works, we set up a set of rules which originate from probability theory and artificial intelligence (AI). This rule set is used to analyze the body of qualitative causal knowledge in many science and industry domains and forms a collection of machine-understandable codes. Then, the expert system recruits a special type of inference engine, i.e. graphical model, to perform inference base on the body of the codes and draw a set of quantitative inferences which support further actions. Other works on expert systems have been proposed over a few years. One representative work is addressed in [87] where, an knowledge-based expert system is designed to make limited and sometimes ambiguous qualitative decisions out of qualitative inputs. Qualitative probabilistic network [109] is used in this system. Knowledge is well-known to represent inconsistent and incomplete information, i.e. knowledge has spatial and temporal properties like other physical systems. The spatial property describes that knowledge represents information on a specific sub-structure of a system and the temporal property states that knowledge represents human understandings at a particular time point. Thus, this knowledge is incomplete and may be updated by complementary discovery. Moreover, knowledge can be inconsistent. In the same domain, there may exist contradicting qualitative statements on dependency, causality over a set of entities. In this dissertation, we also propose and study several methods to handle these properties of qualitative knowledge to reconcile the inconsistent information and to integrate the incomplete information to form a unified framework for probabilistic modeling with qualitative prior knowledge.

1.2 Overview of Data-driven Bayesian Modeling Approach

1.2.1 Bayesian Networks

A Bayesian network (belief network) [82] consists of a graphical structure and a set of conditional probabilities. The graphical structure G is a directed acyclic graph in which nodes represent propositions (or variables) and the edges indicate dependencies between the linked variables of which the strength are quantified by the set of conditional probabilities Θ. If we assume there is a set of ordered variables in the domain of interest, $\mathbf{X} = \{X_1, \dots, X_n\}$, the joint probability distribution over $\mathbf{X}$ can be decomposed into a product of local components as

$$p(\mathbf{X}) = \prod_{i=1}^{n} p(X_i | \pi_i, G, \Theta) \tag{1.1}$$

where π_i denotes the parent nodes of X_i, i.e. π_i is some subset of nodes in $\{X_1, \dots, X_{i-1}\}$ such that X_i and $\{X_1, \dots, X_{i-1}\}$ excluding π_i are conditionally independent given π_i. The primary assumption in Eq. 1.1 is that the order of the variable set is known. This assumption complies with the observations that: i)Human can readily claim causal relationship among variables; and ii) The conditional (in)dependences stem from the causal relationships.

The graphical structure of a Bayesian network is composed by nodes and edges, $G = \{V, E\}$. Each node $X_i \in V$ is a random variable with values x_i and the edges represent the conditional dependencies and independencies among them. The structure of a Bayesian network is defined as a directed acyclic graph (DAG), i.e. there are no directed loops in the network which allow information flow to arrive at the node where they are emitted. Besides the structure of conditional dependence, the strength conditional dependence is encoded in Eq. 1.1, where for every node X_i, the conditional dependence emanate from a set of parent nodes π_i pointing to X_i is quantified by the local probability $P(X_i|\pi_i)$. It states that the state of each variable x_i depends only on the states taken by its parents $x_j \in \pi_i$.

D-Separation Criterion

The various types of conditional dependence and the structure of a Bayesian network can be summarized by d-separation [82]. Consider a set of random variables $\mathbf{X} = \{X_1, X_2, X_3\}$. The two edges connecting the variable pair X_1 and X_2 as well as pair of X_2 and X_3 meet at the intermediate node X_2. According to d-separation criterion, (X_1,X_3) are d-separated, i.e. conditionally independent, if

- X_2 is the midpoint of tail-to-tail or head-to-tail connections and the state of X_2 is known.
- X_2 is the midpoint of head-to-head connections and neither the state of X_2 nor any of its descendants is known.

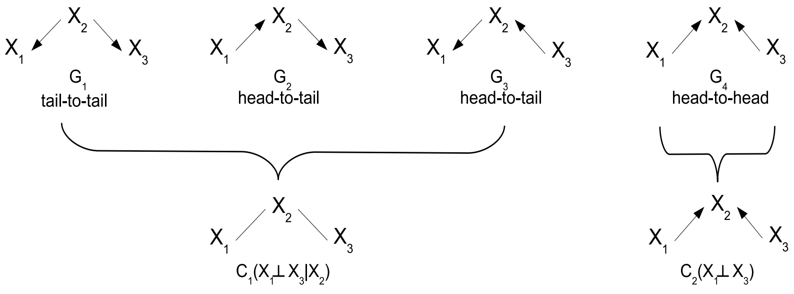


Figure 1.1: D-separation Between X_1,X_2 and X_3 and their equivalent class

Markov Blanket

By applying d-separation criterion on a node X in Bayesian network, we can identify a subset of nodes S in this network so that node X becomes conditionally independent to any other nodes in the network given the nodes in S. This set of nodes is called Markov Blanket of X [82]. Since the states of the parents and children of X evidently give information about this node and its children's parents can be used to explain away X, therefore, a complete Markov blanket of node X consists of its parents, children and the parents of its children. The Markov blanket of a node is important because it identifies all the variables that shield off the node from the rest of the network which means that the Markov blanket of a node is the only knowledge that is needed to predict the state of that node.

Structure Equivalence and Distribution Equivalence

Based on d-separation criterion, we can see that a DAG of a Bayesian network represents the conditional independence encoded in the probability distribution among the set of variables. In Bayesian networks, there are two key concepts: structure equivalence and parameter equivalence [48]. Two Bayesian network are structure equivalent if and only if they have the same set of undirected edges and the same set of collider structures (as shown by G_4 in Fig. 1.1). Two Bayesian network with different structures over $\mathbf{X}$ are distribution equivalent with respect to a family function F if they represent the same joint probability distribution, i.e.

$$p(\mathbf{X}|\Theta_G, G) = p(\mathbf{X}|\Theta_{G'}, G') \tag{1.2}$$

In general, the distribution equivalence with respect to some F implies structure equivalence, but not vice-versa. However, if F is unrestrictive multinomial distribution, structure equivalence also indicates distribution equivalence [48]. It means that if we have two Bayesian network with the same joint probability distribution, their structures must be equivalent. Or on the other words, it is not guaranteed that the conditional dependencies and independencies lead to a unique DAG but instead to many DAGs which altogether describe the same probability distribution equally. This implies that two equivalent DAGs represent the same set of d-separations and therefore also the same probability distribution even though they differ in the direction of some edges. In Figure 1.1, G_1,

G_2 and G_3 present the same d-separation, i.e. X_1 and X_3 are conditionally independent given X_2. Therefore, even though the structures differ in the direction of some edges they are all structure equivalent and distribution equivalence and can be written as $G_1 \sim G_2 \sim G_3$ which indicate that they belong to the same equivalence class $\mathbf{C}_1$. Equivalent structures can be drawn as a partial directed acyclic graph (PDAG) which consists of directed as well as undirected edges. Undirected edges have no direction whereas directed ones are labeled with an irreversible unique direction. The resulting PDAG for the equivalence class C_1 only contains undirected edges, since each edge varies in its direction across the class members. The distribution equivalence across the class of structure equivalence can be demonstrated by using Bayes'rule, the probability distribution of a DAG can be transformed into the distribution of any other member of the same equivalence class, e.g. in Figure 1.1, Eq. 1.2 can be reformulated as

$$\begin{aligned}
C_1: \quad p(X_1,X_2,X_3) &= p(X_1|X_2)p(X_3|X_2)p(X_2) \quad (G_1)\\
&= p(X_2|X_1)p(X_3|X_2)p(X_1) \quad (G_2)\\
&= p(X_1|X_2)p(X_2|X_3)p(X_3) \quad (G_3)\\
\\
C_2: \quad p(X_1,X_2,X_3) &= p(X_2|X_1,X_3)p(X_1)p(X_3) \quad (G_4)
\end{aligned} \tag{1.3}$$

Consequently a Bayesian network model can not necessarily be interpreted as a causal model since putative undirected edges of the corresponding PDAG do not represent causal relationships anymore. For example, in the scenario of genetic regulatory network modeling with a Bayesian network, the edge between molecules in the learned Bayesian network can be reversed in the graph (but not necessarily in the cellular system). Hence for this relationship no unique graphical representation exists and no statement about the causal relationship among these two molecules can be made. In this case, additional prior knowledge is required to assign a direction to the undirected link, i.e. to differentiate the candidate Bayesian networks in the equivalent class. Thus, the problem of structure equivalence can be best addressed by using prior domain knowledge in the learning of Bayesian network. In the following chapters, we present method in which prior cause-effect knowledge can be used to model a Bayesian network.

1.2.2 Bayesian Network Inference

A Bayesian network is a complete model with a graphical structure G over the variables $\mathbf{X}$ and their conditional dependence E, as well as the joint probability Θ. It can be used to answer probabilistic queries about them. For example, the network can be used to find out updated belief of the state of a subset of variables when other variables (the evidence variables) are observed. This process of computing the posterior distribution of variables given evidence is called probabilistic inference. The evidence about recent events or observations is applied to the model by "instantiating" or "clamping" a variable to a state that is consistent with the observation. Then the mathematical mechanics are performed to update the probabilities of all the other variables and their descendants that are connected to the variable representing the new evidence. After inference, the updated probabilities reflect the new levels of belief in all possible outcomes coded in the Bayesian network. The beliefs originally encoded

Algorithm 1: Variable Elimination Algorithm

```
input  : $\mathbf{X}_I$,$\mathbf{X}_E$,$\mathbf{X}_R$,CPDs
output: p($\mathbf{X}_I|\mathbf{X}_E$)
Set the observed variables in all factors to their corresponding observed
values;
while $\boldsymbol{X}_R$ is not Empty do
    Multiply all CPDs has the first variable $Z$ in $\mathbf{X}_R$ and store the results
    in this variable's bulket;
    Sum out the bulket of variable $Z$;
    Remove Z from $\mathbf{X}_R$;
end
Set h=the multiplication of all the factors (h is a function of variables in
$\mathbf{X}_I$ and $\mathbf{X}_E$);
Calculate h/$\sum_{\mathbf{X}_E} h$;
Return
```

in the model are known as prior probabilities, because they are entered before any evidence is known about the situation. The beliefs computed after evidence is entered are known as posterior probabilities, because they reflect the levels of belief computed in light of the new evidence. We can describe the set of variables $\mathbf{X}$ by three subsets: the subset of inquiry variables $\mathbf{X}_I$, the subset of evidence variables $\mathbf{X}_E$ and the rest variables $\mathbf{X}_R$. Thus, the posterior inference of $\mathbf{X}_I$ given $\mathbf{X}_E$ can be calculated as

$$p(\mathbf{X}_I|\mathbf{X}_E) = \frac{\sum_{\mathbf{X}_R} p(\mathbf{X}_I, \mathbf{X}_E, \mathbf{X}_R)}{\sum_{\mathbf{X}_R,\mathbf{X}_I} p(\mathbf{X}_I, \mathbf{X}_E, \mathbf{X}_R)} \tag{1.4}$$

The most common exact inference methods to calculate the values in Eq. 1.4 are variable elimination, which eliminates the $\mathbf{X}_R$ iteratively by distributing the sum over the product; clique tree propagation, which caches the computation so that many variables can be queried at one time and new evidence can be propagated quickly; All of these methods have complexity that is exponential in the network's treewidth. The most common approximate inference algorithms are stochastic Markov Chain Monte Carlo (MCMC) simulation and variational methods.

The most common exact Bayesian network inference approaches are variable elimination algorithm and belief tree propagation algorithm. The variable elimination algorithm sums out variables from a list of factors one by one and the factors are conditional probability distributions in Bayesian network. The VE algorithm can be shown in Algorithm 1. The time to answer any query is exponential in the size (number of terms) in the largest factor (table) that is encountered. The factors come from the original graph (CPDs or potentials), but new factors are created in the process of summing out. The order in which we perform the summation can have a large impact on the size of the intermediate factors. The exact inference in discrete graphical models is NP-hard. A more efficient algorithm for performing exact inference in tree-structured belief network is called message-propagation algorithm [81]. If the network is singly connected, then probabilities can be updated by local propagation in a isomorphic network

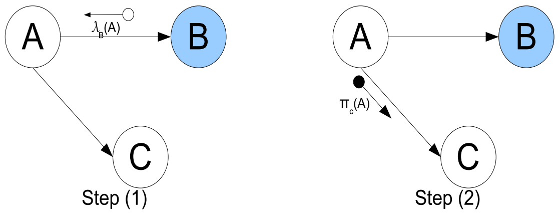


Figure 1.2: Belief Propagation Scheme in Tree-structured Network

of parallel and autonomous processes and the impact of new information can be imparted to all variables in time proportional to the longest path in the network. If the network is multiple connected, the network is required to be transferred into tree structure by introducing "dummy variables" which group together the multiple-connected components and then, we can perform the message propagation in the tree. As shown in Fig. 1.2, two types of messages are transmitted between any pair of variables in the tree, i.e. $\lambda_{ch}(pa)$ and $\pi_{ch}(pa)$ to update the local information stored in each of the node. The messages can be calculated as

$$\lambda_{ch}(pa) = \sum_{ch} \lambda(ch) p(ch|pa) \tag{1.5}$$

and

$$\pi_{ch}(pa) = \alpha \pi_{pa} \prod_{sib} \lambda_{(} sib)(pa) \tag{1.6}$$

where ch indicate child node, pa denotes parent node and sib is the siblings of a node. α is the normalizing factor. These messages are used to update the pre-stored π and λ information in the nodes as

$$\lambda_{ch}(pa) = \lambda(pa) \tag{1.7}$$

and

$$\pi_{ch} = \sum_{pa} p(ch|pa) \pi_{ch}(pa) \tag{1.8}$$

1.2.3 Bayesian Network Learning

A Bayesian network encodes the conditional dependence and independence with the graphical structure G and a set of conditional probability parameters Θ over a set of random variables $\mathbf{X}$. Given a set of training data $\mathbf{D}$, we consider learning a Bayesian network under several settings. First, the network structure might be given in advance or it might have to be inferred from the training data. Second, all the network variables are directly observed in each training example or some might be unobservable. In the case where the network structure is given in advance and the variables are fully observed in the training examples, it is straightforward to learn the conditional probability table. We can estimate the conditional probability table entries by statistical counting as in Naive Bayesian classifier. In case where the network structure is given, but subset of variable values are not observed in the training data. Expectation-Maximization(EM) algorithm [25] can be used to calculate the expect value of statistics presented in the data and estimate the conditional probability table entries by maximizing this expect value. EM algorithm, like gradient descent, finds local maxima on the likelihood surface defined by the network parameters [66, 67]. Russell [95] proposed a gradient ascent procedure searches through the space of hypotheses that corresponds to the set of all possible entries for the conditional probability tables. The objective function $p(\mathbf{D}|h)$ is maximized during the gradient ascent given the hypothesis h. By definition, this corresponds to searching for the maximum likelihood hypothesis for the table entries. Learning Bayesian networks when the structure is unknown can be performed by maximizing some statistically motivated score function [20] to describes how fit each possible structure is to the data. The task of learner is then to find a structure that maximizes the score. Since the structure space is superexponential to the number of the variables in the data, it is an NP-hard problem [18], and thus we need to resort to heuristic methods to select single "good" model or multiple "good" models. In the latter case, model averaging method with Monte Carlo method [50] can be used to generate an efficient and better prediction results. If all the variable values are observed, heuristical search algorithm can be performed to calculate the score metric function for each possible model. If some of the variable values are unobservable, Friedman proposed a powerful iterative maximization algorithm, Structural Expectation-Maximization (SEM) algorithm, to calculate the expected statistics for each possible structure [37]. In this section, we focus on learning Bayesian network structure, especially in case when some of the variable values are unobserved. Namely, we introduce Structural EM algorithm.

Preliminary

The procedure of structural learning can be best described as: Given a data set with N independent examples, $\mathbf{D}=\{D^1, ..., D^N\}$, where each data example is an n-dimensional vector with components $D^l=(d^l_1, \ldots, d^l_n)$ and the element d^l_n of D^l indicates the value of *n-th* variable in the *l-th* data example, we are asked to find a graph structure G and a parameter set Θ that best fit to $\mathbf{D}$. For example, in the context of microarray data analysis, there are N independent microarray experiments, each observing the expression states of n probes or genes. Each node in the learned Bayesian network symbolizes a specific probe or gene and the structure represents the conditional dependency relationships among these molecules regarding the cellular conditions from which microarray samples where taken.

A statistically motivated scoring function S assigns a score $S(G|D)$ to measure the fitness of a graph G with respect to the data. In the following, data are supposed to be multinomial which is consistent with the multinomial nature of the Bayesian network. Meanwhile, we assume the data is exchangeable, i.e. the data sequence obtained by interchanging any two observations in the sequence has the same probability as the original sequence. Interchangeability indicate that the process generating the data do not change in time. Assume that we have n variables, $\mathbf{X}=\{X_1, \ldots, X_n\}$. When each variable X_i can assume r_i different values k and the set of parents pa_i can assume q_i different values j, the local multinomial conditional probability distribution can be represented as a

$r_i \times q_i$ table. Each parameter entry in the table is given by

$$p(X_i|pa_i, G) = \theta_{ijk} \tag{1.9}$$

where θ_{ijk} satisfy the conditions that $0 \leq \theta_{ijk} \leq 1$ and $\sum_{k=1}^{r_i} \theta_{ijk}=1$. The value of θ_{ijk} can be estimated by the frequency of variable X_i takes the value k given the parents pa_i takes value j in the data set $\mathbf{D}$, i.e. $\theta_{ijk}=N_{ijk}/N_{ij}$ where N_{ijk} is the number of the cases d_i^l=k and $pa_i(d^l)$=j. $N_{ij}=\sum_k N_{ijk}$.

Score Function

Frequentist Score

The classical approach for learning is the likelihood maximization. Frequentist deem Θ as unknown parameter and consider the data set $\mathbf{D}$ as represents the joint probability distribution. In this approach, finding the best fit structure G boils down to the problem of finding the structure with the highest likelihood assignment. We have the logarithm of likelihood as

$$\log p(\mathbf{D}|\Theta, G) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \theta_{ijk} \tag{1.10}$$

One problem of using likelihood score function is the problem of overfitting. The learner tends to output a complete since it has the largest number of parameters to fit to the data best. Thus, it is necessary to specify a prior distribution over the discrete structure model space in the objective function to prevent the overfitting of learned structures to the data. Thus, we can re-write the Eq.1.10 as

$$S(G, D, \Theta) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \log \theta_{ijk} - p \tag{1.11}$$

For structure priors, when there is little prior knowledge, a well-principled way to avoid this kind of overfitting is to impose a prior on models which is a penalty function of the number of edges and the size of families to discourage networks that are globally or locally dense. For example, the *Bayesian Information Criterion*(BIC) score [42] calculate the penalty term p as

$$p = \frac{1}{2} |\Theta| N \tag{1.12}$$

where N is the number of samples. Another prominent frequentist scoring function is the *Minimum Description Length*(MDL) which is identical to the negative BIC score but with a origin from coding theory [64]. A comparison of these model selection criteria can be found in [106]. Even though maximum likelihood score is quite useful for structure learning it has to be noted that the maximum likelihood estimate converges to the real joint probability value only when $N \to \infty$, whereas for small N the maximum likelihood score function produces biased results [4].

Bayesian Score

The Bayesian method, rather tries to calculate the most probable structure given the data which is equivalent to weight the models with an a priori distribution.

In addition, the Bayesians deem the parameters as random variables instead of unknown variables. Therefore, the Bayesian score is proportional to the posterior probability of model structure given the data

$$S(G|D) = \frac{p(D|G)p(G)}{p(D)} \tag{1.13}$$

where $p(G)$ is the prior probability of the model structure, $p(D)$ is a normalization constant, and $p(D|G)$ is the marginal likelihood of the data given the structure G. Since in the Bayesian score treats the parameters of a model as random variables characterized by a distribution. This uncertainty over the parameters is expressed by marginalizing the parameters. Thus, the marginal likelihood equals the integral

$$p(D|G) = \int p(D|G,\Theta)p(\Theta|G)d\Theta \tag{1.14}$$

where $p(\Theta|G)$ denotes the prior distribution of the model parameters Θ for a given structure G and $p(D|\Theta, G)$ is the likelihood of the data given a Bayesian network.

If we assume the Bayesian model is multinomial in nature and a conjugate family prior for the model parameter density indicates that $p(\Theta|G)$ follows a Dirichlet distribution. Further, if we assume the nature of our domain satisfy the assumptions such as data completeness, parameter independence and parameter modularity [21], Equation 1.14 can be formulated in closed form (see Appendix A for a more detailed description). The solution of Eq. 1.14 with Dirichlet priori density function outputs an unique scoring function, the *Bayesian Dirichlet*(BD) score [49]. The analytical solution for the BD score can be written as

$$p(D|G) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \tag{1.15}$$

where N'_{ijk} are the parameters of the Dirichlet prior distribution which reflect the prior knowledge on how many times we have observed X_i=k and pa_i=j in the past. N_{ij}=$\sum kN'_{ijk}$ and Γ in the gamma function.

Therefore, the posterior probability of the model structure G given the data set D with BD score can be finalize as

$$p(G|D) = \frac{p(G)}{p(D)}\prod_{i=1}^{n}\prod_{j=1}^{q_i}\frac{\Gamma(N'_{ij})}{\Gamma(N'_{ij}+N_{ij})}\prod_{k=1}^{r_i}\frac{\Gamma(N'_{ijk}+N_{ijk})}{\Gamma(N'_{ijk})} \tag{1.16}$$

There are some open questions in the Eq. 1.16. First, we equally treat each candidate graph in a structure and distribution equivalence class as shown by Eq. 1.3 and the graphs are indistinguishable. In this case, the Bayesian score in Equation 1.16 has to assume identical values for all members of a certain class to ensure score equivalence. This can be realized by calculating N'_{ijk} as

$$N'_{ijk} = N_0 p(X_i = k, pa_i = j) \tag{1.17}$$

where $p(X_i = k, pa_i = j)$ is the marginal prior over the joint states as this ensures the likelihood equivalence property of network structure [49]. In case where insufficient prior knowledge is available, p is often chosen to be uniform as $p(X_i, pa_i)$=1/($|X_i|\,|pa_i|$). The parameter of Dirichlet distribution function N_0 can be set independently. This is called BD equivalent score (BDe). However, in the case, where vast amount of prior knowledge is available, this problem can be resolved by modeling the structure prior $p(X_i = k, pa_i = j)$ according to the prior knowledge which can regularize the BD score to differentiate the structures in an equivalent class. Meanwhile, some of the conditional independences between $\mathbf{X}$ and their directions of the influential relationship can be pre-specified based on the prior knowledge on model structures $p(G)$ as it can reduce the heuristic search space. Second, as in the frequentist approach, the problem of overfitting needs to be addressed by punishing networks complexity. For this, the structure prior $p(G)$ can be used as a penalty term such that the prior probability of a structure G decreases with the structure complexity. The number of edges can be used as a measure for the complexity of structure G such that the structure prior is given as [49]

$$p(G) = ck^{\delta} \tag{1.18}$$

where k is a constant factor $0 < k < 1$ and c is a normalization constant. Note that $p(G)$ used in Eq. 1.18 does not reflect the true prior knowledge, thus can be replaced by those expressions constructed by prior knowledge. In the following chapters, we will study these issues on how to model the $p(X_i = k, pa_i = j)$ and $p(G)$ based on qualitative prior knowledge.

Heuristic Search

The problem of searching for the optimal structure of a Bayesian network is NP-hard [18], thus, we resort to employ heuristic search strategies which can efficiently select the candidate model structures.

Score Decomposability If the data set is complete without either missing values nor hidden variables, the score function in Eq. 1.16 can be decomposed into a set of local scores just like the conditional probability distribution of a Bayesian network decomposes into a product of local probabilities as Eq. 1.1. The score decomposability can be expressed as

$$S(G|D) = \prod_{i=1}^{n} S_i(G|D) \tag{1.19}$$

$S_i(G|D)$ denotes the local score of variable X_i which only depends on its states and the states of its parents pa_i.

According to Eq. 1.19, scoring function can be decomposed into a product of local scoring functions restricted to each family (a variable X_i and its parents pa_i). Each term can be defined as the local score of X_i which depends only on the state of X_i and pa_i. This decomposition property enables a local search procedure which changes one edge at each move and this search strategy can efficiently evaluate the gains made by this change. This implies that at each step only the local scores of those variables whose set of parents has been changed needs to be re-estimated. For a structure G, the structures which differ only in the presence or absence of one edge and satisfy the acyclicity condition (using depth-first search), represent the so-called neighboring structures $nbg(G)$.

Algorithm 2: Procedure for General SEM Algorithm

```
while Loop $n \leq MaxLoop$ or No converge do
    Compute the posterior probability of parameter $p(\Theta^n|G_n, O)$;
    E step: For each $G$ of the neighborhood of $G_n$, compute the expected score;
    Q(G,$G_n$)=$E[\log p(H, O, G_n)|G_n, O]=\sum_{h \in H} p(h|O, G_n) \log p(h, O, G)$;
    M step: Choose $G_{n+1}$=G that maximize Q(G,$G_n$);
    if Q($G_n$,$G_n$)=Q($G_{n+1}$,$G_n$), then;
    return $G_n$
end
```

The structure $G' \in nbg(G)$ which entails the best scoring function is selected as the next candidate structure G_0. This technique is known as local search strategy and a commonly used approach in structure learning. If G_0 is selected as the next structure G depends on the heuristic which used. *Greedy Search* is the simplest heuristic strategy of this kind strategy. In each iteration the space of neighboring structures is visited for an improvement in the score. The neighboring structure G_0 which provides the largest improvement with respect to G becomes the next intermediate structure which is the starting point for the next iteration. Thus, the algorithm always moves across the model space in the direction of the greatest rate of decrease of the error which is quite similar to the gradient descent algorithm for training neural networks. However, a big drawback of this search strategy is that instead of finding the global optimum it might stop at a local optimum. Therefore, a more advanced searching algorithm, such as simulated annealing which an adaptation of the Metropolis-Hastings algorithm [71], are often utilized to avoid this problem. However, since search strategy is not the focus of our study, due to the simplicity, we employ Greedy search in the following studies.

Structural Expectation Maximization (SEM)

One of the hard problems in learning Bayesian network from data raise when the data is incomplete. As outlined above, if the data satisfy the assumption of completeness, the data likelihood given the model $p(D|G, \Theta)$ and the parameter posterior probability $p(\Theta|G, D)$ are decomposable (see Appendix A). However, this is not the case when data set is incomplete, e.g. with missing values or hidden variables. In this case, the posterior probability of parameter is no longer a product of independent terms. For the same reason the marginal data likelihood given the model structure $p(D|G)$ becomes no longer a product of terms. Therefore, we need to directly evaluate the integral in Eq. 1.14. One approach is that, we can find an approximation to the marginal data likelihood with *Maximum A Posterior* parameters [66], i.e.

$$p(O, H|G) = p(D|G) \approx p(D|G, \hat{\Theta}_{MAP}) \tag{1.20}$$

where $\hat{\Theta}_{MAP}$=argmax$\{p(D|\Theta, G)p(\Theta|G)\}$. This approach involves extensive computational complexity since it requires us to find the MAP parameters with EM algorithm for each model we consider before we could score it. When we search in a large space of possible models, this approach become infeasible since the procedure has to invest a tremendous amount of computation before making a single local change in the model structure.

In contrast, Structural EM algorithm attempts to directly optimize the Bayesian score in Eq. 1.13 rather than an asymptotic approximation based on the fact that it is sufficient to find a MAP model by maximizing $p(D|G)p(G)$ since the normalizing term $p(D)$ is the same for all the models we compare. The main idea of Structural EM algorithm is that we can compute and estimate the *expected Bayesian score* and maximize it over iterations instead of the actual score [37]. If we assume that the data set $\mathbf{D}$ consists of observed values $\mathbf{O}$ and hidden or missing values $\mathbf{H}$, the complete data likelihood can be expressed as

$$p(D, G) = p(H, O, G) \approx p(D, G, \hat{\Theta}_{MAP}) \tag{1.21}$$

and the logrithm of expected score function can be described as $E[\log p(H, O, G)]$. The key insight of SEM is that the complete expected data likelihood and hence the expected Bayesian score is decomposable as in the case of complete data. The general outline of SEM algorithm is shown in Algorithm. 2 and can be summarized as follows: Initially, we start with an arbitrary graph G_0 and a guess Θ_0, at *n-th* step, we attempt to learn the MAP parameters under current model given $\mathbf{O}$ and graph G_n to improve the parameters and use the MAP parameters to complete the data by inferencing the missing values in the data, i.e. to compute $p(H|O, \hat{\Theta}_n)$. In this way, we could compute the *expected sufficient statistics* (ESS) for all the models in the neighborhood of current graph, $G'_n \in nbg(G_n)$ as in the complete data case. Then, we select the neighboring graph with the highest expected score $E[p(H, O, G'_n|O, \hat{\Theta}_n)]$ as the structure for *(n+1)-th* iteration, i.e. $G'_n \rightarrow G_{n+1}$. Iteration stops until there is no gain in the score.

1.2.4 Bayesian Model Averaging

Bayesian Model Averaging (BMA) is a technique designed to help account for the uncertainty inherent in the model selection process, something which traditional statistical analysis often neglects. By averaging over many different competing models, BMA incorporates model uncertainty into conclusions about parameters and prediction. Model uncertainty raise from the randomness of a domain. In the practice of learning Bayesian networks from sparse data, it is often true that multiple models with different structures and parameters can fit to the data well. However, the common Bayesian approach-model selection which using greedy local search algorithm, often selects single "good" model and ignore the model uncertainty which is reflected by the posterior probability of models given the data. Full Bayesian approach with Bayesian model averaging (BMA) can be used to compute a posterior probability distribution over all possible models to reflect the true model uncertainty [50, 76]. This approach computes an average of the quantity of interest under each of the models considered, weighted by their posterior model probability. If we assume the data set is D and the quantity of interest is Δ, the posterior probability of Δ given the data is:

$$p(\Delta|D) = \sum_G p(\Delta|G)p(G|D) \tag{1.22}$$

where $p(G|D)$ is the posterior probability of model structure G given the data and $p(\Delta|G)$ indicates the posterior probability of Δ under each model. With

Bayes' rule, the model posterior probability given data can be re-written as

$$p(G|D) = \frac{p(D|G)p(G)}{\sum_G p(D|G')p(G')} \tag{1.23}$$

where

$$p(D|G) = \int_\Theta p(D|G,\Theta)p(\Theta|G)d\Theta \tag{1.24}$$

Thus, Bayesian model averaging provides a coherent mechanism for accounting uncertainty in model selection and avoids over-confident inferences and decisions. In fact, it has been demonstrated that averaging over all the models in this fashion provides better average predictive power, as measured by a logarithmic scoring rule, than using any single model [50, 69]. Although BMA is an attractive solution to the problem of accounting for model uncertainty, there are several practical problems in the implementation of BMA [50].

1. How do we define the prior distribution $p(G)$ and $p(\Theta|G)$?
2. How do we perform the integration over model parameters in Eq. 1.24?
3. How do we marginalize out the hidden variables?
4. How do we perform the summation over model structures in Eq. 1.22?
5. How do we choose the class of models over which to perform averaging?

The first three problems appear also in Bayesian model selection and the last two problems are specific to BMA. With respect to problem 4), the search space is superexponential to the number of variables in the network. To avoid this intractability, it is proposed to use Markov Chain Monte Carlo technique to sample the possible model structures with Metropolis-Hastings (MH) algorithm to resolve the problem. We construct a Markov chain whose state space is all the DAGs and the stationary distribution is the posterior probability $p(G|D)$ [76]. With regarding to the problem 2), when the data is complete and the domain satisfy the five assumptions (See appendix A), the integration of parameters in Eq. 1.24 can be resolved analytically and in the case of incomplete data (problem 3)), the data likelihood surface becomes multimodal, and we have to use iterative methods, such as EM algorithm [25, 67], to find a local maximum of the ML/MAP function. These algorithms need to use an inference algorithm to compute the expected sufficient statistics (or related quantity) for each node. To address problem 5), Draper [27] suggested finding a good model and then averaging over an expanded class of models "near" the good model. Although the problems 2)-5) have been attracting the attentions of research over years, few research have been done in specifying the prior distribution $p(G)$ and $p(\Theta|G)$. In the following chapters, we propose a method which makes use of a qualitative knowledge model to introduce a set of constraints in the hyperspace of the parameters and structures. We show that this method can well model the model uncertainty and produce accurate quantitative predictions based on BMA.

1.3 Overview of Knowledge-based Bayesian Modeling Approach

As outlined in the above sections, a probabilistic network, such as Belief network, consists of a graphical representation of the variables in a domain of interest, and the conditional dependence and independence between the variables. The joint probability distribution is encoded in the probabilistic network by a set of conditional probability distributions of each node given their parents. By Bayesian network inference methods, e.g. message-propagation method, any quantity of interest can be inferred and reasoned in the belief network. Therefore, it is particular important to construct the network automatically. In Section 1.2, Bayesian network learning algorithms have been introduced as to induce Bayesian networks from the data. This is called reverse-engineering approach and we call it *Bottom-up* approach. This approach is the most attractive field of machine learning in that it can automatically analyze the observed data and infer a network structure with parameters to specify the qualitative information and quantitative information among the variables in the domain of interest. The qualitative information represent by a graph model is the (in)dependence and dependence between the variables and the quantitative information indicates the strength (in probability) of the directed influence from parental variables to the child. One prominent problem in the learning with sparse data is that the learned model is often overfitting to the data. In this case, we need to use prior knowledge and full Bayesian method to alleviate and/or avoid the overfitting.

One the other hand, in many scientific and industrial domain, vast amount of qualitative priori knowledge exist in various type of resources, such as scientific publications, domain experts and open access knowledgebase. All these kinds of priori knowledge tend to provide the qualitative and quantitative information on the relationships among the variables in the domain. One kind of such prior knowledge is particular relevant to probabilistic modeling with Belief networks-qualitative causal knowledge (QCK). There are several reasons for its significance: i)Qualitative knowledge is the most accessible, manageable and intuitive information that human reasoning, communicating, and storing in their daily life. ii)Human's intuitions and perceptions focus on causation which represent a more fine-tuned relations than Boolean qualitative relations and iii)The directed connections in the Bayesian networks can be interpreted as causality by nature under proper assumptions. The prominence of this practice is two-fold: Firstly, we can utilize the existing qualitative causal knowledge to construct a comprehensive network under a domain and new knowledge can be discovered by inference and reasoning. Secondly, the probabilistic modeling framework with qualitative causal knowledge can provide a way to specify the prior distribution over Bayesian network structure and parameter space, $p(G)$ and $p(\Theta|G)$, in the reverse-engineering approach.

Therefore, it is particular important to make use of the QCK to automatically construct Bayesian networks. Unfortunately, works in this area have not attracted enough attentions. Some previous works [28, 88, 89] have been proposed to use the qualitative causal knowledge to construct Bayesian network. However, these methods operate at qualitative level, i.e. only qualitative part of a Bayesian network is specified and no quantitative inference are performed in the cases. These framework make use of a knowledge model-qualitative probabilistic network [109] to represent and perform qualitative inference. Due to the lack of the abilities in these approaches to handle the uncertainty in qualitative knowledge, ambiguous and nonsense inference results are often produced. Other works [30,87] try to translate the expert knowledge into numbers which may be a suspicious due to the relative large bias in this method and impractical due to its cost. So, it is imperative to develop a probabilistic modeling framework in which, qualitative information are augmented to produce quantitative inference in an consistent, robust and automatic manner. In this thesis, we developed such framework.

1.3.1 Qualitative Probabilistic Network

Qualitative probabilistic networks was proposed as qualitative abstractions of probabilistic networks [109]. A qualitative probabilistic network encodes a quantitative information which abstract from numeric representation of the probabilistic networks, i.e. (in)dependence and direction of the influence, among a set of variables into an acyclic graph. Instead of numerical joint probabilities, a qualitative probabilistic network associates with its digraph qualitative probabilistic relationships with inequality constraints. Two types of qualitative relationship are defined. Each of them is a probabilistic form of monotonicity constraint over a group of variables. *Qualitative influence* describe the direction of the relationship between two variables. *Qualitative synergies* describe interactions among the influences. Qualitative probabilistic networks use signs to represent the uncertainty of a domain and support an efficient, yet ambiguous, sign-propagation algorithms to justify a reduced form of relative likelihood that imply useful decision-making properties.

Formally, a qualitative probabilistic network can be defined as an acyclic graph, $G=\{V,Q\}$. V represents a set of variables in the domain of interest and Q is a set of edges which describes the independences between the variables in V captured by the d-separation criterion in Section 1.2. Besides the digraph G, a qualitative probabilistic network, includes a set of hyper-arcs indicating the qualitative probabilistic relationships between variables instead of a set of numerical conditional probability distributions in Belief network. There are three types of qualitative probabilistic relationship: qualitative influences, additive synergies and product synergies.

Qualitative Influence

A qualitative influence relationship describes how the values of one node influence the probabilities of the values of the other node. For example, a positive qualitative influence of a node A on its child B expresses that observing higher values for node A makes higher values for node B more likely, regardless of any other direct influences on B, where the concept of "higher" refers to the order on a node's values. The hyperarc of qualitative influence edge, i.e. the sign of qualitative influence, can be expressed as

$$S^{\delta}(A,B,G), \delta \in \{+,-,0,?\} \tag{1.25}$$

which denotes the assertion that a qualitative influence of A on B with sign δ holds in the graph $G=\{V,Q\}$. "+" means a positive qualitative influence, "-" indicates a negative qualitative influence and "0" represents that A and B are virtually independent. "?" denote the situation in which the qualitative relationship between A and B are unknown or ambiguous.

Binary S^+ We say node A positive influence node B stochastically and write $S^+(A, B, G)$ if and only if for all $X \in \mathbf{X}(pred_G(B) - A)$ such that X is consistent with A and $\overline{A}$,

$$P(B|A, X) > P(B|\overline{A}, X) \tag{1.26}$$

Binary S^0 We say node A and node B are independent and write $S^0(A, B, G)$ if and only if for all $X \in \mathbf{X}(pred_G(B) - A)$ such that X is consistent with A and $\overline{A}$,

$$P(B|A, X) = P(B|\overline{A}, X) \tag{1.27}$$

Binary S^- We say node A negative influence node B stochastically and write $S^-(A, B, G)$ if and only if for all $X \in \mathbf{X}(pred_G(B) - A)$ such that X is consistent with A and $\overline{A}$,

$$P(B|A, X) < P(B|\overline{A}, X) \tag{1.28}$$

where $X \in \mathbf{X}(pred_G(B) - A)$ denotes the parents of node B other than A. We could rule out the independence case and restrict to the defintion of S^+ and S^- by replacing the inequality operator $>$ with $\geq$ in Eq. 1.26 and the operator $<$ with $\leq$ in Eq. 1.28.

The intuitive formula of S^δ in Eq. 1.26 to 1.28 is not straightforward in case where the variables in G can take more than two values. For example, in the definition of positive qualitative influence, we want to make assertion that the probability distribution for node B moves towards higher values when A increases. To make such statement, *(Conditional) Cumulative Probability Distribution Function* over B is used to capture the semantics of "higher order".

Conditional Cumulative Probability Distribution Function Let A and B are two random variables in $\mathbf{V}$. Let P be the probability density function and let B's values be $\{b_1 < \ldots < b_n\}$, $n \geq 1$. Then, the function conditional cumulative probability distribution function over B can be defined as

$$\begin{aligned} F_{B|A}(b_i) &= P(b_1 \vee b_2 \vee \ldots \vee b_i|A) \\ &= \sum_{b=b_1}^{b_i} P(B = b|A) \end{aligned} \tag{1.29}$$

An ordering criterion on the values can be *First-order Stochastic Dominance* (FSD).

First-order Stochastic Dominance FSD holds for (conditional) cumulative distribution function F and F', i.e. F FSD F', if and only if for any given value b_i of B, the cumulative function F is less than the cumulative function F'.

$$F\ FSD\ F' \Leftrightarrow F(b_i) \leq F'(b_i) \tag{1.30}$$

Thus, the statement: Higher values for a node B are more likely given the higher values for a node A if and only if the cumulative conditional probability distribution $F_{B|a_i}$ FSD $F_{B|a_j}$ for all values $a_i > a_j$ of A. Namely, the qualitative influence can be defined alternatively by the cumulative probability function in case of multinomial distribution as

Multinomial S^+ Let $F_{B|a_i}(b)$ be the conditional CPD function of node B given A's value a_i and let $F_{B|a_j}(b)$ be the conditional CPD function of node B given A's value a_j, then $S^+(A, B, G)$ is

$$\forall a_i, a_j, a_i > a_j \ F_{B|a_i}(b) \ FSD \ F_{B|a_j}(b) \tag{1.31}$$

The S^- can be defined similarly as

Multinomial S^- Let $F_{B|a_i}(b)$ be the conditional CPD function of node B given A's value a_i and let $F_{B|a_j}(b)$ be the conditional CPD function of node B given A's value a_j, then $S^-(A, B, G)$ is

$$\forall a_i, a_j, a_i > a_j \ F_{B|a_j}(b) \ FSD \ F_{B|a_i}(b) \tag{1.32}$$

Qualitative Synergy

Besides qualitative influences, the hyper-arcs of a qualitative probabilistic network G=$\{V, Q\}$ entail synergies. A synergy describes an interaction among the influences from two or more parents to a third child, i.e. a collider structure in a digraph. There are two types of synergistic interaction, captured by *positive qualitative synergy* and *negative qualitative synergy.* As qualitative influences are directed edges augmented by signs, a qualitative synergy assertion that the variables in set $T \subset V$ is synergistic in direction σ on variable C is written as $Y^\sigma(T, C, G)$.

$$Y^\sigma(T, C, G), \sigma \in \{++, --\} \tag{1.33}$$

where "++" means a positive additive synergy, "$--$" indicates a negative additive synergy and "$\times$" represents productive synergy. A qualitative synergy expresses how the values of two nodes jointly influence the probabilities of the values of a third node.

Binary Positive Synergy ++ Let A, B and C∈V be nodes in G with edges A→C, B→C. Let X=$Pred_G(C)/\{A, B\}$. Then, nodes A and B exhibit a positive synergy on node C, written $Y_G^{++}(A, B, C)$ iff

$$\forall X, P(C|A, B, X) + P(C|\overline{A}, \overline{B}, X) \geq P(C|A, \overline{B}, X) + P(C|\overline{A}, B, X) \tag{1.34}$$

A positive synergy of nodes A and B on a common child C expresses that the joint influence of A and B on C is greater than the sum of their separate influences, regardless of any other direct influences on C.

Binary Negative Synergy $--$ Let A, B and C∈V be nodes in G with edges A→C, B→C. Let X=$Pred_G(C)/\{A, B\}$. Then, nodes A and B exhibit a negative synergy on node C, written $Y_G^{--}(A, B, C)$ iff

$$\forall X, P(C|A, B, X) + P(C|\overline{A}, \overline{B}, X) \leq P(C|A, \overline{B}, X) + P(C|\overline{A}, B, X) \tag{1.35}$$

If the inequality operator in Eq. 1.35 is replaced by "=", then node A and node B exhibit zero synergy, i.e. no synergy between the influenced of A and C on C. The formula of Y^{σ} in Eq. 1.34 to 1.35 is not straightforward in case where the variables in G can take more than two values. In this case, cumulative probability function can be used alternatively as

Multinomial Positive Synergy ++ Let $F_{C|A,B}$ be the cumulative probability distribution function of node C given node A and B. Let $X{=}Pred_G(C)/\{A,B\}$. Then, nodes A and B exhibit a positive synergy on node C, written $Y_G^{++}(A,B,C)$ iff

$$\begin{gathered}\forall a_1, a_2, b_1, b_2, c_0, X, a_1 > a_2, b_1 > b_2 \\ F_{C|a_1 b_1 X}(c_0) + F_{C|a_2 b_2 X}(c_0) \leq F_{C|a_1 b_2 X}(c_0) + F_{C|a_2 b_1 X}(c_0)\end{gathered} \tag{1.36}$$

Multinomial Negative Synergy −− Let $F_{C|A,B}$ be the cumulative probability distribution function of node C given node A and B. Let $X{=}Pred_G(C)/\{A,B\}$. Then, nodes A and B exhibit a negative synergy on node C, written $Y_G^{++}(A,B,C)$ iff

$$\begin{gathered}\forall a_1, a_2, b_1, b_2, c_0, X, a_1 > a_2, b_1 > b_2 \\ F_{C|a_1 b_1 X}(c_0) + F_{C|a_2 b_2 X}(c_0) \geq F_{C|a_1 b_2 X}(c_0) + F_{C|a_2 b_1 X}(c_0)\end{gathered} \tag{1.37}$$

1.4 Summary and Outline

In this thesis, we focus on the discussion about probabilistic modeling with Bayesian network in the knowledge-driven approach. Being creative, we aim to solve the difficult and long-standing problems in data-driven Bayesian approach, namely, how to make use of qualitative prior knowledge (hypothesis) to infer Bayesian network quantitatively which supports quantitative prediction and reasoning and to bridge the gap between qualitative prior hypotheses and quantitative probabilistic representation of the Bayesian networks.

In Chapter 2, we formally propose our method to infer Bayesian network quantitatively from the qualitative knowledge. We do so by constructing a set of consistent qualitative knowledge model and use it to translate the qualitative knowledge into a class of constrained Bayesian networks. This class of Bayesian networks is used in making probabilistic prediction and reasoning. We investigated the robustness of our approach in case there is noise in the knowledge and we proposed several approximation schemes to compute the reasoning. In chapter 3, we extend our approach to model the Bayesian network from a set of inconsistent knowledge. We reconcile the contradicting information by hierarchically formalize the qualitative knowledge model and calculating a prior distribution for each inconsistent knowledge component, i.e. each class of Bayesian networks. In this case, quantitative inference and prediction are computed over all ground Bayesian networks in all classes weighted by their priors. In Chapter 4, we investigate the methods to integrate series of incomplete qualitative knowledge. Due to the fact that our method not only infer the Bayesian model structure but also the probability configurations, the qualitative knowledge integration problem are projected to Bayesian network fusion in a quantitative manner. In this case, Bayesian model fusion under two scenarios is studied, namely, fusion in the parameter space given the structure space and fusion in the parameter space as well as in the structure space.

Chapter 2

Bayesian Modeling with Consistent Qualitative Knowledge

Bayesian reasoning provides a probabilistic approach to inference. In Bayesian framework, quantities of interest are described by probabilities and optimal decisions can be made by reasoning about these probabilities together with the observation or evidence. Bayesian reasoning is important to machine learning because it provides a quantitative approach to weighting the evidence supporting alternative hypotheses. Numerous algorithms have been proposed for learning the Bayesian network structure and parameter from observed data. These algorithms produce a single Bayesian model by maximizing its probability given the training data, i.e. maximum a posterior approximation. In realistic problem, learning Bayesian model by training data requires relative large amount of observed data comparing to the size of network. However, the data basis is often very sparse and it is hardly sufficient to select one adequate model due to the model uncertainty, thus, selecting a single model may induce overfitting to the data and can lead to strongly biased inference results. It is therefore preferable to adopt a full Bayesian approach with model averaging. In contrast to training data, in almost every science and industry domain, there exists an enormous amount of qualitative knowledge which describes the entities and their relationships inexplicitly. This is particular true in biomedicine domain. For example, in the statement:" The risk of lung cancer among smokers is approximate 10 times higher than non-smokers", the qualitative knowledge can be extracted as two entities: smoking, lung cancer and their causal relation, smoking causes lung cancer, as well as some properties which further specify the qualitative relationship, i.e. smoking causes lung cancer by 10 times. Comparing to quantitative experimental data, qualitative knowledge which exist in large amount could be used to construct the priori distribution over the structure and parameter space and thus is able to prevent overfitting. Moreover, model uncertainty is encoded in the nature of qualitative knowledge which enables full Bayesian approach with model averaging. In this section, we propose a novel framework of modeling Bayesian networks and performing quantitative inference with model averaging base on solely qualitative statements. Our method translates the qualitative statement into a set of constraints on the model structure and parameter space by making use of the proposed qualitative knowledge model. Uncertainty in the Bayesian model space is restricted to a subset of models which are consistent with the body of qualitative knowledge. This class of consistent models is used to perform full Bayesian inference which can be approximated by Monte Carlo methods, but is analytically tractable for smaller networks.

2.1 Qualitative Knowledge Model

Now we provide a full formalism of how to translate a set of qualitative statements into probability inequality constraints. Qualitative knowledge models describe the process of transforming the qualitative statements into a set of probability constraints. Our knowledge model is closely related to the concepts of qualitative probabilistic network in Section 1.3.1, but provide more fine-tuned inequality relationships between the conditional probabilities. In this section, we formulate our knowledge model and in the following sections, we proposed Bayesian inference method which makes use of this knowledge model. Note that the inference method is independent of the qualitative knowledge model once the set of probabilistic inequality constraints which are translated from qualitative statements is given. Therefore, we argue that our inference approach could be smoothly transfer to other existing qualitative models such as qualitative probabilistic network [109] and the probabilistic commonsense model [79].

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.

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:

$$\begin{aligned} Depend(A,B) &= 0/1 \\ Influence(A,B) &= 0/1 \end{aligned} \tag{2.1}$$

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}.

Qualitative influences with directions can be defined based on the number of influences imposed from parents to child.

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.

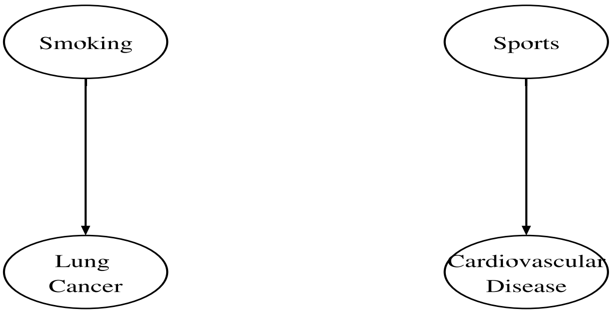
Definition 3.2 Iff presence of parent node A renders presence of child node B more likely, then the parent node is said to have a *single positive influence* on the child node. This can be represented by the inequality

$$P(B|A) \geq P(B|\overline{A}) \tag{2.2}$$

Definition 3.3 Iff 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

$$P(B|A) \leq P(B|\overline{A}) \tag{2.3}$$

Example 3.1



(a) Single Positive Influence

(b) Single Negative Influence

Figure 2.1: Example of Single Positive and Negative Influence

In the statement, *"smoking increases the risk of lung cancer"*, *smoking* is the parent node which has a single positive influence on child node *lung cancer.*

$$P(Lung\ Cancer|Smoking) \geq P(Lung\ Cancer|\overline{Smoking}) \tag{2.4}$$

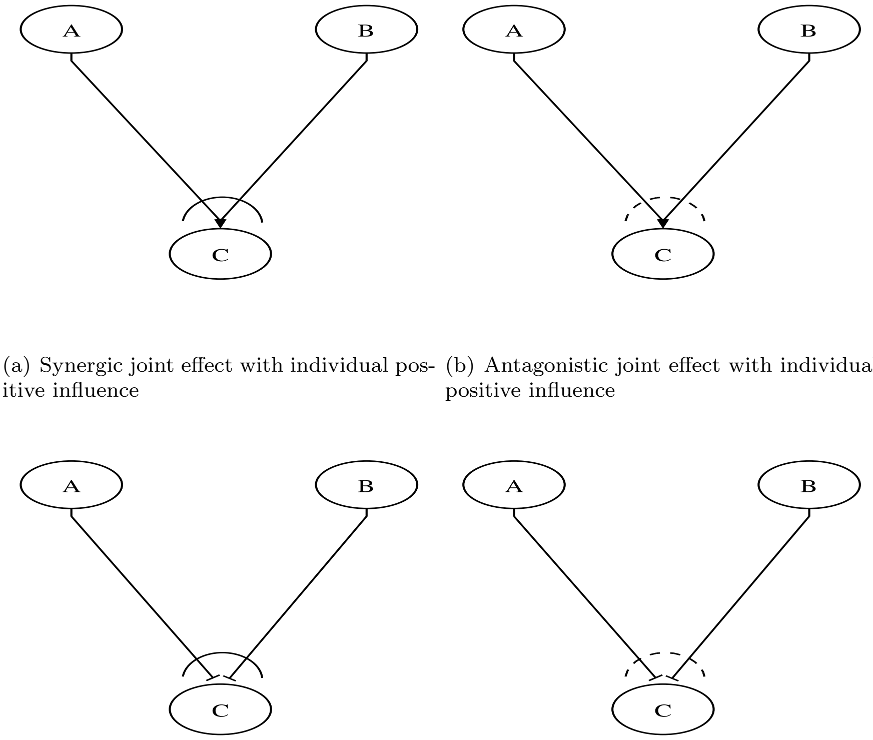
In another statement, *"Sports reduces the risk of cardiovascular disease"*, *Sports* is the parent node which imposes a single negative influence on child node *Cardiovascular disease.*

$$P(Card.\ Disease|Sports) \leq P(Card.\ Disease|\overline{Sports}) \tag{2.5}$$

The graphical representations of the above qualitative statements can be seen in figure 2.1.

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. In figure 2.2, we show an example of a synergic and an antagonistic type of joint influence of



(a) Synergic joint effect with individual positive influence

(b) Antagonistic joint effect with individual positive influence

(c) Synergic joint effect with individual negative influence

(d) Antagonistic joint effect with individual negative influence

Figure 2.2: Example of Joint Effect with Positive and Negative Influence

two parents on one child with which can impose positive or negative individual influences.

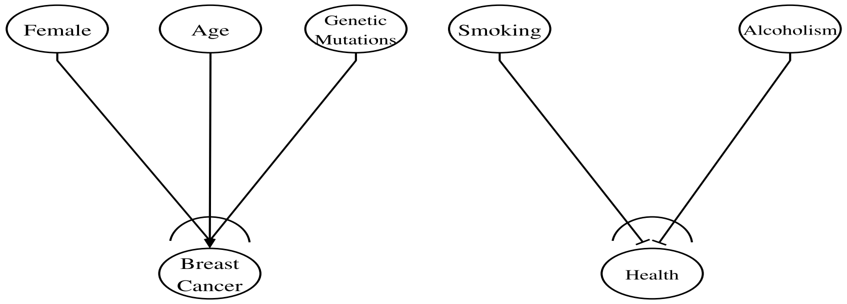
Definition 3.5 If a joint influence from two or more parent nodes generates a combined influential effect larger than the single effect from each individual parent, then the joint influence is referred to as *plain synergic joint influence* or *plain synergy.*

Assume that parent nodes A and B impose positive individual influences on child node C as shown in figure 2.2(a), then the knowledge model can be defined as

$$P(C|A,B) \geq \left\{ \begin{array}{c} P(C|A,\overline{B}) \\ P(C|\overline{A},B) \end{array} \right\} \geq P(C|\overline{A},\overline{B}) \tag{2.6}$$

Example 3.2

Let us consider breast cancer causes. Several risk factors have been identified for breast cancer. According to the American Cancer Society, the three most prominent risk factors are gender, age and genotype. It is stated that *being a woman is the main risk for breast cancer. The chance of getting breast cancer increases as a woman gets older. The most common mutations are those of the BRCA1 and BRCA2 genes. Women with these gene changes have up to an 80%*



(a) Female, Age and Gene Mutation synergically promote breast cancer

(b) Smoking and Alcoholism synergically degrade health

Figure 2.3: Example of Plain Synergy Influence

chance of getting breast cancer during their lifetimes. These knowledge about breast cancer risk factors can also be encoded by a qualitative causality model. According to the statements, the three main risk factors influence breast cancer by positive synergy as shown in figure 2.3(a), i.e. the joint influence of these three factors together is more significant than individual influences from any of these factors alone. We can represent this synergy by the inequalities

$$P(BC|F,A,M) \geq \left\{ \begin{array}{c} P(BC|F,\overline{A},\overline{M}) \\ P(BC|\overline{F},A,\overline{M}) \\ P(BC|\overline{F},\overline{A},M) \end{array} \right\}, \tag{2.7}$$

$$P(BC|F,A,M) \geq \left\{ \begin{array}{c} P(BC|F,A,\overline{M}) \\ P(BC|F,\overline{A},M) \\ P(BC|\overline{F},A,M) \end{array} \right\} \tag{2.8}$$

and

$$P(BC|F,A,M) \geq P(BC|\overline{F},\overline{A},\overline{M}) \tag{2.9}$$

If we assume these risk factors pair-wise symmetric, we can further derive the following inequalities:

$$\begin{array}{c} P(BC|F,A,\overline{M}) \\ P(BC|F,\overline{A},M) \\ P(BC|\overline{F},A,M) \end{array} \geq \left\{ \begin{array}{c} P(BC|F,\overline{A},\overline{M}) \\ P(BC|\overline{F},A,\overline{M}) \\ P(BC|\overline{F},\overline{A},M) \end{array} \right\} \tag{2.10}$$

where BC, F, A and M stands for *Breast Cancer*, *Female*, *Age* and *Mutation.*

Note that often but not always the combined influence refers to the sum of independent influences from each parent node to child node. This defintion has been used by [87]. Without loss of generality, we append this defintion to our classes of qualitative knowledge.

Assume that parent nodes A and B impose negative individual influence on child node C as shown in figure 2.2(c), then the knowledge model can be defined as

$$P(\overline{C}|A,B) \geq \left\{ \begin{array}{c} P(\overline{C}|A,\overline{B}) \\ P(\overline{C}|\overline{A},B) \end{array} \right\} \geq P(\overline{C}|\overline{A},\overline{B}) \tag{2.11}$$

Example 3.3
A report by the Australian Council of Smoking and Health reveals that the combination of drinking alcohol and smoking tobacco leads to a greater risk of developing cancer. The study found people who drink and smoke are up to five times more likely to develop head and neck cancers than those who do not. Therefore, we could represent this knowledge by synergic joint influence with negative individual effects as shown in figure 2.3(b) and inequality

$$P(\overline{H}|S,D) \geq \left\{ \begin{array}{c} P(\overline{H}|S,\overline{D}) \\ P(\overline{H}|\overline{S},D) \end{array} \right\} \geq P(\overline{H}|\overline{S},\overline{D}) \tag{2.12}$$

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.* [87]

Assume in case that parent nodes A and B impose a positive individual influence on child node C as shown in figure 2.2(a), then we define

$$P(C|A,B) \geq P(C|A,\overline{B}) + P(C|\overline{A},B) \geq \left\{ \begin{array}{c} P(C|A,\overline{B}) \\ P(C|\overline{A},B) \end{array} \right\} \geq P(C|\overline{A},\overline{B}) \tag{2.13}$$

Comparing Eq. 2.13 with Eq. 2.6, 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.

Assume in case that parent nodes A and B impose negative individual influences on child node C as shown in figure 2.2(c), then we can define

$$P(\overline{C}|A,B) \geq P(\overline{C}|A,\overline{B}) + P(\overline{C}|\overline{A},B) \geq \left\{ \begin{array}{c} P(\overline{C}|A,\overline{B}) \\ P(\overline{C}|\overline{A},B) \end{array} \right\} \geq P(\overline{C}|\overline{A},\overline{B}) \tag{2.14}$$

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 as shown in figure 2.2(b) the antagonistic influence of A and B can be represented by

$$P(C|\overline{A},\overline{B}) \leq P(C|A,B) \leq \left\{ \begin{array}{c} P(C|A,\overline{B}) \\ P(C|\overline{A},B) \end{array} \right\} \tag{2.15}$$

Assume that parent nodes A and B have independent negative single influence on child node C as shown in figure 2.2(d) the antagonistic influence of A and B can be represented by

$$P(\overline{C}|\overline{A},\overline{B}) \leq P(\overline{C}|A,B) \leq \left\{ \begin{array}{c} P(\overline{C}|A,\overline{B}) \\ P(\overline{C}|\overline{A},B) \end{array} \right\} \tag{2.16}$$

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

$$\begin{array}{rcl} P(C|A,B) & \geq & P(C|\overline{A},B) \\ P(C|A,\overline{B}) & \geq & P(C|\overline{A},\overline{B}) \\ P(C|A,\overline{B}) & \geq & P(C|A,B) \\ P(C|\overline{A},\overline{B}) & \geq & P(C|\overline{A},B) \end{array} \tag{2.17}$$

Once formulated, the Monte Carlo sampling procedure will make sure that all inequalities are satisfied for valid models. 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.

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

$$\begin{array}{rcl} P(B|A) & \geq,\leq & R \times P(B|\overline{A}) + \Delta \\ P(B|A) & \in & [Bd_{min}, Bd_{max}] \\ P(B|\overline{A}) & \in & [Bd'_{min}, Bd'_{max}] \end{array} \tag{2.18}$$

which R is *Influence Ratio*, Δ is *Influence Difference* and Bd, Bd' denote *bound.*
1. Influence Ratio

Definition 3.6 In one qualitative causal influence, the ratio between any two or more configurations of the probabilistic representation of this influence is referred to *influence ratio.*

$$P(B|A) \approx R \times P(B|\overline{A}) \tag{2.19}$$

Example 3.4 In the statement, *"The risk of lung cancer among smokers is approximate 10 times higher than non-smokers"*, *Smoking* is the parent that imposes positive influence on child node *lung cancer* with influence ratio approximately equals to 10.

$$P(LC|S) \approx 10 \times P(LC|\overline{S}) \tag{2.20}$$

where LC denotes *Lung Cancer* and S is *Smoking.*

2. Influence Difference

Definition 3.7 In one qualitative causal influence, the difference between any two or more configurations of the probabilistic representation of this influence is referred to *influence difference.*

In the case that node B impose single influence on node A, there are two probabilistic configurations. The *influence difference* in this case can be expressed as

$$P(B|A) \approx \Delta + P(B|\overline{A}) \tag{2.21}$$

3. Influence Bound

Definition 3.8 In one qualitative causal influence, the absolute value bound of any configuration of the probabilistic representation of this influence is referred to *influence bound.*

In the case that node B impose single influence on node A, there are two probabilistic configurations. The *influence bound* in this case can be expressed as

$$P(B|A) \in [Bd_{min}, Bd_{max}] \tag{2.22}$$

and/or

$$P(B|\overline{A}) \in [Bd'_{min}, Bd'_{max}] \tag{2.23}$$

where Bd and Bd' denote *bound.*

Example 3.5 In the statement, *"The risk of smokers to get lung cancer in their life time is approaximate 1 out of 10".*

$$P(LC|S) \in [0.1 - \delta, 0.1 + \delta] \tag{2.24}$$

where LC denotes *Lung Cancer*, S is *Smoking* and δ is a small quantity reflects the slight uncertainty on the probability bound which is described by verbal *approximate.* Once the qualitative knowledge is translated by the feature set $\{\Pi(Dp, I), \Lambda(\Sigma, \Psi(R, \Delta, Bd))\}$ according to Eq. 2.2 to Eq. 2.18, the distribution of ground models is defined by this knowledge.

2.2 Comparison of Qualitative Knowledge Models

As described in section 1.2, mainly two approaches have been proposed to the concept of qualitative uncertainty modeling, i.e. Wellman approach [108] and

Neufeld approach [79]. In this section, we compare our proposed qualitative knowledge model with these approaches and demonstrate that our approach represents sufficient probability relationships as the two approaches, in addition, provide more fine-grained contraints on the probability configurations.

2.2.1 Compare to Qualitative Probabilistic Network (QPN)

In Wellman approach, a *Qualitative Probabilistic Network(QPN)* is a pair $G = (V, Q)$, where V is the set of variables or vertices of the graph and Q is a set of qualitative relationships among the variables. There are two types of qualitative relationship for modeling the QPN, i.e. *Qualitative Influence(QI)* and *Qualitative Synergy(QS)*. We begin our discussion by comparing the QI definition in Wellman approach to our qualitative knowledge model in section 2.1. Qualitative influence can be thought of as qualitative relations describing the sign(direction) of the relationship between a pair of variables. Accordingly, given the qualitative hypothesis "*A positively influences C*", QI translate it as: Under all contexts $x \in X(pred_G(C) - A)$, A makes C more likely. The definition of QI with binary variables is given in Eq. 1.26 in section 1.2. In case of $X = \emptyset$, i.e. C has single parent A, the definition of Wellman converges to our definition of single positive influence as Eq. 2.2. In case of $X \neq \emptyset$, C has more than one parent, say A and B, then according to QI definition, A always makes C more likely with or without the presence of B, i.e.

$$\begin{aligned} &P(C|A, B) \geq P(C|\overline{A}, B) \\ &P(C|A, \overline{B}) \geq P(C|\overline{A}, \overline{B}) \end{aligned} \tag{2.25}$$

This situation fits to the joint influence in our knowledge model where the relationship between A, B and C can be classified into 4 mutual exclusive catergories in section 2.1, i.e. *Plain Synergy*, *Additive Synergy*, *Antagonism* and *Mixed Joint Influence*. Since it is known that A has positive individual influence on C, i) if B imposes negative influence on C, the qualitative relationship of A, B and C in our knowledge model is defined by *Mixed Joint Influence* as Eq. 2.17. ii) if B imposes positive influence on C as well and forms plain or additive synergy with A, then their qualitative relationship is defined as Eq. 2.6 or 2.13. In either case, we could derive the inequality consistent with Eq. 2.25. If B imposes positive influence on C and forms antagonistic joint influence with A, i.e. the positive influence of A on C is impaired by the positive influence of B on C and if we further assume that the positive influences of A and B on C are symmetric, then the assertion given by QPN, i.e. the presence of A always makes the presence of C more likely regardless to the context of x$\in$ X, is not always valid. For example, due to the strong antagonistic effect between A and B, simutaneous presence of A and B may intensively reduce their positive influence on C, i.e. P(C|A,B), and makes C less likely than the case of single parent, i.e. P(C|A,B)$\leq$ P(C|$\overline{A}$,B) and P(C|A,B)$\leq$ P(C|A,$\overline{B}$). However, in case B is not present, it is true that A makes C more likely, i.e. P(C|A,$\overline{B}$)$\geq$P(C|$\overline{A}$,$\overline{B}$). Therefore, we show that this assertion requires careful inspection on the relative strength between their antagonistic effect and their individual positive influences in this case.

The second type of qualitative relationship introduced in QPN is Qualitative Synergy as defined by Eq. 1.34 to Eq. 1.37 in Section 1.3.1. According to the definition of positive synergy in Eq. 1.36, for multinomial variables A, B and C, an general ordering scheme is introduced on the cumulative distribution function (CDF) to represent the synergy as

$$F(c_0|a_1,b_1) - F(c_0|a_2,b_1) \leq F(c_0|a_1,b_2) - F(c_0|a_2,b_2) \tag{2.26}$$

where a_1, a_2, b_1, b_2 and c_0 are any values in the value-range of A, B and C. $a_1 \geq a_2$, $b_1 \geq b_2$. F is the *cumulative distribution function* and can be formulated as

$$\begin{aligned} F(c_0|\cdot,\cdot) &= \sum_{c_i \leq c_0} P(c_i|\cdot,\cdot) \\ &= 1 - \sum_{c_i > c_0} P(c_i|\cdot,\cdot) \end{aligned} \tag{2.27}$$

Therefore, in case that A, B and C are binary variables, c_0=$\overline{C}$, we can re-write the qualitative relationship by substituting Eq. 2.27 into Eq. 2.26 as

$$P(C|A,B) \geq P(C|\overline{A},B) + P(C|A,\overline{B}) - P(C|\overline{A},\overline{B}) \tag{2.28}$$

where $P(\cdot|\cdot)$ is the probability density function. Comparing the positive synergy of QPN in Eq. 2.28 with the additive synergy in our knowledge model as defined in Eq. 2.13, we conclude that our definition on the additive synergy provides sufficient but not necessary condition to the positive synergy in Wellman approach. Thus, qualitative constraints defined in Eq. 2.13 automatically ensure its satisfaction to the synergy definition in Wellman approach [108]. In addition, our knowledge model provides more fine-grained restriction on the qualitative relationships on the conditional probability distribution which can not be derived by Wellman's model.

Similarly, the negative synergy of QPN defined in Eq. 1.37 can be derived by inversing the sign in Eq. 2.26. Thus, in binary case, the qualitative relationship on conditional probability distribution can be formulated as

$$P(C|A,B) \leq P(C|\overline{A},B) + P(C|A,\overline{B}) - P(C|\overline{A},\overline{B}) \tag{2.29}$$

Comparing Eq. 2.29 to Eq. 2.15, we conclude that our definition on the antagonism is a sufficient but not necessary condition to the negative synergy in Wellman approach. Combining the above facts, we can generalize that our qualitative uncertainty model provide broader and more fine-grained definition on the conditional probability distributions with inequality constraints than those discussed in Wellman approach [108]. Qualitative constraints generated by our knowledge model automatically satisfy the Wellman's definition yet provide more detailed information for better restraining the uncertainty.

2.2.2 Compare to Probabilistic Commonsense Reasoner

Neufeld formalizes the idea of qualitative influence by means of the concept of favoring. He bases his probabilistic commonsense reasoner on a graphical notation which distinguishes the four kinds of qualitative relations that may hold between variables. Assume two binary variables A and B, we can define the qualitative relationships in the Neufeld approach as follows:

Defeasible links Given A, B is more likely to happen, $A \rightarrow B$, if

$$1 > P(B|A) > P(B) \tag{2.30}$$

Logical links Given A, B will surely happen, $A \Rightarrow B$, if

$$1 = P(B|A) > P(B) \tag{2.31}$$

Negative Defeasible links Given A, B is likely to happen, $A \nrightarrow B$, if

$$1 > P(\overline{B}|A) > P(\overline{B}) \tag{2.32}$$

Negative Logical links Given A, B is likely to happen, $A \nRightarrow B$, if

$$1 = P(\overline{B}|A) > P(\overline{B}) \tag{2.33}$$

We compare our qualitative knowledge model to the four qualitative relations given in Neufeld approach [79]. For *Defeasible Links* and *Negative Defeasible Links* in Eq. 2.30 and Eq. 2.32, we prove that they are virtually equivalent to the definition of *Qualitative Influence* in Wellman approach [108] in case of single parent and to the definition of *Single Positive/Negative Influence* in our proposed qualitative knowledge model. For *Logical links* and *Negative logical links* in Eq. 2.31 and Eq. 2.33, we show that they are virtually equivalent to the definition of *Single Positive/Negative Influence* in our model with tight boundary restrictions. We can describe the relationship of P(B|A) and P(B) by computing their difference

$$P(B|A) - P(B) = [P(B|A) - P(B|\overline{A})](1 - P(A)) \tag{2.34}$$

Since 1≥(1-P(A))≥0, the relationship of P(B|A) and P(B) is determined by the relationship of P(B|A) and P(B|$\overline{A}$) eventually. In case P(B|A)≥P(B|$\overline{A}$), i.e. A positively influence B in QPN and in our proposed knowledge model, it is configured as *Defeasible links* in Neufeld approach. In case P(B|A)≤P(B|$\overline{A}$), i.e. A negatively influence B in QPN and in our proposed knowledge model, it is equivalent to the *Negative Defeasible Links* in Neufeld approach.

2.3 Bayesian Modeling based on Consistent Qualitative Knowledge

In this section, we suggest a way to use qualitative relational statements for inference in the Bayesian framework. We proceed from the general Bayesian structural inference in Eq. 1.13 based on data. In addition, we wish to model both structure and parameter space distribution by incorporating qualitative prior knowledge with the data and make quantitative predictions with full Bayesian approach by integrating over the Bayesian model space. We give a detailed recipe to transform knowledge, represented by a set of qualitative statements, into an *a priori* distribution for models.

2.3.1 Modeling with Static Bayesian Networks

Recall that a Bayesian network m represents the joint probability distribution θ of a set of variables $\mathbf{X} = X_1, X_2, ..., X_n$ [48] with a graph structure G, which defines the (in)dependences and influences between variables. Each component of θ is a table of conditional probabilities whose elements define the entries of the corresponding conditional probability tables (CPTs). Hence, a Bayesian network can be written as $m = \{G, \theta\}$. If we believe that 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 $X_E \in \mathbf{X}$, we wish to derive the distribution of the remaining variables $X \in \mathbf{X} \backslash \mathbf{X}_E$. it is provided by their conditional probability given the evidence in light of the model,

$$P(X|E,m) = P(X|E,G,\theta) \tag{2.35}$$

which can be efficiently evaluated by Variable Elimination algorithm or Message-propagation algorithm introduced in Section 1.2.2. Given the past observation data set $\mathbf{D}$, the model m's structure and parameter can be learned by BIC score or BDe score defined in Section 1.2.3.

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. Then a posteriori distribution of models m is then given by

$$P(m|D,\Omega) = \frac{P(D|m,\Omega)P(m|\Omega)P(\Omega)}{P(D,\Omega)} \tag{2.36}$$

The first term in the numerator of Eq. (2.36) 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

$$P(m|D,\Omega) = \frac{1}{Z} P(D|m)P(m|\Omega) \tag{2.37}$$

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. Now, inference in the presence of evidence is performed by building the expectation across models:

$$\begin{aligned} P(X|E,D,\Omega) &= \int_m P(X|E,m)P(m|D,\Omega)dm \\ &= \int_m P(X|E,m)P(D|m)P(m|\Omega)dm \end{aligned} \tag{2.38}$$

In this thesis, we consider the extreme case where no quantitative data is available, $D = \emptyset$. Thus, the model uncertainty is fully described by the qualitative prior knowledge Ω. We will show throughout this thesis that even in this case it is still possible to perform proper quantitative Bayesian inference,

$$P(X|E,\Omega) = \int_m P(X|E,m)P(m|\Omega)dm \tag{2.39}$$

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. (2.39).

In order to determine $P(m|\Omega)$, 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 to define a Bayesian model class. A Bayesian model is determined by a graph structure G 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

$$\begin{aligned} P(m|\Omega) &= P(s_k, \theta|\Omega) \\ &= \frac{P(\theta|s_k,\Omega)P(s_k|\Omega)}{\sum_{a=1}^{K} \int_{\Theta_a} d\theta P(\theta|s_a,\Omega)P(s_a|\Omega)}. \end{aligned} \tag{2.40}$$

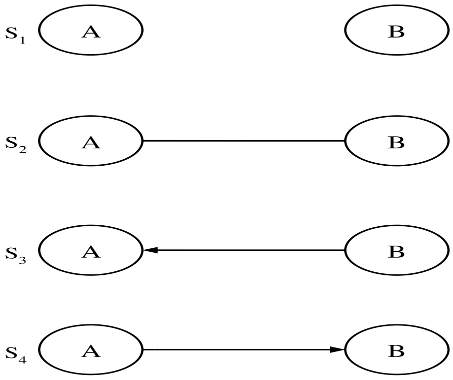
In Eq. 2.40, first the set of allowed structures is determined by means of Ω, followed by the distributions of the corresponding CPT configurations. Then, we calculate the model's posterior probability $P(m|\Omega)$ in Eq. 2.40. Inference is carried out by integrating over the structure space and the structure-dependent parameter space:

$$P(X|E,\Omega) = \sum_{k=1}^{K} \int_{\Theta_k} d\theta P(X|E, s_k, \theta)P(s_k, \theta|\Omega). \tag{2.41}$$

Here we assume Ω to be represented as a list of consistent qualitative statements. In this form, the information can be used in a convenient way to determine the model prior, Eq. 2.40: (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.

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 $P(s_k|\Omega)$. Likewise, only parameter values on that structure, which are consistent with the contents of that statement, are assigned a nonzero probability $P(\theta|s_k, \Omega)$. If no further information is available, the distribution is constant in the space of consistent models.

In the following paragraphs, we first provide a toy example of that procedure, followed by a more thorough introduction, which takes into account recurrent and conflicting statements. We consider a simple toy case in which the body of knowledge Ω consists of a single statement, Ω=*"A activates B"*. We know that there are two random variables A and B, which we assume binary, and we need to consider the set of all possible Bayesian models on (A, B). Fig. 2.4(a) shows the set of possible model structures. In the next step, we use the statement to constrain the space of structures to those consistent with the statement. "A activates B" directly states a causal influence of A on B, hence the bottom graph structure in Fig. 2.4(a) is assigned a nonzero probability: $P(s_4|\Omega) = 1$, $P(s_k|\Omega) = 0,\ k = 1, 2, 3)$.



(a) Constraining the Structure Space

Figure 2.4: Toy Example for Constraining the Bayesian Model Space

A	P$(B = 1\|A)$
0	θ_0
1	θ_1

Table 2.1: Conditional Probability Table

This graph structure encodes the probability distribution

$$P(A, B) = P(B|A)P(A) \tag{2.42}$$

No further information on $P(A)$ is available, however $P(B|A)$ can be further constrained. Table 2.1 shows the corresponding CPT. The CPT entries, i.e., the values of the conditional probabilities form the components of the parameter vector $\theta = (\theta_0, \theta_1)$ of the model class with structure s_4. From the statement we now can infer that the probability of B active when A active is higher than the same probability with A inactive. We obtain the inequality relationship

$$P(B = 1|A = 1) \geq P(B = 1|A = 0), \Rightarrow \theta_1 \geq \theta_0. \tag{2.43}$$

Hence, the set of model parameters consistent with that statement is given by

$$\Theta_4 = \{(\theta_0, \theta_1)|0 \leq \theta_0 \leq 1 \wedge \theta_0 \leq \theta_1 \leq 1\}, \tag{2.44}$$

and the distribution of models in the structure-dependent parameter space becomes

$$P(\theta|s_4, \Omega) = \left\{ \begin{array}{ll} 2 & \theta \in \Theta_4 \\ 0 & else \end{array} \right. . \tag{2.45}$$

All consistent model parameters have been assigned the same probability, which reflects the lack of any further biasing information.

Having derived the Bayesian model class (s_4, Θ_4) consistent with the statement, we can now perform inference by model averaging using Eq. 2.41. Let us assume we observed A active, i.e., $E = \{A = 1\}$, and let us ask what is the probability of having B active under this conditions. We do so by integrating over all models with nonzero probability and averaging their respective inferences, which can be done analytically in this simple case:

$$\begin{aligned} P(B=1|E,\Omega) &= \sum_k P(s_k|\Omega) \int d\theta P(B=1|A=1, s_k, \theta) P(\theta|s_k, \Omega) \\ &= 2\int_{\Theta_4} d\theta P(B=1|A=1,\theta) \\ &= 2\int_0^1 d\theta_0 \int_{\theta_0}^1 d\theta_1 \theta_1 \\ &= 2/3 \end{aligned} \tag{2.46}$$

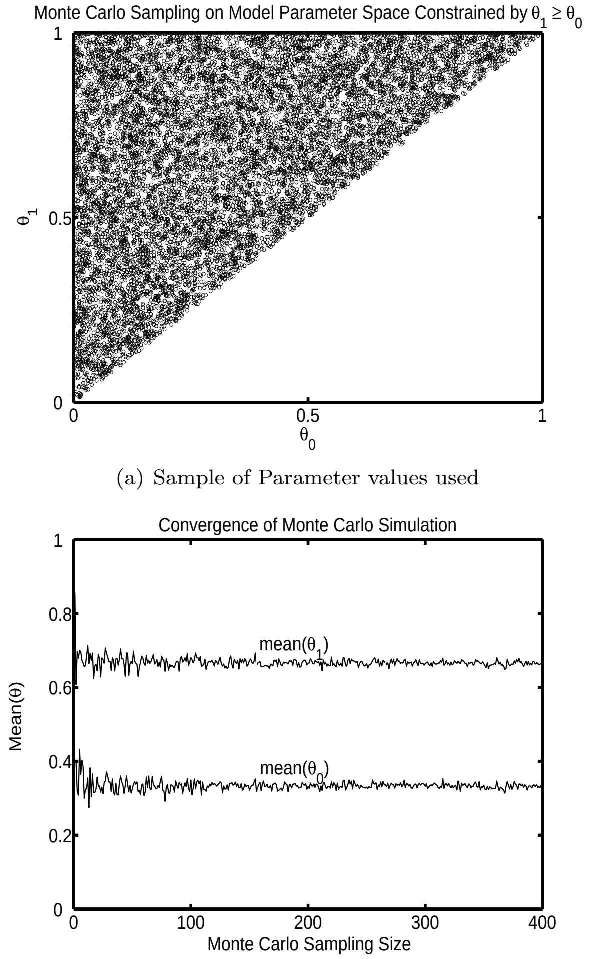
Similarly, the expected value of probability of having B active when A is inactive, i.e. $E = \{A = 0\}$, can be calculated by model averaging as

$$\begin{aligned} P(B=1|E,\Omega) &= \sum_k P(s_k|\Omega) \int d\theta P(B=1|A=0, s_k, \theta) P(\theta|s_k, \Omega) \\ &= 2\int_{\Theta_4} d\theta P(B=1|A=0,\theta) \\ &= 2\int_0^1 d\theta_1 \int_0^{\theta_1} d\theta_0 \theta_0 \\ &= 1/3 \end{aligned} \tag{2.47}$$

It is worth to note that, as long as simple inequalities are considered as statements, the problem remains analytically tractable even in higher dimensions. In general, however, integration during Bayesian inference can become intractable by analytical methods. One way to resolve the integration problem is to use Monte Carlo methods. Throughout this work, we use the *Accept-Reject* algorithm [86] to approximate the constant distribution inside the polyhedron of consistent model parameters. Each dot in Fig. 2.5(a) represents one randomly selected Bayesian model of the toy example, for which the inference has been carried out. Fig. 2.5(b) shows how the empirical means of the inference results for $P(B = 1|A = 1)$ and $P(B = 1|A = 0)$ depend on the Monte Carlo sample size in this simple model.

2.3.2 Modeling with Dynamic Bayesian Networks

However, the conventional Bayesian networks are not able to model the cyclic regulations in molecular networks. We suggest to solve this problem by utilizing



(a) Sample of Parameter values used

(b) Converge of averaged inference results as a function of sample size

Figure 2.5: Monte Carlo Simulation in the 2-dimensional Space

the Dynamic Bayesian Networks (DBN) [23,78]. An example of DBN is shown in Fig. 2.6(a) and it can be defined by a vector of 2-Time-Slice Bayesian Networks (2TBN) over time as shown in Figure 2.6(b). Each 2TBN is a conventional Bayesian model with the structure s and conditional probabilities $\overline{\theta}$ of the DBN and encodes the joint probability of the nodes $\overline{X}$ at time t and $(t-1)$, i.e. $P(\overline{X}_t, \overline{X}_{t-1})$. The joint probability of a DBN over time T is

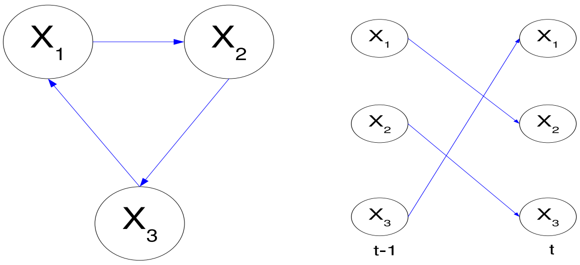
$$P(\overline{X}_1, \ldots, \overline{X}_T) = \prod_{t=1}^{T} P(\overline{X}_t | \overline{X}_{t-1}) \tag{2.48}$$

For *t-th* 2TBN, the joint probability of $\overline{X}$ at time t, i.e. $P(\overline{X}_t)$ can be written as

$$P(\overline{X}_t) \quad = \quad P(X_{1,t}, \ldots, X_{N,t})$$

$$= \prod_{n=1}^{N} P(X_{n,t}) \tag{2.49}$$

where $X_{n,t}$ denotes the *n-th* node at time t. The child nodes in the *t-th* 2TBN, $X_{n,t}$, are independent given the parents, i.e. the nodes at time $(t-1)$. The posterior probability distribution of each node at time t, i.e. $P(X_{n,t})$ in Eq. 2.49 can be calculated by integrating over the parents as in the case of conventional Bayesian network,



(a) DBN Example (b) 2TBN

Figure 2.6: Dynamic Bayesian Example

$$\begin{aligned} P(X_{n,t}) &= \int_{\pi(X_n)} P(X_n|\pi(X_n))P(\pi(X_n))d\pi(X_n) \\ &= \sum_{j=1}^{J} \theta_j P_j^{t-1}(\pi(X_n)) \end{aligned} \tag{2.50}$$

where θ_j denotes the *j-th* entry in the conditional probability table of node X_n given its parents. $P_j^{t-1}(\pi(X_n))$ represent the joint probability of *j-th* configuration of the parents states at time $(t-1)$. The posterior probability distribution of X_n can be used as the priori probability for the next time step. Thus the posterior probability $P(X_{n,t})$ can be calculated iteratively over time $t = \{0, \ldots, T\}$.

In this section, we extend the knowledge-driven Bayesian inference approach to Dynamic Bayesian model. A DBN model, m, can be learned from a time-series data. [62, 78, 115]. As demonstrated in the last section, if there is a set of consistent hypotheses retrieved from a publication which defines a class of models, with the structure and its associated parameter space. The inference with full Bayesian approach is calculated by integrating the inference in each model weighted by its posterior probability given the set of hypothesis as in Eq. 2.41. The inference can be written as

$$\begin{aligned} P(X_n|E,\Omega) &= \sum_{k=1}^{K} \int_{\Theta_k} P(X_n|E,s_k,\theta)P(s_k,\theta|\Omega)d\theta \\ &\approx \frac{1}{K}\sum_{m_k} P(X_n|E,m_k) \end{aligned} \tag{2.51}$$

Since there is no inconsistent knowledge, only the structure which is consistent with the hypotheses is assigned with non-zero probability $P(s_k|\Omega)$. Likewise,

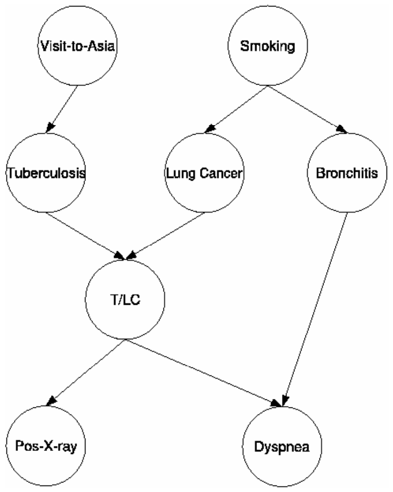


Figure 2.7: ASIA Network Structure

only parameter values on that structure, which are consistent with the contents of the hypotheses, are assigned a nonzero probability $P(\theta|s_k, \Omega)$. If no further information is available, the distribution is constant in the space of consistent models. Now, we can perform inference on the marginal probability of X_n at time t in each DBN model m_k according to Eq. 2.50

$$P(X_{n,t}|E, s_k, \theta) = \sum_{j=1}^{J} \theta_{k,j} P_j^{t-1}(\pi(X_n), E) \tag{2.52}$$

where $\theta_{k,j}$ represents the *j-th* entry of the CPT in *k-th* DBN model. E denotes the evidence of the observed nodes and $P_j^{t-1}(\pi(X_n), E)$ denotes the joint probability distribution of the *j-th* configuration of the parent nodes $\pi(X_n)$ at time $(t-1)$ given the observation E. Therefore, the quantitative inference in Eq. 2.51 can be calculated by

$$\begin{aligned} & P(X_{n,t}|E, \Omega) \\ = \; & \sum_{k=1}^{K} \int_{\Theta} \sum_{j=1}^{J} \theta_{k,j} P_j^{t-1}(\pi(X_n), E) P(s, \theta_{k,j}|\Omega) d\Theta \end{aligned} \tag{2.53}$$

The inference in Eq. 2.53 can be calculated can be performed for each model $m_k = (s_k, \theta)$ over time T and the predictions are averaged over all models in the model class $\widetilde{M}$. We will discuss how to compute this integral in Section 2.4.1.

2.3.3 ASIA Network

An example of Bayesian modeling and inference based on consistent qualitative knowledge is the ASIA network [65]. A popular toy belief model for testing

Table 2.2: ASIA Network Parameters

Node Pair	CPT,θ			
$(Pa., Ch.)$	$P(Ch \mid Pa)$		$P(Ch \mid \overline{Pa})$	
(V,T)	α_1=0.05		α_0=0.01	
(S,LC)	β_1=0.3		β_0=0.01	
(S,B)	γ_1=0.6		γ_0=0.3	
(T/LC,X)	λ_1=0.9		λ_0=0.05	
([Pa_1,Pa_2],$Ch.$)	Pa_1, Pa_2	$Pa_1, \overline{Pa_2}$	$\overline{Pa_1}, Pa_2$	$\overline{Pa_1}, \overline{Pa_2}$
([T/LC,B],D)	ξ_3=0.9	ξ_2=0.7	ξ_1=0.8	ξ_0=0.1
([T,LC],B)	f_3=1	f_2=1	f_1=1	f_0=0

Bayesian algorithms. It is shown in Figure 2.3.2. Each node in the network corresponds to some condition of the patient. The joint probability distribution is defined by conditional probability tables (CPT) of a child node given its parent nodes. We listed the actual parameter, i.e. $\theta = (\alpha, \beta, \gamma, \lambda, \xi)$, in Table 2.2. The qualitative knowledge Ω for each possible relationship between a set of entities in ASIA network are extracted from statements. The statements are retrieved from various resources, such as scientific publications and internet resources. A valid qualitative statement should be able to describe the entities and their causal relationship with baseline and/or extended features defined in section 2.

In ASIA network, entities of interest include *Visit-to-Asia, Tuberculosis, Smoking, Lung Cancer, Bronchitis, Lung Cancer-or-Tuberculosis, Positive-X-Ray* and *Dyspnea*, i.e. *X*={*V, TB, S, LC, B, TLC, XR, D*}. The valid statements are retrieved as below:

1. *Tobacco smoke is the primary cause of lung cancer. Although nonsmokers can get lung cancer, the risk is about 10 times greater for smokers.* (http://www.netdoctor.co.uk)

2. *In the US, 25% of the population smokes. The lifetime risk of developing lung cancer in smokers is approximately 10%.* (http://www.chestx-ray.com/Smoke/Smoke.html)

3. *The study essentially confirmed that the risk of acquiring infection with M. tuberculosis among these travelers (to Asia) was essentially the same as that estimated for the general population in the host countries-that is, an annual risk of infection with M. tuberculosis of 1% to 3%.* [91]

4. *Risk of chronic bronchitis was significantly higher in current smokers than in never smokers, the relative risk RR=2.85; 95% confidence interval CI=[2.45 3.32].* [105]

5. *The National Cancer Institute trials demonstrated that the sensitivity of CXR is 54% when only "suspicious" CXRs are coded as positive, with a specificity of 99%; When "indeterminate" CXRs are considered positive, the Sensitivity of CXR increase to 84%" with a specificity of 90%". However, false negative CXR results continue to be a significant problem.* [40]

6. *Dyspnea is often a presenting symptom for lung cancer patients as a result of direct and indirect effect of the tumor. Studies have shown that 50% of cancer patients in general complain of shortness of breath, with 20% rating it as moderate to severe. 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)

7. *Respiratory symptoms included chronic coughing in 74% of Patients (morning cough in 19%), chronic sputum production in 57% (21% with morning productivity), dyspnea on exertion in 61%, and both at rest and on exertion in 19%.* [84]

We avoid the problem of creditability of the statements by assuming each statement is true, i.e. $P(\Omega) = 1$. To ensure that no inconsistent knowledge complicates our study, we retrieved only consistent statements for each relation in the ASIA network. The corresponding constraints on the network structure and parameter space can be compactly summarized in Table 2.3 and inequalities on CPT entries are thereby formalized based on these constraints.

$$\alpha_0 \le \alpha_1 \qquad \beta_0 \le \beta_1 \qquad \gamma_0 \le \gamma_1 \qquad \lambda_0 \le \lambda_1 \tag{2.54}$$

$$\pi_0 \le \pi_1 \le \pi_2 \tag{2.55}$$

$$\pi_0 = \{\xi_0\} \qquad \pi_1 = \{\xi_1, \xi_2\} \qquad \pi_2 = \{\xi_3\} \tag{2.56}$$

$$2.45\gamma_0 \le \gamma_1 \le 3.32\gamma_0 \qquad 9\beta_0 \le \beta_1 \le 11\beta_0 \tag{2.57}$$

$$\begin{array}{cc} \alpha_1 \in [1\%, 3\%] & \beta_1 \in [9\%, 11\%] \\ \lambda_0 \in [1\%, 10\%] & \lambda_1 \in [54\%, 84\%] \end{array} \tag{2.58}$$

In Table 2.4, ξ is a parameter of four dimension. Besides the direct qualitative knowledge constraint on the 4-dimension space, qualitative knowledge are available to constrain the projection of this probability distribution in a degraded 2-dimension space. The qualitative constraints in the degraded parameter space can be written as

$$\begin{array}{rcl} P(D|B) &=& P(TLC)\xi_3 + (1 - P(TLC))\xi_1 \\ P(D|\overline{B}) &=& P(TLC)\xi_2 + (1 - P(TLC))\xi_0 \\ P(D|TLC) &=& P(B)\xi_3 + (1 - P(B))\xi_2 \\ P(D|\overline{TLC}) &=& P(B)\xi_1 + (1 - P(B))\xi_0 \end{array} \tag{2.59}$$

$$\begin{array}{r} P(D|TLC) \ge P(D|\overline{TLC}) \\ P(D|B) \ge P(D|\overline{B}) \end{array} \tag{2.60}$$

$$\text{P(D|TLC)} \in [60\%, 90\%] \qquad \text{P(D|B)} \in [61\%, 80\%] \tag{2.61}$$

Without further information on the priori probability distribution of P(TLC) and P(B), we can assume a uniform distribution, i.e. P(TLC)=0.5, P(B)=0.5. Thus, constraints in Eq. 2.59 to Eq. 2.61 can be explicitly written as

$$\xi_3 + \xi_1 \ge \xi_2 + \xi_0 \qquad \xi_3 + \xi_2 \ge \xi_1 + \xi_0 \tag{2.62}$$

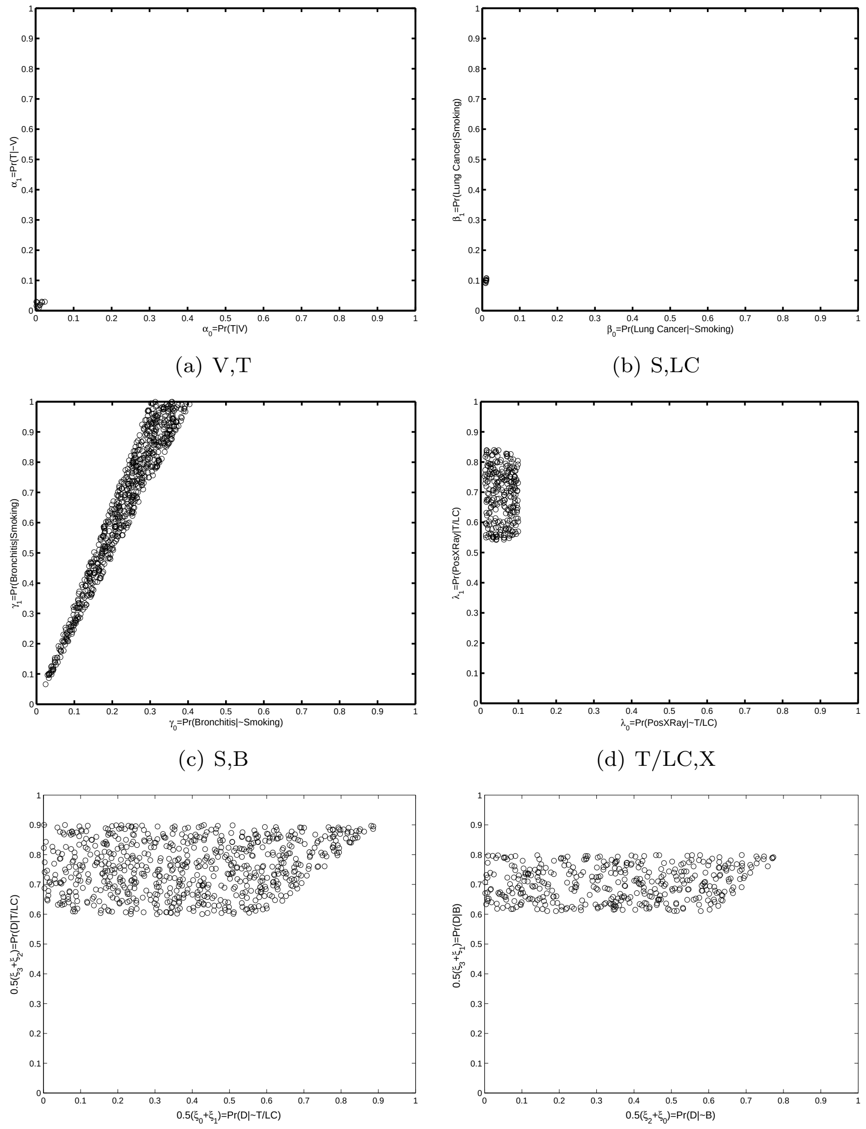
$$\xi_3 + \xi_1 \in [1.22, 1.6] \qquad \xi_3 + \xi_2 \in [1.2, 1.8] \tag{2.63}$$

Table 2.3: Constraints on ASIA Model Structure and Parameter Space.

Node	Dp	I	Σ	R	Δ	BD
V,T	1	1	SP	NULL	NULL	[1%,3%]
S,LC	1	1	SP	[9,11]	NULL	[9%,11%]
S,B	1	1	SP	[2.45,3.32]	NULL	NULL
B,D	1	1	SP	NULL	NULL	[61%,80%]
T/LC,X	1	1	SP	NULL	NULL	[54%,84%] [1%,10%]
T/LC,D	1	1	SP	NULL	NULL	[60%,90%]
(T/LC,B),D	1	1	PLSYN	NULL	NULL	NULL

Table 2.4: ASIA Network Averaged Parameters

Node Pair	CPT,θ			
$(Pa., Ch.)$	$P(Ch \mid Pa)$		$P(Ch \mid \overline{Pa})$	
(V,T)	$\overline{\alpha}_1$=0.02		$\overline{\alpha}_0$=0.01	
(S,LC)	$\overline{\beta}_1$=0.1		$\overline{\beta}_0$=0.01	
(S,B)	$\overline{\gamma}_1$=0.67		$\overline{\gamma}_0$=0.24	
(T/LC,X)	$\overline{\lambda}_1$=0.69		$\overline{\lambda}_0$=0.05	
$([Pa_1,Pa_2],Ch.)$	Pa_1, Pa_2	$Pa_1, \overline{Pa_2}$	$\overline{Pa_1}, Pa_2$	$\overline{Pa_1}, \overline{Pa_2}$
([T/LC,B],D)	$\overline{\xi}_3$=0.88	$\overline{\xi}_2$=0.55	$\overline{\xi}_1$=0.63	$\overline{\xi}_0$=0.24
([T,LC],B)	f_3=1	f_2=1	f_1=1	f_0=0



(a) V,T

(b) S,LC

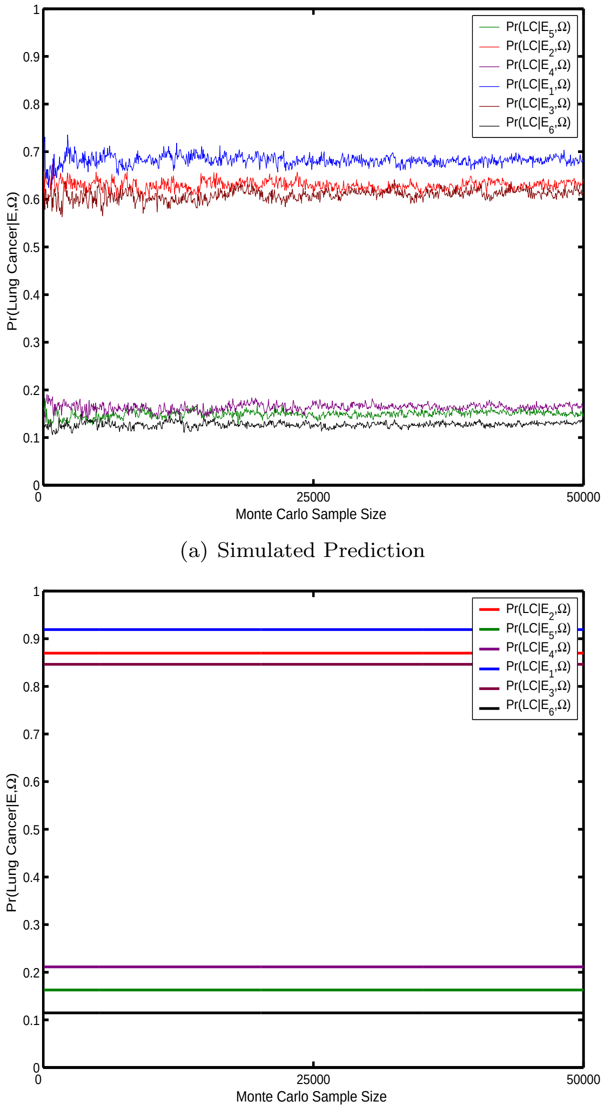
(c) S,B

(d) T/LC,X

(e) T/LC,D

(f) D,B

Figure 2.8: Model Sampling on 2D Parameter Space



(a) Simulated Prediction

(b) True Prediction

Figure 2.9: Prediction Convergence of ASIA Models

Given the qualitative knowledge Ω and associated constraints, we now apply Bayesian inference to predict the incidence of interest in ASIA network in light of construct the Bayesian model class based on the body of qualitative knowledge. As explained in section 2, the qualitative knowledge constraints developed in Eq. 2.54 to Eq. 2.63 are used to define the model priori distribution, i.e. $P(m|\Omega)$. We apply Monte Carlo sampling in each dimension of $\theta = (\alpha, \beta, \gamma, \lambda, \xi)$ as shown in Figure 2.8. The parameter samples are restrained by baseline and extended knowledge features and these samples in each dimension are utilized in our study. They are composed into N=50,000 model samples. By model averaging, we can obtain a single equivalent mean model $\overline{m}$ with mean parameter vector $\overline{\theta}$ which is shown in Table 2.4.

For each selected model sample in Figure 2.8, we perform inferences *in-silico* on the likelihood of a patient having lung cancer given information about the patient's smoking status and clinical evidences including observation of X-ray, Dyspnea, and Bronchitis, i.e. $X_{obs} = \{S, B, XR, D\}$. 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 2.9(a). The evidences are listed below:

$$\begin{aligned} E_1 &= \{S, \overline{B}, XR, D\} & E_2 &= \{S, XR, D\} \\ E_3 &= \{S, B, XR, D\} & E_4 &= \{\overline{S}, \overline{B}, XR, D\} \\ E_5 &= \{\overline{S}, XR, D\} & E_6 &= \{\overline{S}, B, XR, D\} \end{aligned} \tag{2.64}$$

The actual inferences under the same set of evidences is shown in Figure 2.9(b). Comparing Figure 2.9(a) to 2.9(b), we can see that our simulations produce reasonable quantitative predictions on lung cancer probability. The presence of bronchitis could explain away the probability of lung cancer and the presence of smoking increases the risk of getting lung cancer.

2.4 Performance Analysis

2.4.1 Approximating Expected Inference E[P(X|E,m)]

Bayesian networks are directed acyclic graphs in which the nodes represent variables and the arcs signify direct dependencies between the linked variables. Bayesian network can be used to represent generic knowledge of a domain, such as domain expert, qualitative prior knowledge and quantitative observations. As we have shown in section 2.3.3, the belief network can be turned into a computational architecture for directing and activating the data flow in the computations by translating, storing and manipulating the qualitative prior knowledge.

There are various methods to perform belief propagation, i.e. inference. The most common exact inference methods are *variable elimination*, which eliminates (by integration or summation) the non-observed non-query variables one by one by distributing the sum over the product; *clique tree propagation* [81], which caches the computation so that many variables can be queried at one time and new evidence can be propagated quickly; and *recursive conditioning*, which allows for a space-time tradeoff and matches the efficiency of variable elimination when enough space is used. All of these methods have complexity that is exponential in the network's *treewidth*. The most common approximate inference algorithms are stochastic *Markov Chain Monte Carlo (MCMC)* simulation, *mini-bucket elimination* which generalizes loopy belief propagation, and variational methods.

In this section, we will investigate the approximation to the belief propagation in Eq. 2.41 and Eq. 2.53 respectively. Moreover, we would like to determine the degree of performance loss in the approximation.

Approximating Inference in Static Bayesian Networks

In section 2.3, we have derived the belief propagation in the presence of model uncertainty which is constrained by qualitative prior knowledge. The inference in static Bayesian networks, Eq. 2.41, is calculated by taking the averaged inference of each ground static Bayesian model given the new information. We re-write the equation here as

$$P(X|E,\Omega) = \sum_{k=1}^{K} \int_{\Theta_k} d\theta P(X|E,s_k,\theta) P(s_k,\theta|\Omega). \tag{2.65}$$

We can see that $P(X|E,\Omega)$ is actually the expected value of the ground Bayesian network inference given the model uncertainty $P(s,\theta|\Omega)$.

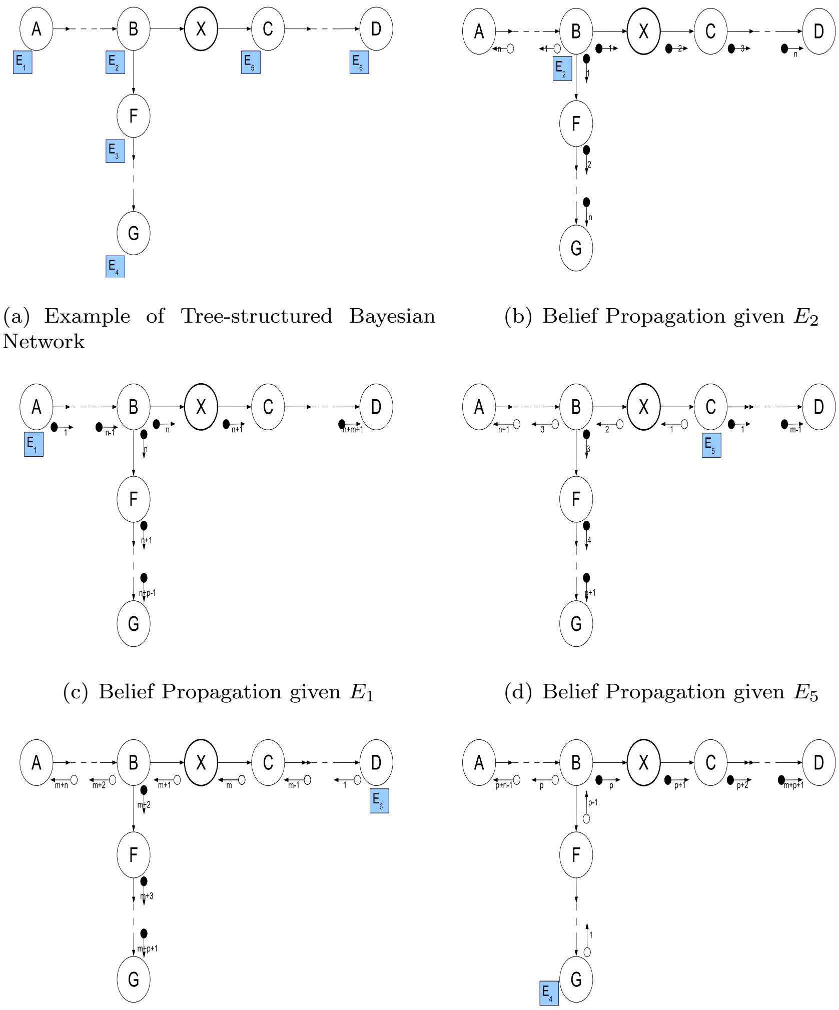
$$P(X|E,\Omega) = E_{P(s,\theta|\Omega)}[P(X|E,s,\theta)] \tag{2.66}$$

Thus, the problem we are addressing is the evaluation of terms of the form $E[f(X)]$. For this approximation problem, we have few choices [37]. The simplest approximation has the form

$$E[f(X)] \approx f(E[X]) \tag{2.67}$$

However, this approximation is only exact when $f(X)$ is *linear* in terms of its argument X [37]. In our case, the function $f(X)$ corresponds to the Bayesian inference function and its argument X is actually the parameter configurations θ in the ground Bayesian network. Unfortunately, the inference function is not always (in most of the cases) a linear function of the network parameters.

In the following, we explore the loss of generalization accuracy by using Eq. 2.67 to approximate $E[P(X|E,s,\theta)]$ based on three types of Bayesian network, i.e. *Tree-structured Belief Network*, *Single-connected Belief Network* and *Multiple-connected Belief Network*. In this thesis, we use Message-propagation algorithm [81] to compute $P(X|E,s,\theta)$.



(a) Example of Tree-structured Bayesian Network

(b) Belief Propagation given E_2

(c) Belief Propagation given E_1

(d) Belief Propagation given E_5

(e) Belief Propagation given E_6

(f) Belief Propagation given E_4

Figure 2.10: Belief Inference in Static Bayesian Network

Tree-structured Belief Networks

We consider tree-structured influence network, i.e. one in which every node except root has exactly one incoming link. We allow each node to represent a multinomial variable which may represent a collection of mutually exclusive hypotheses, such as *Sunny*, *Cloudy*, *Raining* or a set of possible observations, e.g. gene expression level: *High*, *Low*. An example of Tree-structured Bayesian network is shown in Fig. 2.10(a). In this network, a variable is labeled by a capital letter. i.e. $V = \{A, B, C, D, F, G, X\}$. The possible values of each variable are denoted by subscripted letter, e.g. a_1, a_2. Each directed link is quantified by a fixed conditional probability distribution (CPD), e.g. the link from B to X can be described as $\Theta_X = P(X|B)$ with entries: $\theta_{i,j} = P(X = x_i|B = b_j)$. Now, lets assume a set of independent new information $E = \{E_1, E_2, E_3, E_4, E_5, E_6\}$ are inserted into the network at different positions. Our task is to perform inference on the probability of variable X given the evidences in E.

Firstly, we calculate the inference given E_2. Based on the message-propagating algorithm in section 1.2.2, each node in the network stores its prior probability $\pi(V = v_i) = P(V = v_i|D_V^+)$ and the data likelihood $\lambda(V = v_i) = P(D_B^-|V = v_i)$ at initial equilibrium state. The marginal belief of node V in its value v_i can be calculated as

$$BEL(V = v_i) = \alpha\pi(V = v_i)\lambda(V = v_i) \tag{2.68}$$

Firstly, we consider the inference process in the single-connected belief networks.

Single-connected Belief Networks

Upon the arrival of information E_2, the influence of E_2 propagates to its neighbors and activate the updating processing. Two kinds of parameters are stored with each node of the network, i.e. π and λ. The influence of the new information will spread through the network and messages are transmitted through the tree. In Fig. 2.10(b), as soon as node B received the new information, B transmits the π-*message* (black token) to its children and transmit λ-*message* (white token) to its fathers. In the next phase, the triggered fathers and children absorb these tokens and manufacture the appropriate number of tokens for their neighbors, i.e. π-*message* for their children and λ-*message* for their fathers. The links through which the absorbed tokens have entered do not receive new tokens, thus, reflecting the feature that a π-*message* is not affected by a λ-*message* cross the same link. The *message-passing* procedure in case of entering information E_2 is illustrate in Fig. 2.10(b). The numerical annotation of the messages denote the index the updating steps. Assume there are n nodes between B and *(A,G,D)*, the propagation procedure reach new equilibrium after 2n steps. Now we derive the belief of X at equilibrium state in various cases where new information enters the network at different locations by calculating the parameters updated by the information.

Immediate Upstream

In Fig. 2.10(b), the new information E_2 entered the network at node B, i.e. father of node X. In the first propagation step, B transmits message, $\pi_X(B)$, to X. According to section 1.2.2, $\pi_X(B)$ can be written as

$$\pi_X(B) = \alpha\pi(B) \tag{2.69}$$

The π parameter of node X can be calculated as

$$\begin{aligned}\pi(X) &= \sum_B P(X|B)\pi_X(B) \\ &= \sum_B P(X|B)\alpha\pi(B) \qquad (2.70)\end{aligned}$$

and belief of X can be derived as

$$\begin{aligned}BEL(X) &= \pi(X)\lambda(X) \\ &= \sum_B P(X|B)\alpha\pi(B)\lambda(X) \qquad (2.71)\end{aligned}$$

where $\pi(B)$ stands for the fact that multivariate variable B are set to a particular value by the information E_2 and $\lambda(X)$ denotes the λ parameter of node X at initial equilibrium state. α is normalizing constant.

Non-immediate Upstream

In Fig. 2.10(c), the new information E_1 entered the network at node A, i.e. a non-immediate node in the upstream of X. In the first propagation step, A transmits message $\pi_{A+1}(A)$ to its child, say A+1. If we assume there are n nodes from A to X, at *n-th* step, node B are triggered by the incoming message from its father, say $B-1$, and transmits the message $\pi_X(B)$ to node X. Similarly, the π-messages are consequently propagated to the descents of X and B until the network reaches new equilibrium. The belief of node X in this case can be computed sequentially as

$$\pi_{A+1}(A) = \alpha\pi(A) \prod_{sib(A)} \lambda_{sib(A)} \qquad (2.72)$$

where $\pi(A)$ is the π parameter stored in node A at initial equilibrium state and the π parameter of node $A+1$ can be updated by

$$\pi(A+1) = \sum_A P(A+1|A)\pi_{A+1}(A) \qquad (2.73)$$

Thus, we could write the updating message transmitted from node A+i to node $A+i+1$ at *i-th* step as

$$\pi_{A+i+1}(A+i) = \alpha\pi(A+i) \prod_{sib(A+i+1)} \lambda_{sib}(A+i) \qquad (2.74)$$

where operator *sib(*)* denotes the set of siblings of the child node $A+i$ and parameter $\pi(A+i+1)$ can be calculated as

$$\pi(A+i+1) = \sum_{A+i} P(A+i+1|A+i)\pi_{A+i+1}(A+i) \qquad (2.75)$$

After n steps, node B is activated and transmit message $\pi_X(B)$ to X which can be computed as

$$\pi_X(B) = \alpha\pi(B)\lambda_F(B) \qquad (2.76)$$

where $\lambda_F(B)$=$\sum_F \lambda(F)P(F|B)$. The π parameter of X can be written as

$$\pi(X) = \sum_B P(X|B)\pi_X(B) \qquad (2.77)$$

Substituting Eq. 2.74 and Eq. 2.75 into Eq. 2.77, we can derive the belief of X as

$$\begin{aligned} BEL(X) &= \pi(X)\lambda(X) \\ &= \alpha \sum_{B} P(X|B) \cdots \sum_{A} P(A+1|A)\pi(A)\lambda(X)\lambda_F(B) \end{aligned} \qquad (2.78)$$

where $\pi(A)$ stands for the impact of the information E_1 on node A and $\lambda(X)$ denotes the configurations of node X at initial equilibrium state. α is the normalizing factor.

Immediate Downstream

In Fig. 2.10(d), the new information E_5 entered the network at node C, i.e. the child of node X. In the first propagation step, C transmits message, $\lambda_C(X)$, to X which can be written as

$$\lambda_C(X) = \sum_{C} \lambda(C)P(C|X) \qquad (2.79)$$

where $\lambda(C)$=$\{0, \ldots, 1, \ldots, 0\}$ reflects the features of the inserted information E_5. Thus, the λ parameter of X can be updated by $\lambda_C(X)$.

$$\lambda(X) = \lambda_C(X) \qquad (2.80)$$

$\pi(X)$ is not affected by the new information E_5 since there is no impact from the upstream of X imposed in the network. Thus, the belief of node X can be written as

$$\begin{aligned} BEL(X) &= \pi(X)\lambda(X) \\ &= \pi(X) \sum_{C} \lambda(C)P(C|X) \end{aligned} \qquad (2.81)$$

where $\lambda(C)$ are set to a particular value by the information E_5.

Non-immediate Downstream

In Fig. 2.10(e), the new information E_6 entered the network at node D, i.e. a non-immediate node in the downstream of X. In the first propagation step, D transmits message $\lambda_D(D-1)$ to its parent, say $D-1$. If we assume there are m nodes from D to X, at *m-th* step, node C are triggered by the incoming message from its child, say $C+1$, and transmits the message $\lambda_C(X)$ to node X. If there are n nodes between node X and node A, at *(m+n)-th* step, A is activated by its child, say A+1 and start transmitting π-*message* on the same link. Similarly, the π-messages are consequently propagated to the descents of A and B until the network reaches new equilibrium. The belief of node X in this case can be calculated by $\lambda(X)$ and $\pi(X)$. As Eq. 2.74 and 2.75, λ message can be calculated sequentially as

$$\lambda_{D-i}(D-i-1) = \sum_{D-i} \lambda(D-i)P(D-i|D-i-1) \qquad (2.82)$$

and the λ parameter of node (D-i) can be calculated as

$$\lambda(D-i) = \prod_{k} \lambda^{k}_{D-i}(D-i-1) \qquad (2.83)$$

where $D-i$ is the *i-th* node in the upstream of node D and $D-i-1$ denotes the parent node of node $D-i$. If node $D-i-1$ has more than one child node, i.e. $D-i$ is its *k-th* child node, then, $\lambda(D-i)$ is a product of the λ-message from its children as Eq. 2.83. $\lambda(X)$ can be written as

$$\begin{aligned}\lambda(X) &= \sum_{C} \lambda(C)P(C|X) \\ &= \sum_{C} P(C|X) \cdots \sum_{D} P(D|D-1)\lambda(D)\end{aligned} \tag{2.84}$$

where $\lambda(D)$ represent the new information E_6. Meanwhile, $\pi(X)$ is not affected by the new information and we could derive BEL(X) as

$$\begin{aligned}BEL(X) &= \pi(X)\lambda(X) \\ &= \pi(X) \sum_{C} P(C|X) \cdots \sum_{D} P(D|D-1)\lambda(D)\end{aligned} \tag{2.85}$$

Non-immediate Siblings

In Fig. 2.10(f), the new information E_4 entered the network at node G, i.e. a non-immediate node in the downstream of X's sibling F. In the first propagation step, G transmits message $\lambda_G(G-1)$ to its parent, say $G-1$. If we assume there are p nodes from G to X, at (p-1)-th step, node B are triggered by the incoming message from its parent F, and transmits the message $\lambda_B(B-1)$ to its parent $B-1$ as well as transmit $\pi_X(B)$ to its child node X. $\lambda(X)$ is not affected by the information E_4. The effect of new information of E_4 on node X is reflected by the update of $\lambda(F)$ which is enclosed in the $\pi_X(B)$ message transmitted from B to X. Based on Eq. 2.82 and 2.83, we can calculate $\lambda(F)$ as

$$\begin{aligned}\lambda_{F+1}(F) &= \sum_{F+1} \lambda(F+1)P(F+1|F) \\ &= \sum_{F+1} P(F+1|F) \cdots \sum_{G} P(G|G-1)\lambda(G) \\ &= \lambda(F)\end{aligned} \tag{2.86}$$

where $F+1$ is the child of node F and $G-1$ denotes the parent of node G. Meanwhile, $\pi_X(B)$ can be calculated according to Eq. 2.76 and $\lambda_F(B)$ can be calculated as

$$\begin{aligned}\lambda_F(B) &= \sum_{F} \lambda(F)P(F|B) \\ &= \sum_{F} P(F|B) \sum_{F+1} P(F+1|F) \cdots \sum_{G} P(G|G-1)\lambda(G)\end{aligned} \tag{2.87}$$

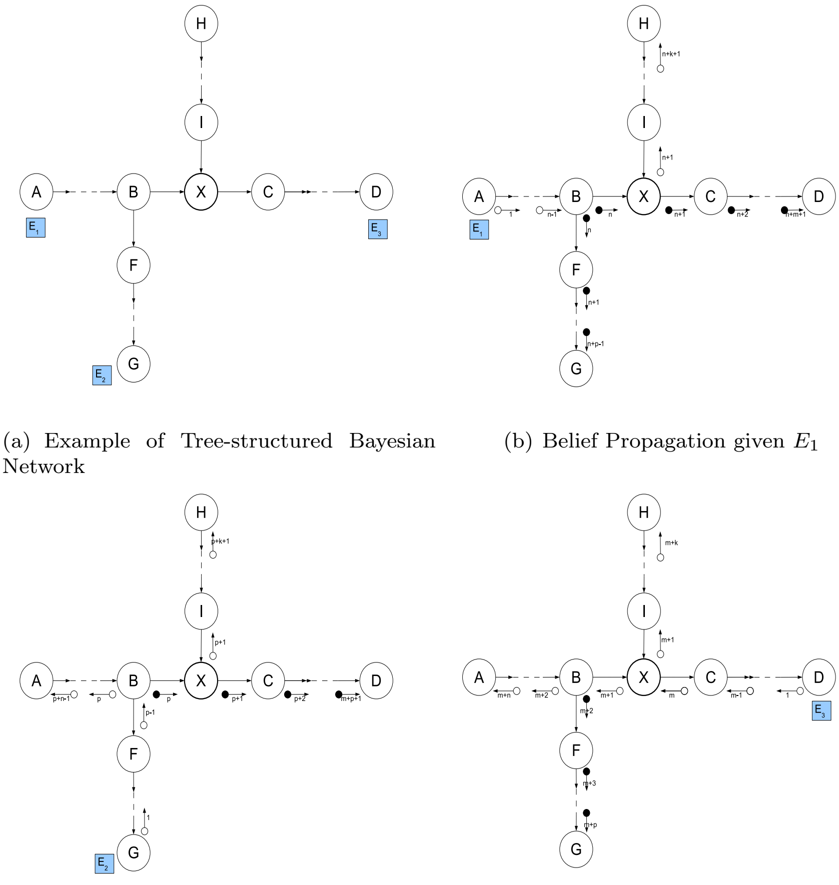
where $\lambda(G)$ represent the new information E_4. Therefore, $\pi(X)$ can be formalized as

$$\begin{aligned}\pi(X) &= \sum_{B} P(X|B)\pi_X(B) \\ &= \sum_{B} P(X|B) \cdots \sum_{A} P(A+1|A)\pi(A)\lambda_F(B)\end{aligned} \tag{2.88}$$

The belief of X can be written as

$$\begin{aligned} BEL(X) &= \pi(X)\lambda(X) \\ &= \sum_B P(X|B)\cdots\sum_A P(A+1|A)\pi(A)\sum_F P(F|B) \\ &\quad \cdots\sum_G P(G|G-1)\lambda(G)\lambda(X) \qquad (2.89) \end{aligned}$$

where $\lambda(G)$ represent the new information E_4 and $\lambda(X)$ is the λ parameter of X at initial equilibrium state.



(a) Example of Tree-structured Bayesian Network

(b) Belief Propagation given E_1

(c) Belief Propagation given E_2

(d) Belief Propagation given E_3

Figure 2.11: Belief Inference in Static Bayesian Network

Single-connected Belief Network with Multiple Parents
Now, we discuss the belief propagation scheme in the single-connected network which allows a child node has more than one parent. In Fig. 2.11(a), new information $\{E_1, E_2, E_3\}$ are inserted into the network at various locations. We derive the belief of node X respectively.

Non-immediate Upstream
In Fig. 2.11(b), new information E_1 enter the network at node A. Assume there

are n nodes from A to X, then at n-th step, node B is triggered to send $\pi_X(B)$ message to node X. Thus, the belief of node X is updated by $\pi_X(B)$. According to section 1.2.2, BEL(X) can be calculated as

$$BEL(X) = \alpha\lambda_C(X)\sum_B\sum_I P(X|B,I)\pi_X(B)\pi_X(I) \tag{2.90}$$

$\pi_X(B)$ can be calculated as Eq. 2.76 and $\lambda_C(X)$ can be computed as Eq. 2.79. Thus, the belief of node X can be written as

$$\begin{aligned} BEL(X) &= \alpha\sum_C \lambda(C)P(C|X)\sum_B\sum_I P(X|B,I) \\ &\quad \cdots\sum_A P(A+1|A)\pi(A)\lambda_F(B)\pi_X(I) \end{aligned} \tag{2.91}$$

where $\lambda(C)$ is set at initial equilibrium state and $\pi(A)$ is determined by the new information E_1.

Non-immediate Siblings

In Fig. 2.11(c), new information E_2 enter the network at node G. Assume there are p nodes from G to X, then at p-th step, node B is triggered to send $\pi_X(B)$ message to node X. Thus, the belief of node X is updated by $\pi_X(B)$. According to section 1.2.2, BEL(X) can be calculated as

$$BEL(X) = \alpha\lambda_C(X)\sum_B\sum_I P(X|B,I)\pi_X(B)\pi_X(I) \tag{2.92}$$

where $\pi_X(B)$ can be calculated as Eq. 2.76 and $\lambda_F(B)$ can be computed as Eq. 2.87. Thus, the belief of node X can be written as

$$\begin{aligned} BEL(X) &= \alpha\sum_C \lambda(C)P(C|X)\sum_B\sum_I P(X|B,I)\sum_F P(F+1|F) \\ &\quad \cdots\sum_G P(G|G-1)\lambda(G)\pi_X(I) \end{aligned} \tag{2.93}$$

where $\lambda(C)$ is set at initial equilibrium and $\lambda(G)$ is decided by new information E_2.

Non-immediate Downstream

In Fig. 2.11(d), new information E_3 enter the network at node D. Assume there are m nodes from D to X, then at m-th step, node C is triggered to send $\lambda_C(X)$ message to node X. Thus, BEL(X) can be calculated as

$$BEL(X) = \alpha\lambda_C(X)\sum_B\sum_I P(X|B,I)\pi_X(B)\pi_X(I) \tag{2.94}$$

where $\lambda_C(X)$ can be calculated as Eq. 2.82 and 2.83 iteratively as

$$\lambda_C(X) = \sum_C P(C|X)\cdots\sum_D P(D|D-1)\lambda(D) \tag{2.95}$$

Thus, the belief of node X can be written as

$$BEL(X) = \alpha\sum_C P(C|X)\cdots\sum_D P(D|D{-}1)\lambda(D)\sum_B\sum_I P(X|B,I)\pi_X(B)\pi_X(I) \tag{2.96}$$

where $\lambda(D)$ is decided by new information E_3.

Multiple-connected Belief Networks

The efficiency of message passing algorithm in single-connected network has been shown in the above. The question raised thereafter is whether similar propagation schemes can be applied to less restrictive networks, i.e. *Multiple-connected network* where multiple parents of common children may processes common ancestors, thus forming loops in the underlying network. If we ignore the existence of loops and permits the nodes to continue communicating with each other as if the network were singly connected, messages may circulate indefinitely around these loops, and the process will not converge to the correct state of equilibrium.

A straightforward way of handling the loop would be to appoint a local interpreter for the loop by collapsing nodes (B,C). This method works well on small loops [55], but as soon as the number of variables exceeds 3 or 4, compounding requires handling huge matrices which impede the natural conceptual structure embedded in the original network.

A second method of propagation is based on "stochastic relaxation" [41]. Similar to *Boltzman machine* [32]. Each processor examines the states of the variables within its screening neighbors, computes a belief distribution for the values of its host variable, and then randomly selects one of these values with probability given by the computed distribution. The value chosen will subsequently be interrogated by the neighbors upon computing their beliefs, and so on. This scheme is guaranteed convergence, but it usually requires very long relaxation times before reaching the equilibrium state.

A third method, *conditioning* [58], is based on transferring the connectivity of a multiple-connected network to a singly-connected network. We instantiate a variable to a particular value which enables single-connected belief propagation techniques. Assume binomial node, we set this variable, e.g. node Y, to 1 and calculate the belief of node X, *BEL*$_1$*(X)* as single-connected network. Meanwhile, we set this variable to 0 and compute the belief of node X, *BEL*$_0$*(X)*. Two results are combined by the marginal posterior probability $P(X|Y)$. Although *conditioning* method provides a working solution in practical cases, the number of messages may grow exponentially with the number of nodes required for breaking up all loops in the network.

Finally, a preprocessing approach permanently changes the multiple-connected network into a *star-decomposable* and *tree*-decomposable tree by introducing *dummy variable.*

Approximating Inference in Dynamic Bayesian Networks

For the belief propagation in dynamic Bayesian networks, we would like to predict future outcomes given all the observations up to the present time, $y_{1:t} = \{y_1, \ldots, y_t\}$ which is a common task in on-line analysis. In this thesis, we only consider discrete-time systems, hence t is always an integer. Since we will generally be unsure about the future, we will try to compute a probability distribution over the possible future observations; We denote this by $P(y_{t+h}|y_{1:t})$, where $h > 0$ is the horizon, i.e. how far into the future we want to predict. Sometimes we have some control over the system we are monitoring, e.g. some external stimuli. In this case, we would like to predict future outcomes as a function of our inputs. Let $E_{1:t}$ denote our past inputs, and $E_{t+1:t+h}$ denote our next

h inputs. Now the task is to compute $P(y_{t+h}|E_{1:t+h}, y_{1:t})$. Classical approaches to time-series prediction use linear models, such as ARIMA, ARMAX, etc. [45], or non-linear models, such as neural networks (either feed-forward or recurrent) or decision trees [19]. For discrete data, it is common to use n-gram models (see e.g. [56]) or variable-length Markov models [70, 94]. There are several problems with the classical approach. First, we must base our prediction of the future on only a finite window into the past, say where is the lag, if we are to do constant work per time step. If we know that the system we are modeling is Markov with an order, we will suffer no loss of performance, but in general the order may be large and unknown. Recurrent neural nets try to overcome this problem by using internal state, but they are still not able to model long-distance dependencies [3]. Second, it is difficult to incorporate prior knowledge into the classical approach: much of our knowledge cannot be expressed in terms of directly observable quantities, and black-box models, such as neural networks, are notoriously hard to interpret. Third, the classical approach has difficulties when we have multi-dimensional (multi-variate) inputs and/or outputs. For instance, consider the problem of predicting (and hence compressing) the next frame in a video stream using a neural network. Actual video compression schemes (such as MPEG) try to infer the underlying "cause" behind what they see, and use that to predict the next frame. For attacking these problems in general, Kevin [78] has suggested the state-space models, which we use in out thesis to perform temporal predictions. Further, in this thesis, since our focus is on how to model uncertainty based on qualitative hypotheses for (Dynamic Bayesian) networks, we assume the state-space model to be Markovian, i.e. Dynamic Bayesian networks. In this case, the future prediction y_{t+1} (h=1) is soley dependent on the current states of the system, y_t, therefore, the prediction boils down to compute the probability distribution of $P(y_{t+1}|E_{1:t+h}, y_t)$ as shown in Eq. 2.53.

The inference in Eq. 2.53 can be calculated in two ways. Firstly, the inference can be performed for each model $m_k = (s_k, \theta)$ over time T and the predictions are averaged over all models in class $\widetilde{M}$. Secondly, an averaged model, i.e. *equivalent mean model*, can be calculated by calculating the average of the parameters over all possible models under each model structure, $m_k \in \widetilde{M}$, and then the inference is computed as the averaged inference of each mean model with different structure as

$$\begin{aligned} & P(X_{n,t}|E,\Omega) \\ = & \sum_{k=1}^{K} \int_{\Theta} \sum_{j=1}^{J} \theta_{k,j} P_j^{t-1}(\pi(X_n), E) P(s, \theta_{k,j}|\Omega) d\Theta \\ = & \sum_{k=1}^{K} \sum_{j=1}^{J} Pr_j^{t-1}(\pi(X_n), E) \int_{\Theta} \theta_{k,j} P(s, \theta_{k,j}|\overline{\Omega}) d\Theta \end{aligned} \tag{2.97}$$

Approximating E[P(X|E,m)] by P(X|E,$\overline{m}$)

Back to the approximation problem in Eq. 2.67, the inference function, f(X), equals to BEL($\overline{\theta}$) in Eq. 2.71 to Eq. 2.96. It is obvious to conclude that the inference on node X given evidence E and each Bayesian network $m_k(s_k, \overline{\theta}_k)$ in the class $\overline{M}$=$\{m_k|k = 1, \ldots, K\}$ is a multinomial nonlinear function in terms of $\overline{\theta}_k$

and the distance between the observation and information insertion. For example, in Fig. 2.10(c), information E_1 is inserted on node A and belief propagation is observed at node X. The belief of node X can be written as

$$BEL(X) = \alpha \sum_B P(X|B) \cdots \sum_A P(A+1|A)\pi(A)\lambda(X)\lambda_F(B) \tag{2.98}$$

If we assume binomial distribution, $\pi(A)$=$\{0,1\}$ stands for the fact that information E_1 set the multivariate node A to a particular value. $\lambda(X)$ denotes the configurations of node X at initial equilibrium state, e.g. $\lambda(X)$=$\{1,1\}$ and $\lambda_F(B)$ represents the belief impact of node F onto node B. α is a normalizing factor. Assume we have obtained a class of Bayesian networks according to section 2.3, we describe the *k-th* Bayesian network with structure s(Fig. 2.10(c)) and *k-th* parameter configurations $\overline{\Theta}_k$. $\overline{\Theta}_k$ denotes a vector of conditional probability distribution of each node given its parents in the network, i.e. $\overline{\Theta}_k$=$\{\Theta^k_{X|B}, \ldots, \Theta^k_{(A+1)|A}\}$. For example, $\Theta^k_{X|B}$ is a conditional probability table (CPT) and can be represented by a matrix as

$$\Theta^k_{X|B} = \begin{pmatrix} P^k(X|B) & P^k(\overline{X}|B) \\ P^k(X|\overline{B}) & P^k(\overline{X}|\overline{B}) \end{pmatrix} \tag{2.99}$$

Each item in $\Theta^k_{X|B}$ is a random variable with constraint of summation of each row in the matrix equals to 1, i.e. P(X|B)+P($\overline{X}$|B)=1 and P(X|$\overline{B}$)+P($\overline{X}$|$\overline{B}$)=1. Thus, the belief inference in Eq. 2.99 can be described by the conditional probability table as

$$\begin{aligned} BEL(X) &= \pi(A)\lambda(X)\lambda_F(B)\Theta^k_{X|B} \cdots \Theta^k_{(A+1)|A} \\ &= \pi(A)\lambda(X)\lambda_F(B)\prod_{i=1}^{n}\Theta^k_i \end{aligned} \tag{2.100}$$

where n denotes the number of steps for belief to propagate from A to X as shown in Fig. 2.10(c). It can be expanded as

$$\begin{aligned} BEL(X) &= \pi(A)\lambda(X)\lambda_F(B)\prod_{i=1}^{n}\Theta^k_i \\ &= \pi(A)\lambda(X)\lambda_F(B)\sum_{m=1}^{2^n}(\theta^k_1 \cdots \theta^k_n)_m \end{aligned} \tag{2.101}$$

where $(\theta^k_1 \cdots \theta^k_n)_m$ denotes the *m-th* configuration of the element-wide product $\prod_{i=1}^{n}\Theta^k_i$ corresponding to the non-zero value in $\pi(A)$ and θ^k_i is an element of matrix Θ^k_i. Eq. 2.101 is a polynomial function in form of $O(\theta^n)$.

We construct a polynomial function $f(\overline{X}, N)$ to mimic the belief propagation in Eq. 2.101 with arguments $\overline{X}$ and n.

$$f(\overline{X}, n) = \sum_{m=1}^{M}(X_m)^n \tag{2.102}$$

where $\overline{X}$ is a vector of random variables, $\overline{X}$=$\{X_1, X_2, \ldots, X_M\}$ and M=2^n. For example,

$$\begin{aligned} n &= 1, f = X_1 + X_2 \\ n &= 2, f = X_1^2 + X_2^2 + X_3^2 + X_4^2 \end{aligned}$$

In case of n=1, f is equivalent to one step forward/backward belief propagation and is a linear function of $\overline{X}$. Whereas in case of n=2, f is a multivariate quadratic function.

In the following discussion, we study correlation between distributions of function f and $\overline{X}$ and analyze the approximation accuracy of expected value of function $E[f(\overline{X})]$ by $f(E[\overline{X}])$. The expected value of f can be written as

$$E[f(\overline{X}, n)] = \frac{1}{K}\sum_{k=1}^{K}\sum_{m=1}^{M}(X_m^k)^n \tag{2.103}$$

and the function f of the expected argument can be written as

$$f(E[\overline{X}], n) = \sum_{m=1}^{M}(\frac{1}{K}\sum_{k=1}^{K}(X_m^k))^n \tag{2.104}$$

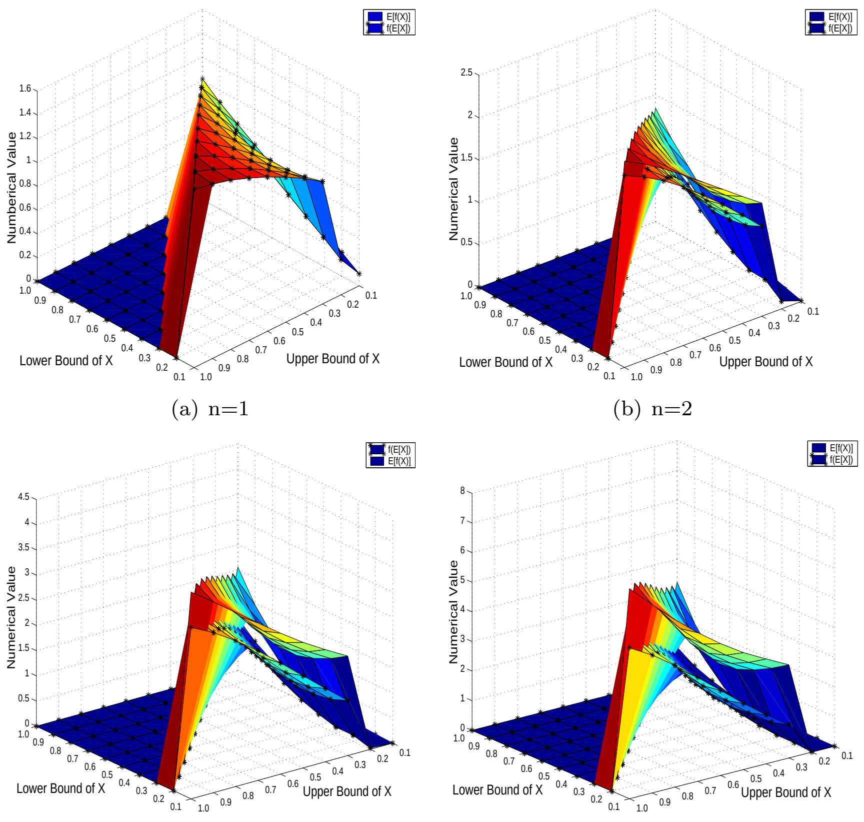
In this study, we draw K samples of the vector $\overline{X}$ uniformly distributed in [0,1]. We constrain $\overline{X}$ by setting its upper and lower bounds as $\{b_1 \leq X_m \leq b_2 | m = 1, \ldots, M\}$ and vary the bound with restraints $\{0 \leq b_1 \leq b_2 \leq 1\}$. The values of $E[f(\overline{X}, n)]$, $f(E[\overline{X}], n)$ based on various settings of $\overline{X}$ are shown in Fig. 2.12. In Fig. 2.12, the deviation between $E[f(X)]$ and $f(E[X])$ increases proportionally to n for the same set of $\overline{X}$ which demonstrate that message-passing belief propagation scheme will raise the aberration in proportion to the number of inference steps n (If we deem f as belief inference function in Eq. 2.101).

On the other hand, the various distribution of $\overline{X}$ in the range of [0,1] may affect the distance between $E[f(X)]$ and $f(E[X])$. We investigate the values of $E[f(X)]$ and $f(E[X])$ based on different distribution of $\overline{X}$ in [0,1] and numerical results are shown in Fig. 2.13. From this result, we can conclude that $E[f(X)]$ and $f(E[X])$ approach to each other as the uncertainty of variables in $\overline{X}$ decreases, i.e. with tighter constraints in the space of $\overline{X}$. If we consider f as belief inference function, the expected inference results converge to its approximation with decreasing uncertainty on the parameter space. Therefore, if we set up a tight enough constraints by applying our qualitative knowledge model in the parameter space, it is probably reasonable to approximate the expected inference results by Eq. 2.67.

2.4.2 Robust Analysis

From Eq. 2.41 and Eq. 2.53, it is clear that generalization accuracy largely depends on the accuracy of the model uncertainty including the structure space uncertainty, $P(s|\Omega)$ and the structure-dependent parameter space uncertainty, $P(\theta|s, \Omega)$. In the bottom-up probabilistic modeling framework based on the data, these two distributions are learned given the data and usually the item with maximized score is selected as the learned model which are used in the future generalization task, such as maximum a posterior probability, i.e. $\widehat{s} =$ argmax$\{P(s|D)\}$ and $\widehat{\theta}$=argmax$\{P(\theta|s, D)\}$.

In the top-down probabilistic modeling framework, the model uncertainty is constructed based on the qualitative prior knowledge and the qualitative knowledge model which are defined in section 2.1. The qualitative prior knowledge are feed into the qualitative knowledge model and the semantics of the knowledge



(a) n=1

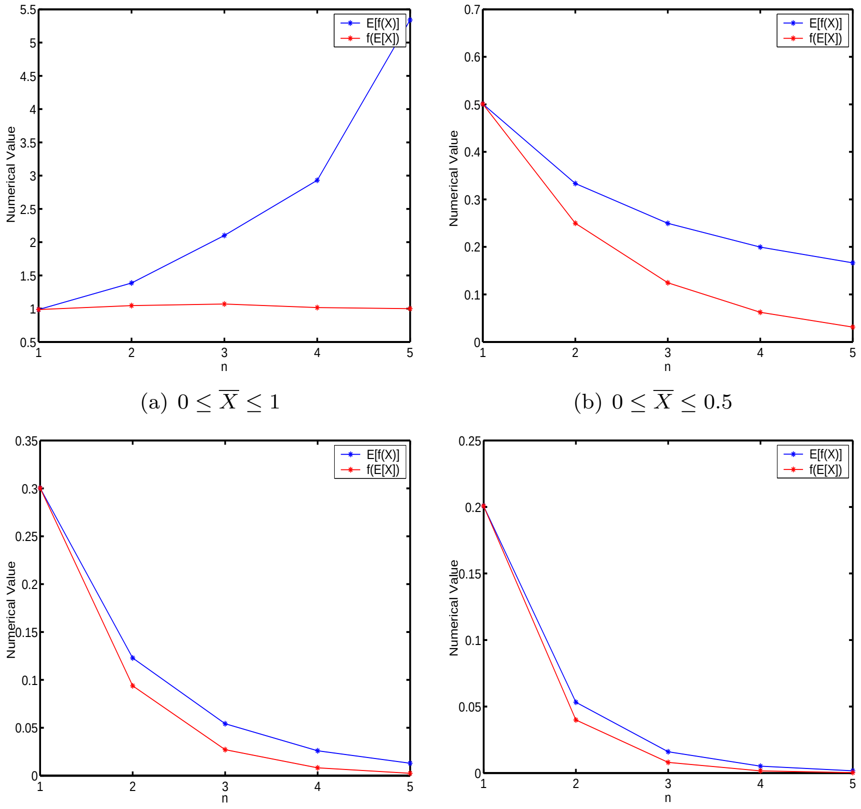
(b) n=2

(c) n=3

(d) n=4

Figure 2.12: E[f(X)] and f(E(X))

are translated into a set of structural representations and a set of inequalities in the structure-dependent parameter space by a vector of knowledge features. Comparing to the bottom-up approaches, full Bayesian approach assigns non-zero probability to a set of models which are consistent with the body of the knowledge instead of taking single model for granted. Each model is used to generate inference and the expected value of these inferences is calculated by model averaging weighted with model uncertainty. In case of consistent qualitative knowledge, the models are assumed to be uniformly distributed in the model space which are confined by the set of structural representations and parameter inequalities. Thus, the generalization accuracy of our top-down approach largely depends on the accuracy of the constructed model uncertainty given the qualitative prior knowledge, i.e. $P(s|\Omega)$ and $P(\theta|s,\Omega)$. Moreover, as we have discussed in section 2.4.1, if we approximate the expected inference across model space by the inference in the expected model, i.e. $P(X|s,E[\theta],\Omega)$, the generalization accuracy will be further adjusted by this linear/non-linear approximation. Therefore, the quantitative inference accuracy is exclusively dependent on the accuracy of the constructed model uncertainty and the approximation of the inference. In section 2.4.1, we have studied the generalization accuracy given consistent qualitative knowledge as a consequence of the inference approxima-



(a) $0 \leq \overline{X} \leq 1$

(b) $0 \leq \overline{X} \leq 0.5$

(c) $0 \leq \overline{X} \leq 0.3$

(d) $0 \leq \overline{X} \leq 0.2$

Figure 2.13: E[f(X)] and f(E(X)) with various $\overline{X}$

tion. In this section, we will discuss the generalization accuracy given consistent qualitative knowledge as a consequence of the accuracy of model uncertainty. Noisy information often exists largely in a qualitative statement. For example, a domain expert may feel comfortable to express the probability of getting lung cancer for a smoker *likely* ranges from 10% to 15%. In this statement, parameter uncertainty on the conditional probability of lung caner given smoking is given by the boundary information: [10%,15%]. However, the word *likely* express a second-order uncertainty on the boundary information, i.e. uncertainty over the bounded uncertainty. In this thesis, we refer this second-order uncertainty to noise in the knowledge and use Gaussian distribution function to represent this kind of noise. We perform our robust analysis in 2-dimensional parameter space by varying the second-order model uncertainty to different extents. Similar analysis can be carried out in higher-dimensional parameter space.

Robust Analysis in 2-Dimensional Parameter Space

Lets consider a simple qualitative hypothesis:*A causes B*. The structure can be represented by s_4 in Fig. 2.4(a) and the 2-dimensional conditional parameter space is composed by $Y = P(B|A)$ and $X = P(B|\overline{A})$. Now, if we use the expected model to perform quantitative inference, the expected model with averaged parameter can be described as m=$\{s_4,\overline{Y},\overline{X}\}$ where $\overline{Y}$ and $\overline{X}$ denote the averaged value of $P(B|A)$ and $P(B|\overline{A})$. The distance between any two parameters can be defined as

$$d = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2} \tag{2.105}$$

Now, the expected inference given the evidence $e = \{A = 1\}$ reads

$$\begin{aligned} E[P(B = 1|e, X, Y)] &= \int_X \int_Y P(B = 1|e, X, Y)P(X, Y)dXdY \\ &= \int_0^1 dX \int_0^1 YdY \\ &= E[Y] \end{aligned} \tag{2.106}$$

Similarly, the expected inference given the evidence $e = \{A = 0\}$ can be calculated as

$$\begin{aligned} E[P(B = 1|e, X, Y)] &= \int_X \int_Y P(B = 1|e, X, Y)P(X, Y)dXdY \\ &= \int_0^1 XdX \int_0^1 dY \\ &= E[X] \end{aligned} \tag{2.107}$$

Since $P(B = 1|e, X, Y)$ is a linear function of X and Y in this case, the expected inference equals to the inference in the expected model, i.e.

$$E[P(B = 1|e, X, Y)] = P(B = 1|e, E[X], E[Y]) \tag{2.108}$$

The *Root Mean Square Error* (RMSE) between the expected inference and the actual inference can be described by the averaged distance between the inference in expected model and the actual inference,

$$RMSE = E[\sqrt{\sum_{e_1,e_2} (P(B|e, E[X], E[Y]) - P(B|e, X_i, Y_i))^2}] \tag{2.109}$$

where X_i and Y_i are the uniformly distributed actual parameters in the parameter space and $e_1 = \{A = 1\}$, $e_2 = \{A = 0\}$. Substituting Eq. 2.106 and 2.107 into 2.109, we have

$$RMSE = E[\sqrt{(E[X] - X_i)^2 + (E[Y] - Y_i)^2}] \tag{2.110}$$

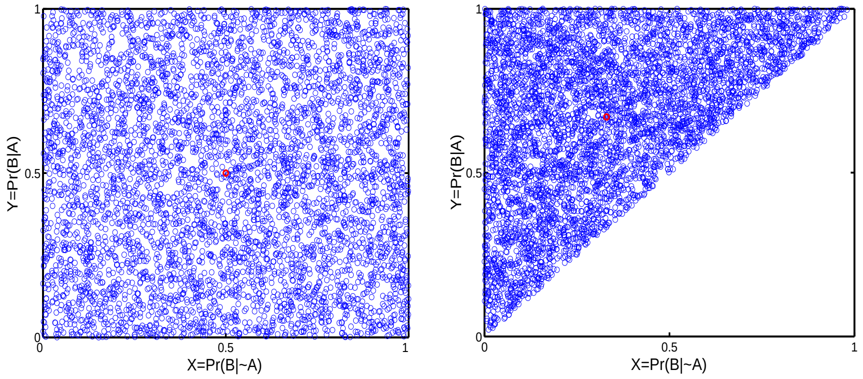
Comparing Eq. 2.110 to Eq. 2.105, we can conclude that the RMSE of the inference results is proportional to the averaged distance between the actual parameters and expected parameter values, i.e.

$$RMSE = E[d(X_i, Y_i, E[X], E[Y])] \tag{2.111}$$

The averaged distance between the actual parameter and the averaged parameter can be calculated by Monte Carlo method.

$$E[d(X_i, Y_i, E[X], E[Y])] \approx \frac{1}{M} \sum_{i=1}^{M} d(X_i, Y_i, E[X], E[Y]) \tag{2.112}$$

In extreme case, there is no constraint imposed onto the parameter space, therefore, the mean parameters, $E[X]$=0.5 and $E[Y]$=0.5, are shown by the red cycle in Fig. 2.14(a). Meanwhile, the actual parameter can be located anywhere in the parameter space as shown by the blue cycles in Fig. 2.14(a). By applying Eq. 2.112, RMSE=0.3863. In the second case, we translate the qualitative hypothesis "*A causes B*" by the baseline knowledge feature in section 2.1. *A cause B* can be translated as A active makes B active more likely. A constraint is introduced onto the parameter space by inequality $Y_i \geq X_i$. The mean parameters can be calculated as Eq. 2.46 and 2.47 which equals to $E[X] = \frac{1}{3}$ and $E[Y] = \frac{2}{3}$. It is shown by the red cycle in Fig. 2.14(b) as well the actual parameters are uniformly distributed across the up-triangular area of the parameter space as shown by the blue cycles in Fig. 2.14(b). By applying Eq. 2.112, RMSE=0.30. In the third case, we assume the qualitative hypothesis includes not only the baseline information but also the extended information *ratio*. An straightforward example of such hypothesis can be "*A causes B by more that R_1 times but less than R_2 times*". Then, the parameter space is restrained by inequality $R_1X \leq Y \leq R_2X$ as shown in Fig. 2.15(a) and the mean parameters can be calculated as



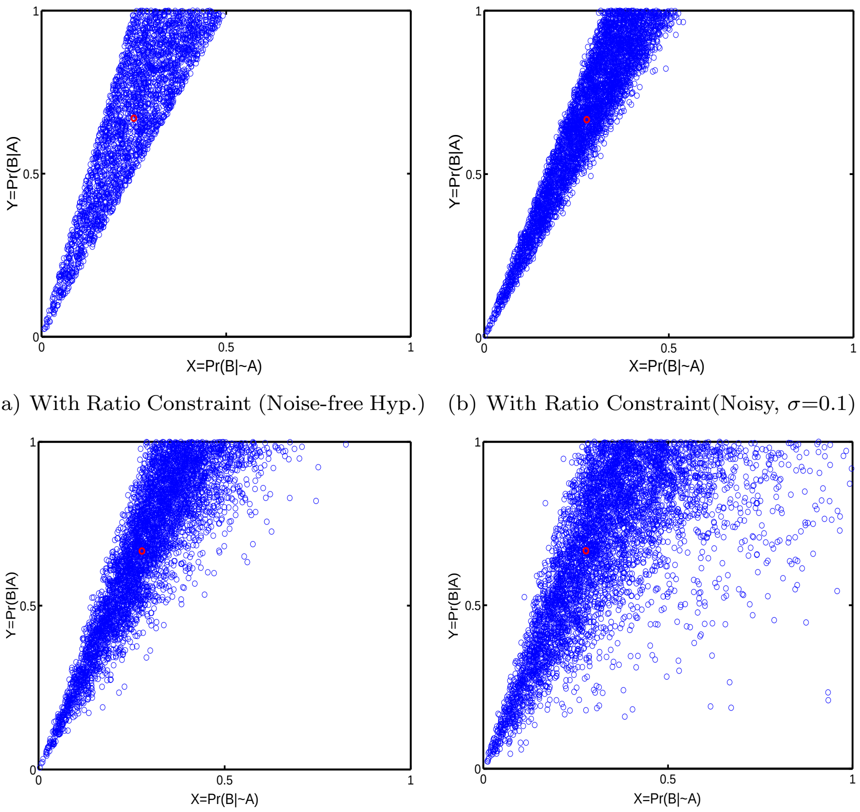
(a) Without Constraints (b) With Baseline Constraint

Figure 2.14: 2-Dimensional Parameter Space with Consistent Hypothesis

$$\begin{aligned} E[X] &= P\int_0^1 dY \int_{\frac{Y}{R_2}}^{\frac{Y}{R_1}} XdX \\ &= \frac{P}{6}(\frac{1}{{R_1}^2} - \frac{1}{{R_2}^2}) \end{aligned} \tag{2.113}$$

and

$$\begin{aligned} E[Y] &= P\int_0^1 YdY \int_{\frac{Y}{R_2}}^{\frac{Y}{R_1}} dX \\ &= \frac{P}{3}(\frac{1}{R_1} - \frac{1}{R_2}) \end{aligned} \tag{2.114}$$



(a) With Ratio Constraint (Noise-free Hyp.) (b) With Ratio Constraint(Noisy, σ=0.1)

(c) With Ratio Constraint(Noisy, σ=0.3) (d) With Ratio Constraint(Noisy, σ=0.5)

Figure 2.15: 2-Dimensional Parameter Space with Consistent Hypothesis (Cont.)

where P is the normalizing constant which satisfy

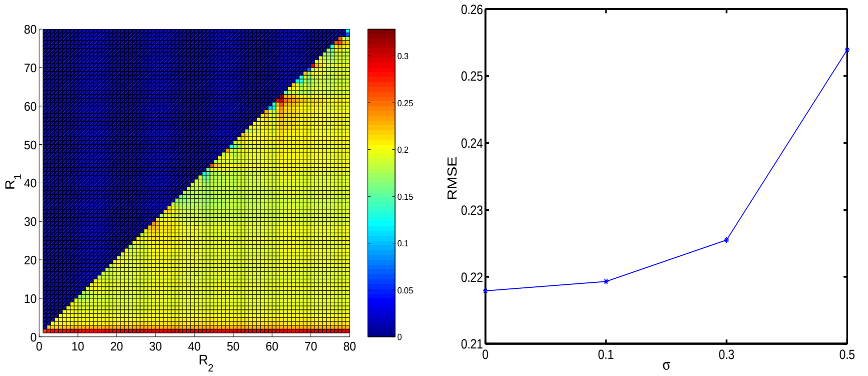
$$P\int_0^1 dY \int_{\frac{Y}{R_2}}^{\frac{Y}{R_1}} dX = 1 \tag{2.115}$$

Then, we have P=$\frac{2R_1R_2}{(R_2-R_1)}$. By substituting P into Eq. 2.113 and 2.114, we derive the mean parameters $E[X] = \frac{R_1+R_2}{3R_1R_2}$ and $E[Y] = \frac{2}{3}$ which is shown by the red cycle in Fig. 2.15(a). The RMSE of the inference with mean parameters to the actual inference can be computed by

$$RMSE \approx \frac{1}{M}\sum_{i=1}^{M}\sqrt{(\frac{2}{3}-Y_i)^2+(\frac{R_1+R_2}{3R_1R_2}-X_i)^2} \tag{2.116}$$

If we further assume the hypothesis is noise-free, i.e. the actual parameters are uniform distributed in the constrained area of parameter space exclusively, the RMSE is a function of R_1 and R_2. We compute the averaged RMSE by varying R_1 and R_2 in the range [1,80] and the result is shown in Fig. 2.16(a). When $R_1 = 1$ and $R_2 \Rightarrow \infty$, the constrained area of parameter space set by

R_1 and R_2 converge to the up-triangular area in the second case as well as $E[X] \Rightarrow \frac{1}{3}$ and $RMSE \Rightarrow 0.30$. No parameters are sampled in the area where $R_1 \geq R_2$ since it violates the body of the qualitative hypothesis. Thus, the averaged RMSE in this area is set to zero. The mean RMSE of the lower-triangle parameter space in Fig. 2.16(a) equals to 0.23. If the hypothesis is noisy, i.e. the actual parameters may not be exclusively distributed in the area constrained by the inequality, instead, the actual parameters are guided by a noisy version of the inequality. We represent the noise by a Gaussian distribution over the constrained parameter space. The noisy constraint can be described as



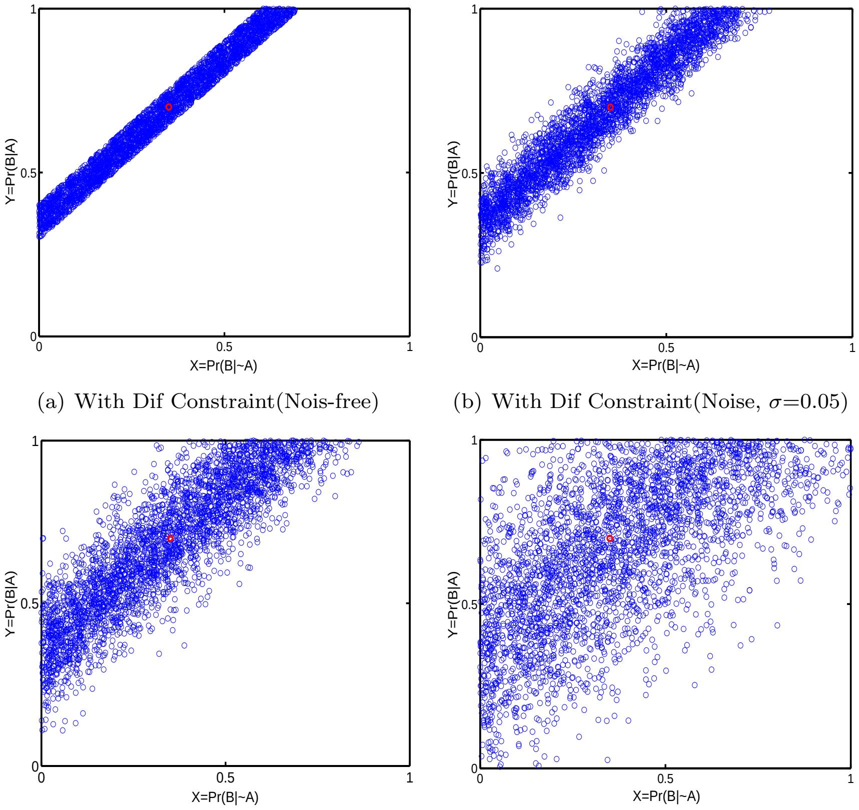
(a) Varying R_1 and R_2 (Noise-free Hyp.) (b) Varying Noise Strength σ (R_1=2,R_2=3)

Figure 2.16: RMSE Distance Measure with Ratio Constraint

$$(R_1 + \sigma N_1)X \leq Y \leq (R_2 + \sigma N_2)X \tag{2.117}$$

where N_1 and N_2 represent samples from the Gaussian distribution $N(0,1)$ with zero mean and unit variance and σ is a constant representing the noise strength. We first draw a vector of noise $\overline{N} = \{N_i | i = 1, \ldots, K\}$ from $N(0,1)$. For each N_i, we construct the inequality constraint in Eq. 2.117 respectively and sample M parameters in the restrained parameter space. Thus, we obtain a set of parameter samples $\{(X_m, Y_m) | m = 1, \ldots, K \times M\}$ which represent the actual parameter distribution confined by the noisy qualitative hypothesis. In Fig. 2.15(b) to Fig. 2.15(d), the noisy actual parameter samples with R_1=2 and R_2=3 are drawn with various noise strength. We calculate the averaged RMSE in these figures respectively as shown in Fig. 2.16(b). We can see that the average RMSE distance between actual inference and the expected inference degrade proportionally to the strength of the Gaussian noise σ. When σ <0.5, there are no parameter samples located in the lower-triangle of the parameter space, i.e. the actual parameter samples are consistent with the body of qualitative hypothesis "*A causes B*", and in these cases RMSE degrades smoothly with the increasing σ. However, when σ=0.5, a subset of the actual parameter samples in Fig. 2.15(d) are drawn from the lower-triangle of the parameter space which are inconsistent with the above hypothesis and the RMSE degrades dramatically thereafter.

In the fourth case, we assume the qualitative hypothesis includes not only the baseline information but also the extended information *difference*. A simple



(a) With Dif Constraint(Nois-free)

(b) With Dif Constraint(Noise, σ=0.05)

(c) With Dif Constraint(Noise, σ=0.1)

(d) With Dif Constraint(Noise, σ=0.2)

Figure 2.17: 2-Dimensional Parameter Space with Consistent Hypothesis (Cont.)

example of such hypothesis could be ”*A causes B by more than* D_1 *but less than* D_2”. Then, the parameter space is restrained by inequality $X + D_1 \leq Y \leq X + D_2$ as shown in Fig. 2.15(a) and the mean parameters can be calculated as

$$E[X] \quad = \quad P\{\int_0^{1-D_2} XdX \int_{X+D_1}^{X+D_2} dY + \int_{1-D_2}^{1-D_1} XdX \int_{X+D_1}^{1} dY\} \tag{2.118}$$

and

$$E[Y] \quad = \quad P\{\int_0^{1-D_2} dX \int_{X+D_1}^{X+D_2} YdY + \int_{1-D_2}^{1-D_1} dX \int_{X+D_1}^{1} YdY\} \tag{2.119}$$

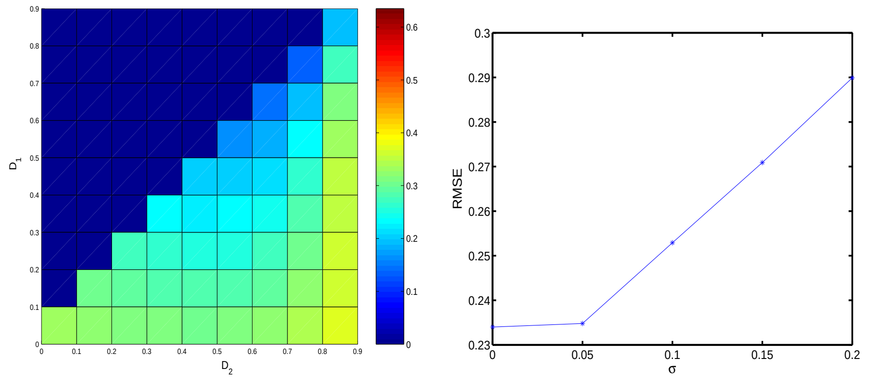
where P is the normalizing constant which satisfy

$$P\{\int_0^{1-D_2} dX \int_{X+D_1}^{X+D_2} dY + \int_{1-D_2}^{1-D_1} dX \int_{X+D_1}^{1} dY\} = 1 \tag{2.120}$$

Then, we have P=$\frac{2}{(D_2-D_1)(2-D_1-D_2)}$. By substituting P into Eq. 2.118 and 2.119, we derive $E[X] = \frac{(1-D_1)^3-(1-D_2)^3}{3(2-D_1-D_2)(D_2-D_1)}$ and $E[Y] = \frac{3-(D_2^2+D_1D_2+D_2^2)}{3(2-D_1-D_2)}$ which is shown by the red cycle in Fig. 2.15(a). The RMSE of the inference with mean parameters to the actual inference can be computed by

$$RMSE \approx \frac{1}{M}\sum_{i=1}^{M}\sqrt{(\frac{3-(D_2^2+D_1D_2+D_2^2)}{3(2-D_1-D_2)}-Y_i)^2+(\frac{(1-D_1)^3-(1-D_2)^3}{3(2-D_1-D_2)(D_2-D_1)}-X_i)^2} \tag{2.121}$$

If we further assume the hypothesis is noise-free, i.e. the actual parameters are uniform distributed in the constrained area of parameter space exclusively, the RMSE is a function of D_1 and D_2. We compute the averaged RMSE by varying D_1 and D_2 in the range [0,1] and the result is shown in Fig. 2.18(a). When $D_1 = 0$ and $D_2 = 1$, the constrained area of parameter space set by D_1 and D_2 equals to the up-triangular area in the second case as well as $E[X] = \frac{1}{3}$, $E[Y] = \frac{2}{3}$ and $RMSE = 0.30$. No parameters are sampled in the area where $D_1 \leq 0$ and/or $D_2 \leq 0$ since it violates the body of the qualitative hypothesis. Thus, the averaged RMSE in this area is set to zero. The mean RMSE of the lower-triangle parameter space in Fig. 2.18(a) equals to 0.29. If the hypothesis is noisy, i.e. the actual parameters may not be exclusively distributed in the area constrained by the inequality, instead, the actual parameters are guided by a noisy version of the inequality. We represent the noise by a Gaussian distribution over the constrained parameter space. The noisy constraint can be described as



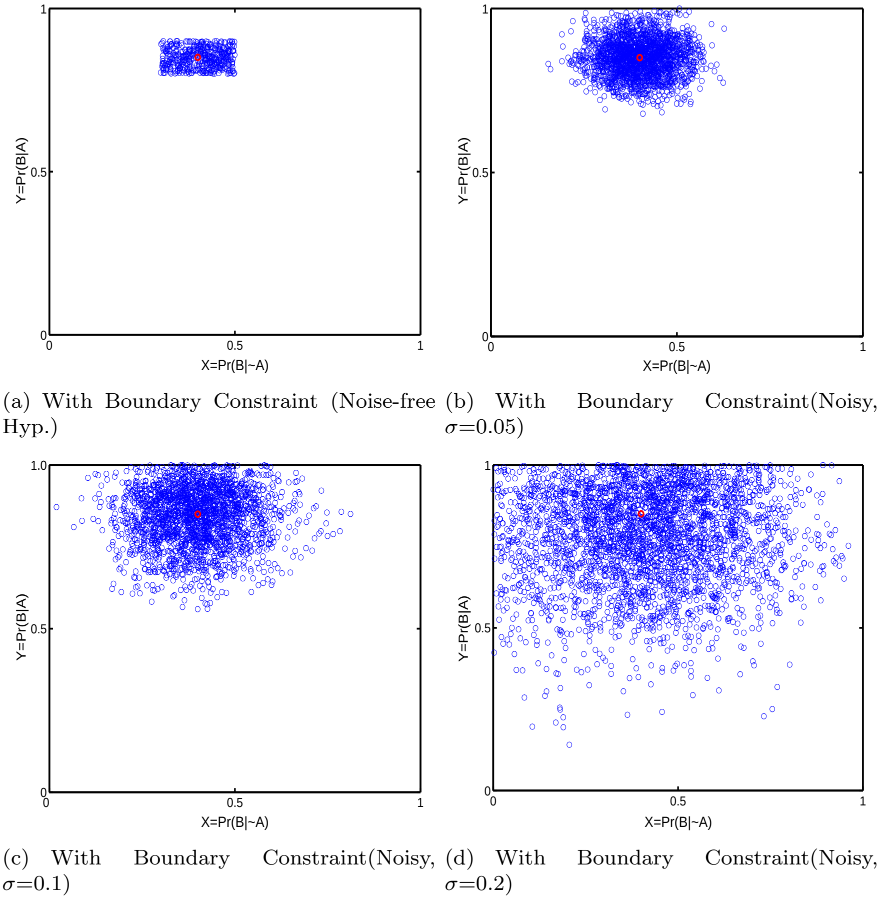
(a) Varying D_1 and D_2 (Noise-free Hyp.)

(b) Varying Noise Strength σ (D_1=0.3,D_2=0.4)

Figure 2.18: RMSE Distance Measure with Dif Constraint

$$(D_1 + \sigma N_1) + X \leq Y \leq (D_2 + \sigma N_2) + X \tag{2.122}$$

where N_1 and N_2 represent noise samples from the Gaussian distribution $N(0,1)$ with zero mean and unit variance and σ is a constant representing the noise strength. We first draw a vector of noise $\overline{N} = \{N_i|i = 1,\dots,K\}$ from $N(0,1)$. For each N_i, we construct the inequality constraint in Eq. 2.122 respectively and sample M parameters in the restrained parameter space. Thus, we obtain a set of parameter samples $\{(X_m, Y_m)|m = 1, \ldots, K \times M\}$ which represent the actual parameter distribution confined by the noisy qualitative hypothesis. In Fig. 2.17(a) to Fig. 2.17(d), the actual parameter samples with D_1=0.3 and D_2=0.4 are drawn with various noise level. We calculate the averaged RMSE in these figures respectively as shown in Fig. 2.18(b). We can see that the average RMSE distance between actual inference and the expected inference degrade proportionally to the strength of the Gaussian noise σ. When σ <0.2, there are no parameter samples located in the lower-triangle of the parameter space, i.e. the actual parameter samples are consistent with the body of qualitative hypothesis "*A causes B*", and in these cases RMSE degrades smoothly with the increasing σ. However, when σ=0.2, a subset of the actual parameter samples in Fig. 2.17(d) are drawn from the lower-triangle of the parameter space which are inconsistent with the above hypothesis and the RMSE degrades dramatically thereafter.



(a) With Boundary Constraint (Noise-free Hyp.)

(b) With Boundary Constraint(Noisy, σ=0.05)

(c) With Boundary Constraint(Noisy, σ=0.1)

(d) With Boundary Constraint(Noisy, σ=0.2)

Figure 2.19: 2-Dimensional Parameter Space with Consistent Hypothesis (Cont.)

In the last case, we assume the qualitative hypothesis includes not only the baseline information but also the boundary information. A simple example of such hypothesis could be ”*A causes B, the measure of B when A active is more than B_1 but less than B_2, the measure of B when A is not active is more than B'_1 but less than B'_2*”. Then, the parameter space is restrained by a set of inequality $X \leq Y$, $B_1 \leq Y \leq B_2$ and $B'_1 \leq X \leq B'_2$. If $B_1 \geq B'_2$, then restrained area of parameter space are shown in Fig. 2.19(a) and the mean parameters can be calculated as

$$E[X] \quad = \quad P\int_{B'_1}^{B'_2} XdX \int_{B_1}^{B_2} dY \tag{2.123}$$

and

$$E[Y] \quad = \quad P\int_{B'_1}^{B'_2} dX \int_{B_1}^{B_2} YdY \tag{2.124}$$

where P is the normalizing constant which satisfy

$$P\int_{B'_1}^{B'_2} dX \int_{B_1}^{B_2} dY = 1 \tag{2.125}$$

Then, we have P=$\frac{1}{(B_2-B_1)(B'_2-B'_1)}$. By substituting P into Eq. 2.123 and 2.124, we derive $E[X] = \frac{(B'_1+B'_2)}{2}$ and $E[Y] = \frac{B_1+B_2}{2}$ which is shown by the red cycle in Fig. 2.19. The RMSE of the inference with mean parameters to the actual inference can be computed by

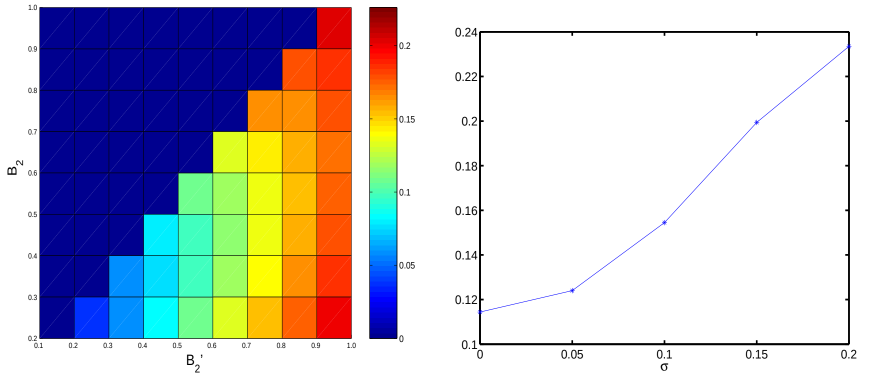
$$RMSE \approx \frac{1}{M}\sum_{i=1}^{M}\sqrt{(\frac{(B_1+B_2)}{2} - Y_i)^2 + (\frac{(B'_1+B'_2)}{2} - X_i)^2} \tag{2.126}$$

If we further assume the hypothesis is noise-free, i.e. the actual parameters are uniform distributed in the constrained area of parameter space exclusively as in Fig. 2.19(a), the RMSE is a function of B_1, B_2, B'_1 and B'_2 with the inter-relation of $B_2 \geq B_1 \geq B'_2 \geq B'_1$. We compute the averaged RMSE by varying B_1, B_2 and B'_1 and B'_2 in the range [0,1]. For demonstration purpose, we reduced the number of variables by clamping B_1 and B'_2 ($B_1 = B'_2$) and set $B'_1 = 0$. Thus, the size (S) of the constrained area is exclusively controled by B'_2 and B_2, i.e.

$$S = (B'_2 - B'_1) * (B_2 - B_1) = B'_2 * (B_2 - B'_2). \tag{2.127}$$

and the result is shown in Fig. 2.20(a). No parameters are sampled in the area confined by B_1, B_2, B'_1 and B'_2 since it violates the body of the qualitative hypothesis. Thus, the averaged RMSE in this area is set to zero. The mean RMSE of the lower-triangle parameter space in Fig. 2.20(a) equals to 0.15. If the hypothesis is noisy, i.e. the actual parameters may not be exclusively distributed in the area constrained by the inequality, instead, the actual parameters are guided by a noisy version of the inequality. We represent the noise by a Gaussian distribution over the constrained parameter space. The noisy constraint can be described as

$$\begin{aligned} (B_1 + \sigma N_1) \leq Y \leq (B_2 + \sigma N_2) \\ (B'_1 + \sigma N'_1) \leq X \leq (B'_2 + \sigma N'_2) \\ B'_2 \leq B_1 \end{aligned} \tag{2.128}$$



(a) Varying B_2 and B_2' (Noise-free Hyp.) (b) Varying Noise Strength $\sigma(B_1{=}0.3{,}B_2{=}0.6{,}B_1' = 0{,}B_2'{=}0.3)$

Figure 2.20: RMSE Distance Measure with Dif Constraint

where N_1, N_2, N_1' and N_2' represent noise samples from the Gaussian distribution $N(0,1)$ with zero mean and unit variance and σ is a constant representing the noise strength. It is obvious that the model uncertainty is affected by the noise which controls the wideness of the constrained area in the parameter space, i.e. $|B_2 - B_1|$ and $|B_2' - B_1'|$. We first draw a vector of noise $\overline{N} = \{N_i|i = 1, \ldots, K\}$ from $N(0,1)$. For each N_i, we construct the inequality constraint in Eq. 2.128 respectively and sample M parameters in the restrained parameter space. Thus, we obtain a set of parameter samples $\{(X_m, Y_m)|m = 1, \ldots, K \times M\}$ which represent the actual parameter distribution confined by the noisy qualitative hypothesis. In Fig. 2.17(a) to Fig. 2.17(d), the actual parameter samples with B_1=0.8, B_2=0.9, B_1'=0.3 and B_2'=0.4 are drawn with various noise level. We calculate the averaged RMSE in these figures respectively as shown in Fig. 2.20(b). We can see that the average RMSE distance between actual inference and the expected inference degrade proportionally to the strength of the Gaussian noise σ. When σ <0.2, there are no parameter samples located in the lower-triangle of the parameter space, i.e. the actual parameter samples are consistent with the body of qualitative hypothesis "*A causes B*", and in these cases RMSE degrades smoothly with the increasing σ. However, when σ=0.2, a subset of the actual parameter samples in Fig. 2.17(d) are drawn from the lower-triangle of the parameter space which are inconsistent with the above hypothesis and the RMSE degrades dramatically thereafter.

By comparing the averaged RMSE given the baseline parameter constraint and different extended parameter constraints in Fig. 2.16(a), 2.18(a) and 2.20(a), we conclude that the baseline feature provide a primitive and rough constraints on the model uncertainty which results in relative low generalization accuracy, but it is the most robust constraint among the others. Moreover, boundary feature gives the tightest constraints on the model uncertainty which generate high generalization accuracy, however, it is very sensitive to the noisy information.

2.5 Empirical Study

In this section, we apply the proposed method to the real-world applications where qualitative statements of the molecular interactions in breast cancer bone metastasis and breast cancer cell proliferation are extracted from biological publications. We use our knowledge model to translate these statements into a set of constrained class of Dynamic Bayesian networks to predict incidence of bone metastasis and proliferation rate respectively.

2.5.1 TGFβ-mediated Breast Cancer Bone Metastasis Network

Breast Cancer Bone Metastasis (BCBM)

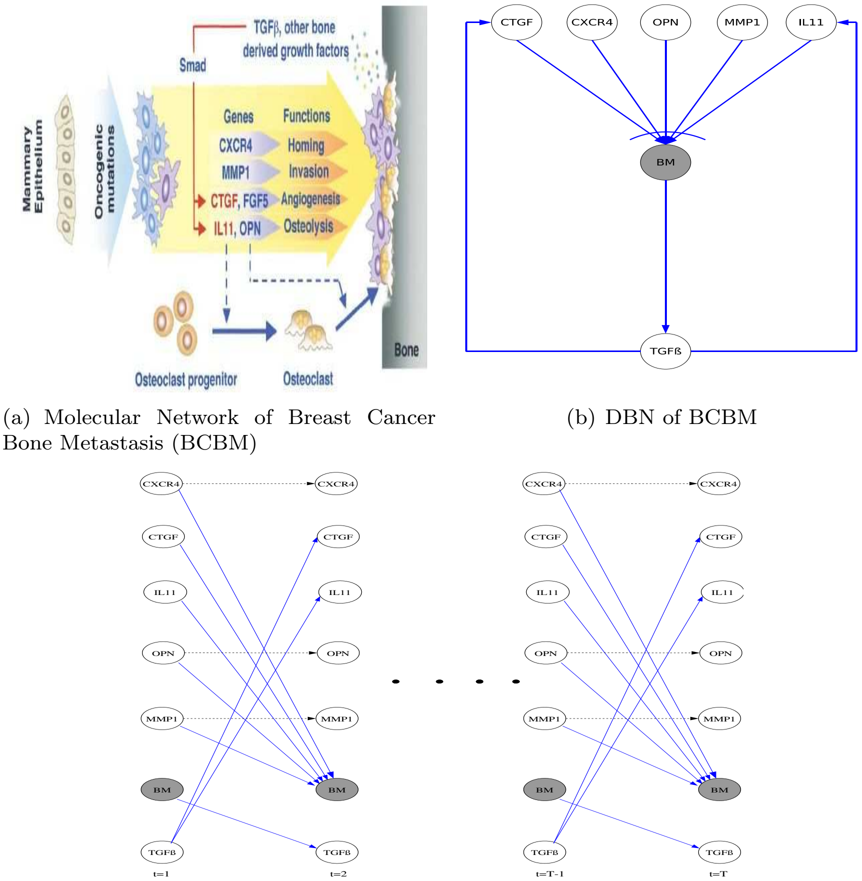
Kang and his colleagues, [60], have identified several key genes responsible for promoting breast carcinoma metastasis to bone. Among this functionally overexpressed gene set, a subset was further investigated by transfection studies, namely, matrix metalloproteinase 1 *(MMP1)*, interkeulin 11 *(IL11)*, a chemokine receptor for SDF-1 *(CXCR4)* and connective tissue-derived growth factor *(CTGF)*. In addition, Osteopontin *(OPN)* was considered in the study due to its consistent overexpression in highly metastatic cell lines derived from MDA-MB-231 cancerous cells. The genes were found to promote bone metastasis in a cooperative manner. Combined transfection of various gene combinations out of these genes into the parental MDA-MB-231 human breast carcinoma cell line resulted in enhanced but different metastatic patterns comparing to the parental MDA-MB-231 populations. Besides these genes, transforming growth factor-β ($TGF\beta$) which is abundantly stored in bone matrix [75, 92], is released during osteolysis which supports a vital cycle of breast cancer bone metastasis [75]. The breast cancer bone metastasis(BCBM) network is shown in figure 2.21(a).

In addition to revealing these molecular relationships, the bone metastatic activity of the different transfected cell lines could be quantified by injecting them into a mouse model and evaluating the evoked incidence of bone metastasis over time. The Kaplan Meier curves for bone metastatic activity are plotted in figure 2.24. In this example application, we aim at predicting the quantitative values of bone-metastatic activity from the set of relationships reflected in figure 2.21(a).

Qualitative Knowledge Model of the BCBM Network

The original biological observation statements from Kang are summarized in the appendix B. We could mainly draw the following conclusions from these qualitative statements:

1. *Osteolytic bone metastasis in breast cancer increases the level of TGFβ since abundant TGFβ is released from bone matrix during the osteolytic bone metastasis.*
2. *TGFβ activates CTGF and IL11 which forms a positive feedback loop in bone metastasis formation.*
3. *CXCR4, CTGF, IL11, OPN and MMP1 cooperatively promote osteolytic bone metastasis which forms synergic effects on bone metastasis in ad-*



(a) Molecular Network of Breast Cancer Bone Metastasis (BCBM)

(b) DBN of BCBM

(c) 2TBN of BCBM

Figure 2.21: Structure of BCBM Network

dition to a positive effect of each individual gene in causing osteolytic metastasis.

These statements can be translated into probability inequalities by applying the knowledge model introduced in the last section: They are of the type *single positive influence* and *plain synergic joint influence*. The first two statements can be translated into a set of probability inequality constraints by definition 3.2

$$\begin{aligned} P(TGF\beta|BM) &\geq P(TGF\beta|\overline{BM}) \\ P(IL11|TGF\beta) &\geq P(IL11|\overline{TGF\beta}) \\ P(CTGF|TGF\beta) &\geq P(CTGF|\overline{TGF\beta}) \end{aligned} \tag{2.129}$$

We use $\Pi = \{CXCR4, CTGF, IL11, OPN, MMP1\}$ to denote the parent nodes of bone metastasis. If we further assume that these five nodes are pair-wise symmetric to the child node, then we can translate the last qualitative statement

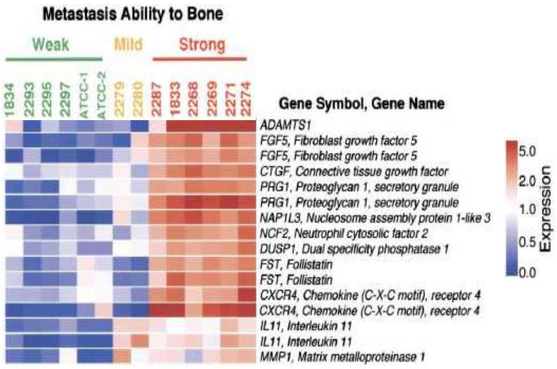


Figure 2.22: Gene Expression Profiles of BCBM Cell Lines

into probability inequality constraints by definition 3.5

$$P(BM|\pi_5) \geq P(BM|\pi_4) \geq P(BM|\pi_3) \geq P(BM|\pi_2) \geq P(BM|\pi_1) \geq P(BM|\pi_0) \tag{2.130}$$

where π_n denotes the set of all joint parent states with exactly n parents being active and $P(BM|\pi_n)$ represents the subset of CPT entries with n active parents. The set of inequalities in Eq. 2.129 and Eq. 2.130, denoted by Δ, are derived from the set of statements about relationships which is a part of the total qualitative knowledge available in [60]. In addition to these qualitative statements we also make use of a set of qualitative statements on the gene expression levels in Π, which provide their prior probability. Qualitative statements about the gene expression levels of *(CTGF)*, *(CXCR4)*, *(MMP1)* and *(IL11)* can be extracted from the gene expression profiles in figure 2.22 [60]. From this, we could conclude the following qualitative statements on the initial expression levels of gene *CXCR4, OPN* and *MMP1* as

$$\begin{gathered} CXCR4 = 1.0 \\ 0.5 < CTGF < 1.0 \\ MMP1 < IL11 < CTGF \end{gathered} \tag{2.131}$$

By using the fact that expression levels range between 0 and 5 in this experiment, we can obtain statements about the activation probabilities of these genes as

$$\begin{gathered} P(CXCR4) = 0.2; \\ 0.1 < P(CTGF) < 0.2; \\ P(MMP1) < P(IL11) < P(CTGF). \end{gathered} \tag{2.132}$$

These probability inequalities are referred to as Φ. The complete qualitative knowledge Ω translated from the qualitative statements on causal relations between molecules and phenotypic entities in [60] is given by $\Omega = \{\Delta, \Phi\}$.

Dynamic Bayesian Network Structure for BCBM

A cyclic graphical representation of the BCBM molecular network is given in figure 2.21(a) [60]. It reflects the graph structure that is consistent with the statements analyzed in the previous section. The curve above *BM* indicates a plain synergic joint influence between its parent nodes.

Since the BCMB biological network is cyclic, we next created the corresponding DBN model, which is shown in figure 2.21(b). The black dotted links between two nodes indicate that their states are time invariant, i.e. probabilities of these genes being overexpressed are time invariant. The resulting DBN is uniquely defined by its structure and parameters.

DBN Parameter Space

For the fixed DBN structure used, the model uncertainty is equivalent to model parameter uncertainty. It is given by the parameters of the four CPTs listed in table 2.5 and table 2.6, which are encoded in the network graph structure. The total vector of parameters for the conditional probabilities is given by $\theta = (\beta, \gamma, \lambda, \zeta, \xi)$. The parameter sets $(\beta, \gamma, \lambda, \zeta)$ are derived from the statements Δ, i.e., they are constrained by the probability inequalities Eq. (2.129) and (2.130). ξ refer to constraints on the expression values of the genes. They are derived from Φ and are given by Eq. 2.132.

According to Δ defined in Eq. 2.129 and Eq. 2.130, the model uncertainty on CPTs can be expressed as

$$\beta_0 \leq \beta_1 \qquad \gamma_0 \leq \gamma_1 \qquad \lambda_0 \leq \lambda_1 \tag{2.133}$$

and

$$\alpha_0 \leq \alpha_1 \leq \alpha_2 \leq \alpha_3 \leq \alpha_4 \leq \alpha_5 \tag{2.134}$$

where

$$\begin{aligned}
\alpha_0 &= \{\zeta_0\} \\
\alpha_1 &= \{\zeta_1, \zeta_2, \zeta_4, \zeta_8, \zeta_{16}\} \\
\alpha_2 &= \{\zeta_3, \zeta_5, \zeta_6, \zeta_9, \zeta_{10}, \zeta_{12}, \zeta_{17}, \zeta_{18}, \zeta_{20}, \zeta_{24}\} \\
\alpha_3 &= \{\zeta_7, \zeta_{11}, \zeta_{13}, \zeta_{14}, \zeta_{19}, \zeta_{21}, \zeta_{22}, \zeta_{25}, \zeta_{26}, \zeta_{28}\} \\
\alpha_4 &= \{\zeta_{15}, \zeta_{23}, \zeta_{27}, \zeta_{29}, \zeta_{30}\} \\
\alpha_5 &= \{\zeta_{31}\}
\end{aligned} \tag{2.135}$$

Inference of Bone Metastatic Activity

Given qualitative knowledge $\Omega = (\Delta, \Phi)$ in Eq. 2.129, 2.130 and 2.132, we now apply Bayesian inference to predict the incidence of bone metastasis on the basis of the qualitative knowledge by model averaging. We try to infer *in silico*, what is the likelihood of bone metastasis (node BM) for different transfectant cell lines, in light of the body of qualitative knowledge used to construct the Bayesian model class. The transfection of a gene X corresponds to a measurement or piece of evidence E in the model. Transfection of X means that it is constantly overexpressed, hence we obtain as evidence E: $P(X = 1) = 1$ for all time steps of the dynamic Bayes net simulation.

Based on the evidence of having various transfectants E, we wish to infer the resulting metastatic activity, $P(BM|E, \Omega)$ using Eq. 2.97. We do so once by making use of the statements about relationships only, Δ, and once by taking

(a) CPT of TGFβ given Bone Metastasis

BM	β_j=P(TGFβ |BM)
0	β_0
1	β_1

(b) CPT of CTGF given TGFβ

TGFβ	γ_j=P(CTGF|TGFβ)
0	γ_0
1	γ_1

(c) CPT of IL11 given TGFβ

TGFβ	λ_j=P(IL11|TGFβ)
0	λ_0
1	λ_1

Table 2.5: CPTs for Single Positive Influence links

CXCR4	CTGF	IL11	OPN	MMP1	ζ_j=P(BM|CXCR4,CTGF,IL11,OPN,MMP1)
0	0	0	0	0	ζ_0
0	0	0	0	1	ζ_1
0	0	0	1	0	ζ_2
0	0	0	1	1	ζ_3
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	1	1	ζ_{31}

Table 2.6: CPT of Bone Metastasis given related genes

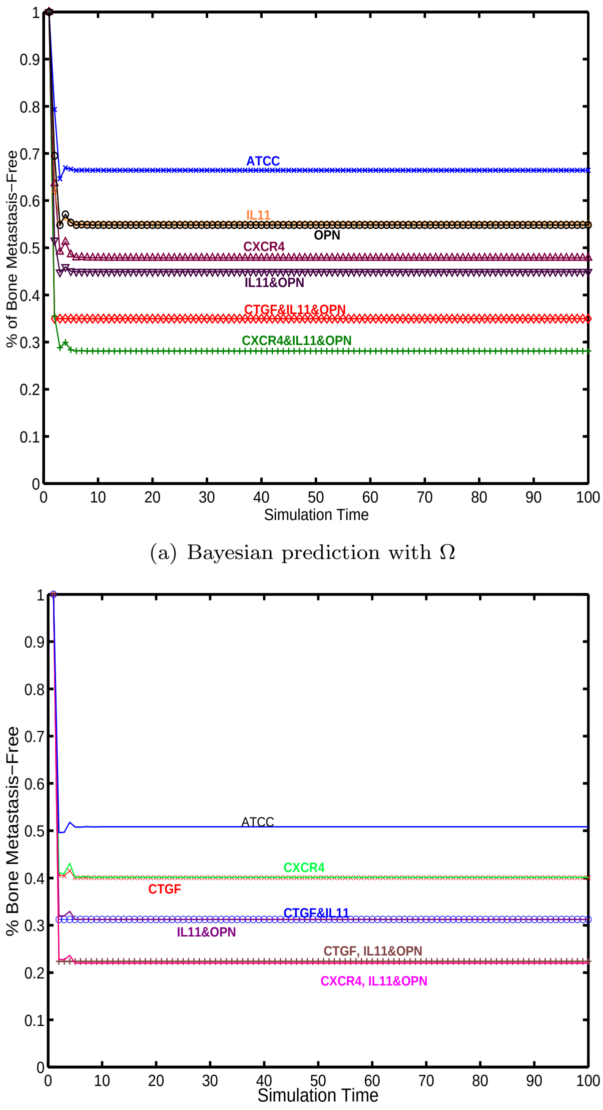
into account the full statement set Ω. In the first case, we marginalize over the full range of all gene expression levels not clamped by an *in silico* transfection experiment. In the second case we marginalize over gene expression values only in the range specified in Eq. 2.132.

Since the model parameter space $\theta = (\beta, \gamma, \lambda, \zeta, \xi)$ is rather high-dimensional, $\beta, \gamma, \lambda \in [0,1]^2$, $\zeta \in [0,1]^{32}$ and $\xi \in [0,1]^5$, we use Monte Carlo sampling to approaximate the integration. We use $N = 500,000$ model samples in the Monte Carlo simulation. For each simulation, the selected Bayesian model is iterated over time until the posterior probability of bone metastasis converges.

In-silico Prediction on Bone Metastasis

Seven *in silico* transfection experiments are performed in which certain genes out of probe gene set (CXCR4,CTGF,IL11,OPN,MMP1) in MBA-MD-231 human breast cancer cell are manually transfected to be overexpressed. We show the convergence of model uncertainty on parameter space by Monte Carlo simulation based on qualitative knowledge.

500,000 Monte Carlo samples are simulated and for each sample 100 DBN iterations were performed. The total simulation time was approximate 10 minutes on a PentiumIV PC. Figure 2.26 shows how the empirical mean of metastastic activity of ATCC cell as calculated with 100 DBN iterations convergence with the number of Monte Carlo samples. Figure 2.23 shows how the mean estimate

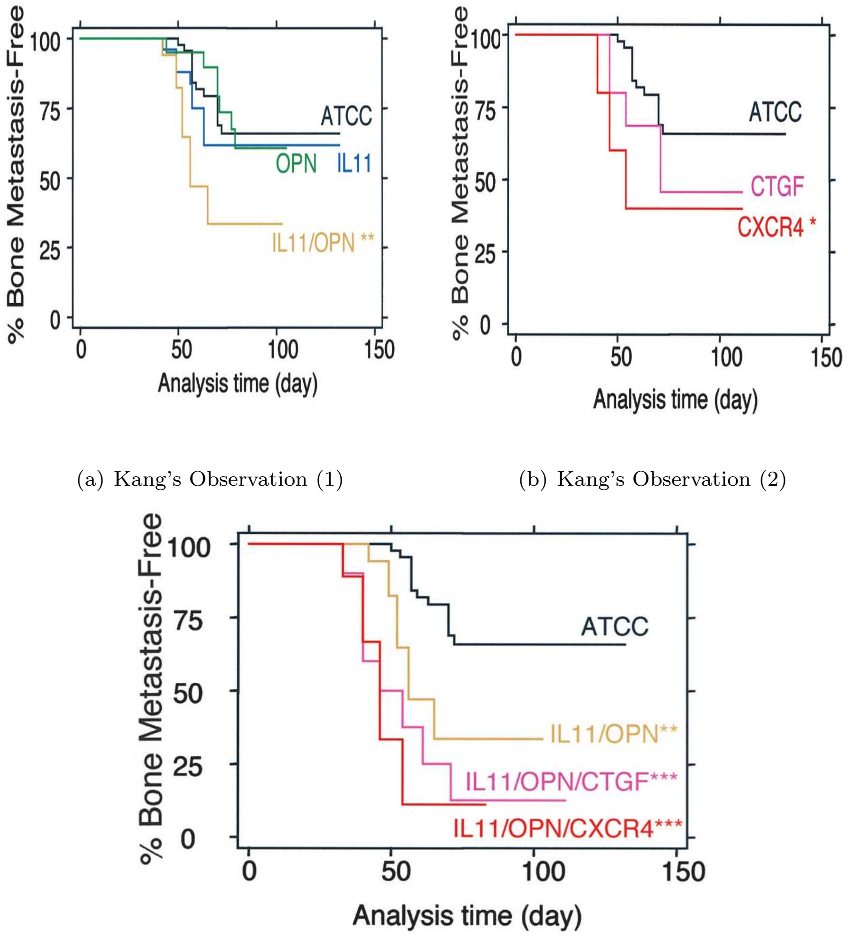


(a) Bayesian prediction with Ω

(b) Bayesian prediction with only Δ

Figure 2.23: Bone Metastasis Prediction with Qualitative Knowledge

of bone metastatic activity evolves over time in the DBN simulation. The time course of each simulation is labeled with the overexpressed transfectant genes' names in the figure. The simulation with the non-transfectant parental ATCC cell is labeled "ATCC". When comparing figure 2.23 and figure 2.24, we can see that our simulations produce reasonable quantitative predictions on bone metastasic activity under each biological experimental setting. Table 2.7 summarizes the simulated and measured values for the bone metastatic activities of different cell lines. The results are closer to the true values, when additional information about ranges of gene expression levels is taken into account (Ω used, fig. 2.23(a)) than with statements about relationships alone (only Δ used, fig.



(a) Kang's Observation (1)

(b) Kang's Observation (2)

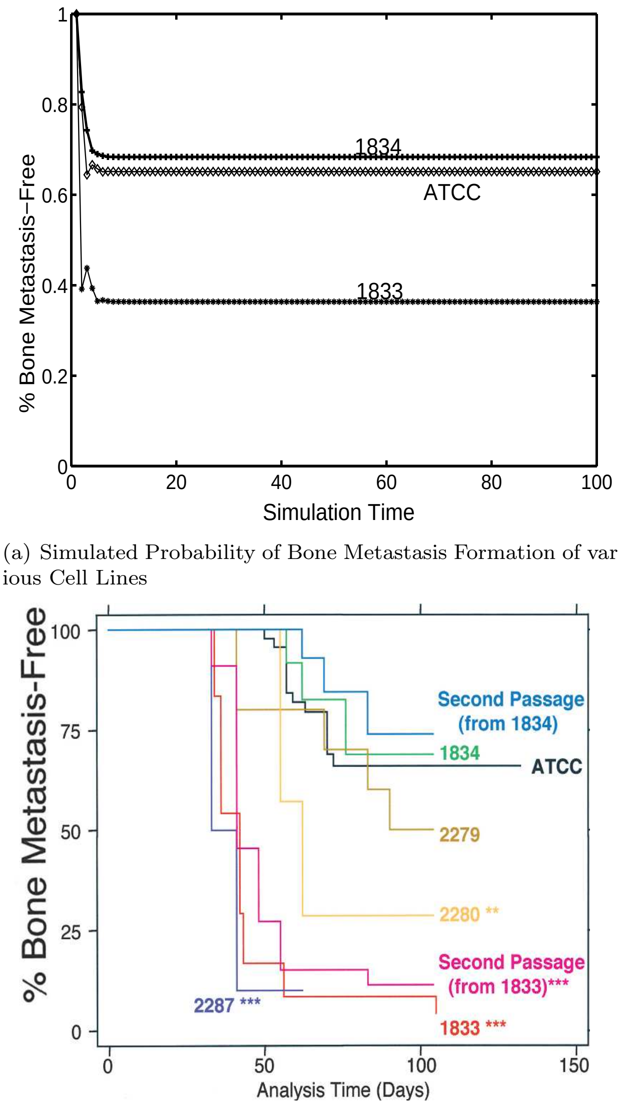
(c) Kang's Observation (3)

Figure 2.24: Bone Metastasis Observations in Kang's Experiment

2.23(b)) Hence, the beauty of this approach is that it seems to allow realistic quantitative predictions by qualitative prior knowledge alone.

2.5.2 TGFβ-mediated Mammary Epithelial Cell Cytostasis Program and Breast Cancer

We apply our method to investigate the complex function of transforming growth factorβ (TGFβ) in controlling the mammary epithelial cell proliferation [11]. However, tumor cells which are relieved from TGFβ growth constraints might then overproduce this cytokine to create a local immunosuppressive environment that fosters tumor growth and exacerbates the invasiveness and metastasis [12, 102]. In the first experiment, we model the core representation of the molecular interaction network of TGFβ cytostatic program with dynamic Bayesian model and perform in-silico quantitative prediction on the probability of cell growth given TGFβ. We showed that TGFβ initially acts as growth



(a) Simulated Probability of Bone Metastasis Formation of various Cell Lines

(b) Biology Observation of Bone Metastasis Formation of various Cell Lines

Figure 2.25: Bone Metastasis Probability of Different Cell Lines

suppressor in normal cells based on its expression level in the cell. In the second experiment, we predict the probability of cell growth in breast cancer with specific loss of TGFβ cytostatic response in the pathway.

TGFβ-mediated Cytostatic Program Modeling

TGFβ can activate cytostatic gene responses in G1 phase and impede the completion of the ongoing cell cycle. TGFβ responses in human epithelial cell lines from skin, lung and mammary gland originals have revealed a shared cyto-

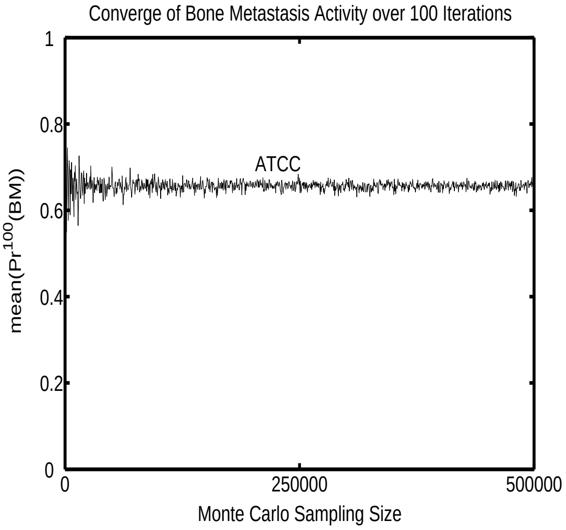


Figure 2.26: Convergence of Bone Metastasis Prediction by Monte Carlo Simulation

Experiments	Prediction with ξ	Prediction w/o ξ	True Values
ATCC	35%	50%	30%
IL11	44%	60%	38%
CTGF	44%	60%	54%
OPN	45%	60%	39%
CXCR4	52%	60%	59%
IL11&OPN	55%	70%	66%
IL11&OPN&CTGF	65%	78%	88%
IL11&OPN&CXCR4	72%	78%	89%

Table 2.7: Numerical Prediction of Bone Metastasis in different simulations

static program that minimally includes activation of the cyclin-dependent kinase (CDK) inhibitors, p15, p21 and p27 and repression of the growth-promoting transcription factor c-MYC. Several feedback loops serve to integrate this program and providing tight control and robustness signals. A set of qualitative hypotheses with regarding to this network can be extracted from a group of publications [17,22,46,85,90,93,97,98,102] as $\overline{\Omega} = \{\Omega_i | i = 1 \dots 9\}$ where

1. Ω_1: *CDK2, CDK4 and CDK6 drive progression through the G1 phase of the cell cycle. In G1, CDK4/6 activation requires association with D-type cyclins whereas cyclin E binding activates CDK2. [102]*;

2. Ω_2: *The cyclin-dependent kinase inhibitor p15, is induced by treatment with TGFβ, suggesting p15 may act as an effector of TGFβ-mediated cell cycle arrest. [46]*;

3. Ω_3: *TGFβ elevates expression of CDK4/6-specific inhibitor p15 and induces the release of p27 from CDK4 and CDK6 complex and this release*

concides with the increased binding of p27 from CDK4 to CDK2 in vivo, suggesting that the the release of CDK4-bound p27 in TGFβ treated cells in caused by the surge in p15 levels. [90];

4. Ω_4: *TGFβ can induce the cyclin-dependent kinase inhibitor p21 through a p53-independent pathway. [22]*;
5. Ω_5: *TGFβ can induce the cyclin-dependent kinase inhibitor p27 which associates with cyclinE-CDK2 complex in vivo and prevents their activation. [85, 93]*;
6. Ω_6: *A complex containing Smad3, E2F4/5, DP1 and p107, in response to TGFβ, associates with Smad4 and recognize a composite Smad-E2F site on c-MYC for repression. [17]*;
7. Ω_7: *TGFβ signalling prevents recruitment of c-MYC to the p15 transcription initiator by Miz-1. Two separate TGFβ-dependent inputs keep tight control over p15 activation: Smad-mediated transcription and relief of repression by c-MYC. [98]*;
8. Ω_8: *Transcript factor c-MYC is directly recruited to the p21 promoter by the DNA-binding protein Miz-1. This interaction blocks p21 induction by p53 and other activators. [97]*;
9. Ω_9: *TGFβ activates Ras and ErbB2 which induces formation of proliferative structures in noninvasive early stage mammary epithelial lesions. [99]*

Based on $\overline{\Omega}$, the dynamic Bayesian network of TGFβ-mediated cytostatic program can be shown in Figure 2.27(a) and can be unrolled over the time into a serie of 2TBNs as shown in Figure 2.27(b). In this experiment, we describe the compact representation of the cytostatic network with the core molecules, thus, smad proteins and other co-expressor, e.g. E2F4/5, p53, p107, p300, ID1, ID2 and Miz-1 [17, 22, 46, 85, 90, 93, 97, 98, 102] are excluded for simplicity. However, for later study, these molecules can be consistently integrated [8, 10].

The parameters are described by the conditional probability tables (CPT) in Table 2.8 and 2.9. According to the qualitative knowledge model [13, 15], the parameter α can be modeled by *Single Negative Influence*, i.e. $\alpha_0 \geq \alpha_1$, and σ can be modeled by by *Single Positive Influence*, $\sigma_1 \geq \sigma_0$ and parameters (β,γ) can be described by *Mixed Joint Influence*, i.e.

$$\begin{aligned} \beta_0 \geq \beta_1, \beta_2 \geq \beta_3, \beta_2 \geq \beta_0 \\ \beta_3 \geq \beta_1, \gamma_0 \geq \gamma_1, \gamma_2 \geq \gamma_3 \\ \gamma_2 \geq \gamma_0, \gamma_3 \geq \gamma_1 \end{aligned} \tag{2.136}$$

The parameters (λ,ρ) can be modeled by *Plain Synergy with Positive Individual Influence*, i.e.

$$\rho_7 \geq \left\{ \begin{array}{c} \rho_3 \\ \rho_5 \\ \rho_6 \end{array} \right\} \geq \left\{ \begin{array}{c} \rho_1 \\ \rho_2 \\ \rho_4 \end{array} \right\} \geq \rho_0$$

$$\lambda_7 \geq \left\{ \begin{array}{c} \lambda_3 \\ \lambda_5 \\ \lambda_6 \end{array} \right\} \geq \left\{ \begin{array}{c} \lambda_1 \\ \lambda_2 \\ \lambda_4 \end{array} \right\} \geq \lambda_0$$

(a) α

TGFβ	Pr(C-MYC|TGFβ)
0	α_0
1	α_1

(b) σ

TGFβ	Pr(RAS|TGFβ)
0	σ_0
1	σ_1

(c) γ

TGFβ	c-MYC	Pr(p21|TGFβ,c-MYC)
0	0	γ_0
0	1	γ_1
1	0	γ_2
1	1	γ_3

(d) β

TGFβ	c-MYC	Pr(p15|TGFβ,c-MYC)
0	0	β_0
0	1	β_1
1	0	β_2
1	1	β_3

(e) η

cyclinE	CDK2	p21	p27	Pr(cyclinE-CDK2|cyclinE,CDK2,p21,p27)
0	0	0	0	η_0
0	0	0	1	η_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	1	η_{15}

(f) λ

TGFβ	p15	p27-cyclinD-CDK46	Pr(p27|TGFβ,p15,p27-cyclinD-CDK46)
0	0	0	λ_0
0	0	0	λ_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	λ_7

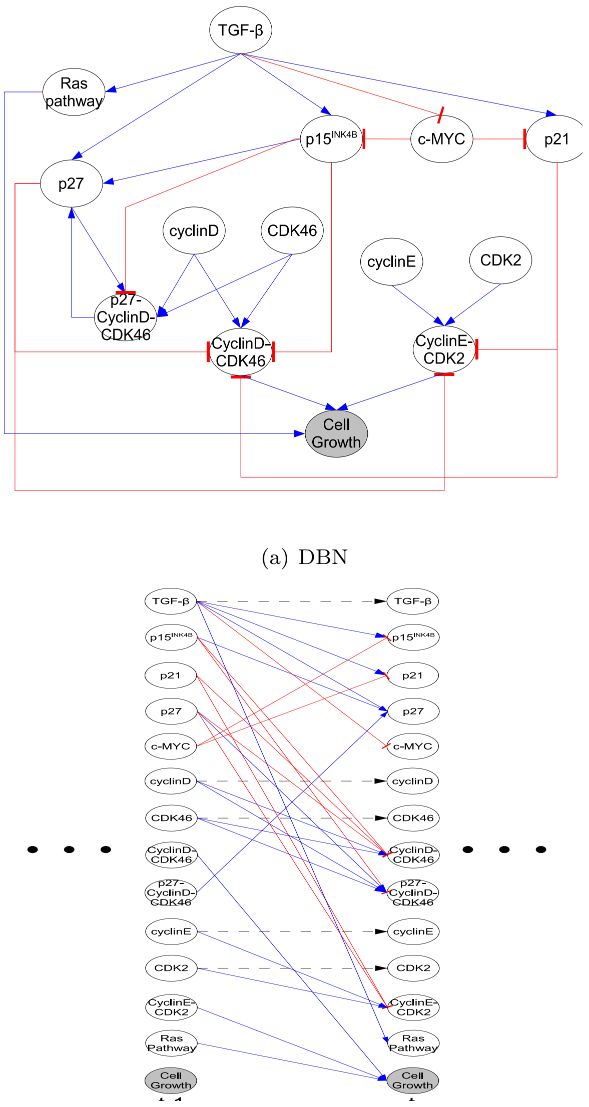
(g) ϕ

cyclinD	CDK46	p27	p15	Pr(p27-cyclinD-CDK46|cyclinD,CDK46,p27,p15)
0	0	0	0	ϕ_0
0	0	0	1	ϕ_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	1	ϕ_{15}

(h) θ

cyclinD	CDK46	p15	p21	p27	Pr(cyclinD-CDK46|cyclinD,CDK46,p15,p21,p27)
0	0	0	0	0	θ_0
0	0	0	1	0	θ_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	1	1	θ_{31}

Table 2.8: Parameters of TGFβ-mediated Cytostatic Program



(a) DBN

(b) 2TBN

Figure 2.27: Structure of TGFβ-mediated Cytostatic Program

The parameters (θ,ϕ,η) can be defined by a set of constraints hierarchically. Firstly, the parameters can be modeled by *Mixed Joint Influence* since there are multiple input signals to activate and to repress the molecules *CyclinD-CDK46*,*CyclinE-CDK2* and *p27-CyclinD-CDK46* from their parents. Therefore, the parameters can be classified according to the number of repressors being overexpressed. Secondly, the parameters in each class can be further defined by *Plain Synergy with Positive Individual Influence* based on the number of activators being overexpressed. For example, the parameter η can be firstly classified into four classes of parameters based on the configuration of p21 and p27, i.e. G_0=$\{\eta_0, \eta_4, \eta_8, \eta_{12}\}$, $G_{1,1}$=$\{\eta_1, \eta_5, \eta_9, \eta_{13}\}$, $G_{1,2}$=$\{\eta_2, \eta_6, \eta_{10}, \eta_{14}\}$ and

cyclinE-CDK2	cyclinD-CDK46	RAS	Pr(CellGrowth|cyclinE-CDK2,cyclinD-CDK46,RAS)
0	0	0	ρ_0
0	0	1	ρ_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	1	1	ρ_7

Table 2.9: Parameters of TGFβ-mediated Cytostatic Program(Cont.)

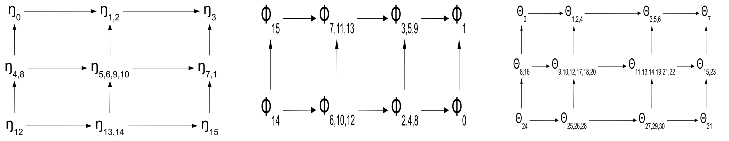
G_2=$\{\eta_3, \eta_7, \eta_{11}, \eta_{15}\}$. If we assume the inhibitive effects of p21 and p27 on the molecular *CyclinD-CDK46* are symmetric (In general, it is possible to model the unsymmetrical effects), we could merge the parameters of $G_{1,1}$ and $G_{1,2}$ into one class, G_1. With the same configuration of the parents, the parameters across the classes can be constrained as

$$\begin{array}{llll} \eta_0 \geq \eta_{1,2} & \eta_4 \geq \eta_{5,6} & \eta_8 \geq \eta_{9,10} & \eta_{12} \geq \eta_{13,14} \\ \eta_{1,2} \geq \eta_3 & \eta_{5,6} \geq \eta_4 & \eta_{9,10} \geq \eta_{11} & \eta_{13,14} \geq \eta_{15} \end{array} \tag{2.137}$$

Secondly, within each class, the parameters can be further classified by the number of activators being overexpressed as

$$\begin{gathered} \eta_{4,8} \geq \eta_0, \eta_{5,6,9,10} \geq \eta_{1,2}, \eta_{7,11} \geq \eta_3 \\ \eta_{12} \geq \eta_{4,8}, \eta_{13,14} \geq \eta_{5,6,9,10}, \eta_{15} \geq \eta_{7,11} \end{gathered} \tag{2.138}$$

The constraints in Eq. 2.137 and Eq. 2.138 can be best illurstrated by a tree as shown in Figure 2.28(a). Similarly, ϕ and θ can be modeled as Figure 2.28(b) and 2.28(c) where the edge indicates the *greater than* operator.



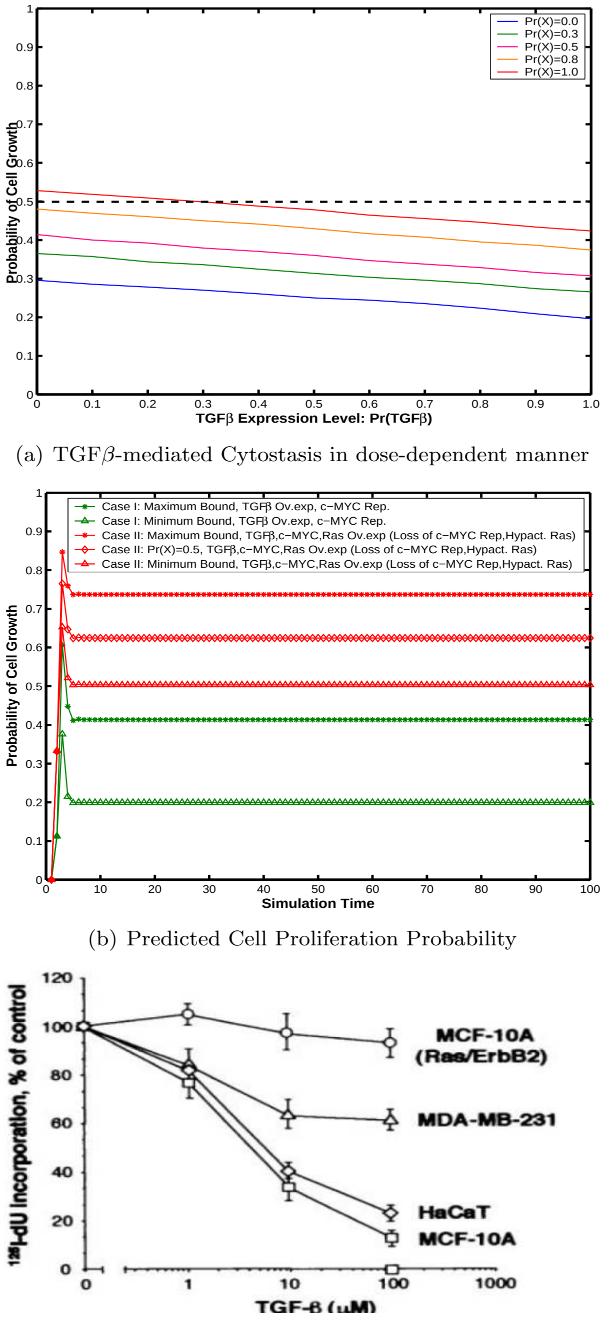
(a) Tree of η (b) Tree of ϕ (c) Tree of θ

Figure 2.28: Parameter Tree of TGFβ-mediated Cytostatic Program

The inference on cell growth is computed for each possible model with parameters Π=$\{\alpha, \sigma, \gamma, \rho, \phi, \beta, \eta, \theta, \lambda\}$ by Eq. 2.97. However, since the parameter space is rather high dimension, we can use Monte Carlo methods to approximate the integration. For each parameter in Π,we simulate K=500,000 samples constrained by Eq. 2.136 to 2.137 and Figure 2.28(a) to 2.28(c), and these CPT samples together with the structure s in Figure 2.27(a) define a consistent model class, $\widetilde{M} = \{m_k(s, \alpha_k, \gamma_k, \rho_k, \phi_k, \beta_k, \eta_k, \theta_k, \lambda_k) | k = 1, \ldots, K\}$.

TGFβ-mediated Cell Growth Control

It has been long observed that TGFβ acts as tumor suppressor in normal cells and low-grade tumorigenic cells with inactivated Ras pathway. In this experiment, we predict the probability of cytostasis at steady state, i.e. at the convergence of DBN belief propagation, by varying the expression level of TGFβ



(a) TGFβ-mediated Cytostasis in dose-dependent manner

(b) Predicted Cell Proliferation Probability

(c) Experiment Observations

Figure 2.29: Parameter Tree of TGFβ-mediated Cytostatic Program

as shown in Figure 2.29(a). We showed that i)The cell growth by TGFβ is dose-dependent process, i.e. increasing TGFβ expression level monotonically reduces the probability of cell growth and vice-versa; ii) The cell growth control by TGFβ are cell context-based process in that different cell lines with distinct expression level of X demonstrate various cell growth capability; iii)Based on a specific cell context, reducing the TGFβ expression level which is equivalent to the loss of the TGFβ responses with mutated effectors along the pathway may cause the turnover of original normal cells with dominant cell growth(CG) control signals, i.e. P($\overline{CG}$)≥P(CG), to low-grade tumorigenic cells with dominant cell growth signals, i.e. P($\overline{CG}$)≤P(CG).

Loss of TGFβ Responses in Breast Cancer

Loss of growth inhibitory responses to the TGFβ in cancer cells may results from mutational inactivation of TGFβ receptors or the Smad transcript factors. In breast cancer, however, it has been identified that the loss of TGFβ growth inhibition often occurs without a loss of these signaling factors. In stead, the repression of a key cell growth promotion component c-MYC is selectively lost [16]. In this experiment, we predict the cell growth probability of various breast cancer cell lines with the loss of c-MYC repression by TGFβ and hyperactive oncogenic Ras pathway. We assume the rise of tumorigensis is caused by unrestricted cell growth which is true for most epithelial cancers [107]. Three breast immortalized and tumorigenic cell lines, MCF-10A, MCF-10A(Ras/ErbB2) and MDA-MB-231 with distinct response of c-MYC repression to TGFβ and Ras oncogenic transfection [16] are simulated. The loss of c-MYC repression in MCF-10A(Ras/ErbB2) and MDA-MB-231 cells can be modeled by clamping the prior probability of c-MYC to overexpress, i.e. P(c-MYC)=1. The hyperactivity of the Ras pathway in MDA-MB-231 cells and MCF-10A(Ras/ErbB2) cells can be represented by clamping the prior probability of Ras to overexpress, i.e. P(RAS)=1. Since the level of X are unknown, we represent this uncertainty by varying the prior probability of cell cycle control molecules X between 0 and 1, i.e. P(X)∈[0,1], to reflect the broad distinction of the cell context. We simulated the minimum/maximum bounds on the probability of cell proliferation under two circumstances: i)The c-MYC repression by TGFβ is preserved and Ras/ErbB2 are not transfected to overexpress which is the case of normal human mammary epithelial cells MCF-10A; ii)The c-MYC repression by TGFβ is lost and Ras/ErbB2 are transfected to overexpress which is the case of MCF-10A(Ras/ErbB2) and MDA-MB-231 cells. The simulation results are shown in Figure 2.29(b). In case i), depending on P(X), the probability of cell proliferation ranges in [19%-41%] whereas in case ii), it results in [50%-74%]. It has been observed that the overall expression level of cyclinD and cyclinE in MCF-10A are undetectable [61,116] which indicates that P(X) of MCF-10A cell can be sufficiently close to zero, thus, its simulated proliferation probability shall be close to the minimum bound of case i). In contrast, in MDA-MB-231 cells, expression level of cyclinE is only moderately overexpressed [61] and cyclinD relatively overexpress [6,116] which implies that P(X) is somewhere between the maximum express level(1) and the minimum expression level(0), thus the simulated proliferation probability of MDA-MB-231 cells is between 50% and 74%; For demonstration, we simulated with P(X)=0.5 which results in 63% cell proliferation. In MCF-10A(Ras/ErbB2), the overexpression of cyclinD is strongly induced by ErbB2 and the Ras significantly stimulates the cyclinE/CDK2 activity [6, 51] which presumptively indicates that P(X) can be sufficiently large comparing to 1, thus, its simulated proliferation probability shall be close to the maximum bound of case ii). For comparison, the observations on the cell proliferation probability formed by these cell lines are shown in Figure 2.29(c). [16]

2.6 Summary

In this chapter, we formally proposed the knowledge-driven probabilistic networks modeling framework. We formulated the qualitative knowledge model and use it to capture the structural dependence and influence between the variables under concern and translate the cause-effect relationship described by the qualitative statements into a set of inequality constraints over the model parameter space. The structural and parameter constraints eventually forms the model piror distribution. In this way, we can build a class of (Dynamic) Bayesian networks which are consistent with the body of the qualitative prior knowledge.

We employ full Bayesian approach to calculate the average quantity of interest, e.g. inference, over the class of Bayesian networks. The integral over model structure space and parameter space can be computed by Monte Carlo integration techinique with Accept-Reject algorithm. We sampled the model uniformly from the model space and select those which are consistent with the structural and parameter constraints. One computational difficulty here is that in order to get a "good" approximation to $E[P(X|E,\Omega)]$, we ususally sample a large amount of valid models and compute the inference in each of them which causes extremy computation complexity. Hereby, we have proposed an simple and efficient solution to this problem, i.e. we approximate mean inference over the model space $E[P(X|E,\Omega)]$ by performing inference in the mean model which is the averaged model over model space, $P(X|E,\overline{m},\Omega)$. This approximation scheme is exact when inference function $P(X|E)$ is a linear function with respect to the model parameters which is often true in the case of forward prediction in Dynamic Bayesian networks. However, this method will introduce biased results in case of static Bayesian networks. We have analyzed the bias as a function of the message-propagation steps n and of the constraints over the parameter space. We concluded that for small n and relative restrictive constraints (relative small parameter sub-space), our approximation scheme is usually reasonable.

Also, we have studied the robustness of our approach. Noisy information exists largely in a qualitative statement. For example, a domain expert may feel comfortable to express the probability of getting lung cancer for a smoker *likely* ranges from 10% to 15%. In this statement, parameter uncertainty on the conditional probability of lung caner given smoking is given by the boundary information: [10%,15%]. However, the word *likely* express a second-order uncertainty on the boundary information, i.e. uncertainty over the bounded uncertainty. In this thesis, we refer this second-order uncertainty to noise in the knowledge and use Gaussian distribution function to represent this kind of noise. For each possible knowledge feature in our knowledge model, we studied the effect of such noise on our constructed model uncertainty and its consequent generalization accuracy.

Finally, we applied our method to solve real-world problems in Bioinformatics. We applied our knowledge model to capture the qualitative cause-effect relationship about the molecular interaction in signaling transduction pathway in breast cancer bone metastasis and cell proliferation. We construct a class of constrained Dynamic Bayesian networks to quantitatively predict the bone metastasis percentage and cell proliferation rate based on various interference on the expression level of the molecules in the network. We compare our in-silico simulation to the wet-lab experimental observations and show that our approach can make reasonable quantitative prediction based on only qualitative causal statements.

Further improvements on our approach could be: For high-dimensional parameter space and extreme restrictive constraints, it is computationally expensive to use Accept-Reject algorithm to sample in the model space. Therefore, more efficient sampling techniques can be used here to improve the efficiency of the algorithm, such as Monte Carlo Markov Chain (MCMC) algorithm or Gibbs sampling algorithm.

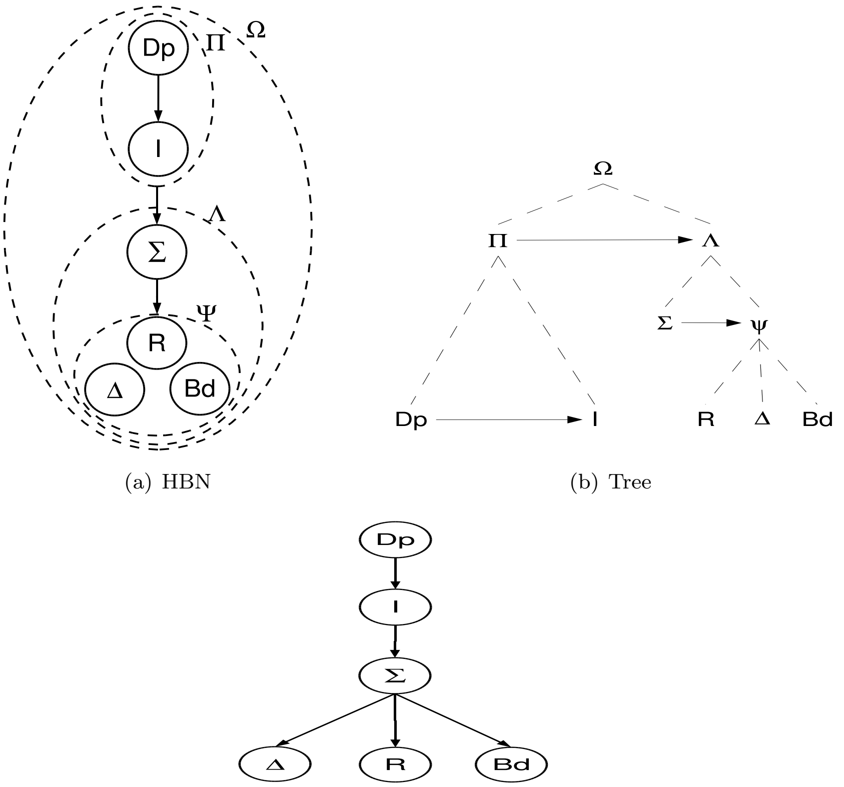
Chapter 3

Bayesian Modeling with Inconsistent Qualitative Knowledge

One significant drawback of qualitative knowledge is its potential inconsistency. In the same domain, there may exist contradicting qualitative statements on dependency, causality and parameters over a set of entities. Therefore, methods for integrating and learning semantics of inconsistent qualitative knowledge and making use of it as prior background knowledge in modeling Bayesian networks and performing quantitative prediction are definite beneficial to the Bayesian framework. In this section, we propose a novel framework for this purpose. 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 based on the hierarchical knowledge model. These knowledge components define 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.

3.1 Hierarchical Knowledge Feature Model

The qualitative knowledge for a general belief network can be represented hierarchically into a tree structure, i.e. qualitative knowledge feature tree (QKFT). Once the QKFT is formulized, constraints on structure and parameter can be generated by going through the tree top-down as shown in figure 3.1. We show that the prior probability of a knowledge component can be calculated as a product of the conditional probabilities of these dependent knowledge features. 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 state-



(a) HBN

(b) Tree

(c) BN

Figure 3.1: Hierarchical Bayesian Network on Qualitative Knowledge

ments. In these cases, only baseline knowledge model will be used to generate constraints in model space to perform inference by model averaging.

The dependent qualitative knowledge feature set can be represented by a hierarchical Bayesian network (HBN) [44]. 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 $\Omega = \{\Pi(Dp, I), \Lambda(\Sigma, \Psi(R, \Delta, Bd))\}$,

$$\Omega = \{\Pi(Dp, I), \Lambda(\Sigma, \Psi(R, \Delta, Bd))\} \tag{3.1}$$

where $\Sigma = (SP, SN, PlSyn, AdSyn, Ant, MxSyn)$. The hierarchical knowledge model is shown in Figure 3.1(a) and a tree hierarchy in Figure 3.1(b). The equivalent Bayesian network is shown in Figure 3.1(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.

$$Pr(\Omega) = Pr(\Pi)Pr(\Sigma|\Pi)Pr(\Psi|\Sigma) \tag{3.2}$$

where

$$\begin{aligned} Pr(\Pi) &= Pr(Dp)Pr(I|Dp) \\ Pr(\Sigma|\Pi) &= Pr(\Sigma|I) \\ Pr(\Psi|\Sigma) &= Pr(R|\Sigma)Pr(\Delta|\Sigma)Pr(Bd|\Sigma) \end{aligned} \tag{3.3}$$

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_1
S_3	null	null	null	null	null	null	w_3

Table 3.1: Feature-vector of Statements

$Pr(\Psi|\Sigma) = Pr(R|\Sigma)Pr(\Delta|\Sigma)Pr(Bd|\Sigma)$, $Pr(\Pi) = Pr(Dp)Pr(I|Dp)$ and $Pr(\Sigma|\Pi) = Pr(\Sigma|I)$.

3.1.1 Inconsistent Knowledge Integration

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 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 Table 3.1. The conditional probability of the features can be calculated straightforwardly by

$$\begin{array}{cc} Pr(I|Dp) = (w_1 + w_2)/w_a & Pr(\overline{I}|Dp) = w_3/w_a \\ Pr(r_1|\Sigma = SP) = w_1/w_b & Pr(r_2|\Sigma = SP) = (w_1 + w_2)/w_b \end{array} \tag{3.4}$$

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 3.1(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, $\widetilde{S} = \{S_1, S_2, S_3\}$ can be transformed to the qualitative knowledge $\widetilde{\Omega} = \{\Omega_1, \Omega_2, \Omega_3\}$ with discrete-valued features. Once we have calculated the conditional probabilities of knowledge features, the prior probability of qualitative knowledge can be computed according to Eq. 3.2. Thus the inconsistent knowledge components are ready to be reconciled. The qualitative knowledge in Table 3.1 can be described by $\widetilde{\Omega}$:

$$\begin{aligned}
\Omega_1 &= \{1, 1, SP, [10, 25], null, null\} \\
\Omega_2 &= \{1, 1, SP, [25, \infty], null, null\} \\
\Omega_3 &= \{1, 0, null, null, null, null\} \qquad (3.5)
\end{aligned}$$

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,

$$\begin{aligned}
Pr(\Omega_1) &= Pr(Dp)Pr(I|Dp)Pr(SP|I)Pr(r_1|SP) \\
&= 2/9 \\
Pr(\Omega_2) &= Pr(Dp)Pr(I|Dp)Pr(SP|I)Pr(r_2|SP) \\
&= 4/9 \\
Pr(\Omega_3) &= Pr(Dp)Pr(\overline{I}|Dp) \\
&= 1/3 \qquad (3.6)
\end{aligned}$$

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 based on Inconsistent 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.

3.2.1 Modeling with Static Bayesian Networks

A Bayesian model m represents the joint probability distribution of a set of variables $X = \{x_1, x_2, ..., x_N\}$ [48]. 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. Being different from Section 2.3.1, we 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

$$Pr(m|D,\widetilde{\Omega}) = \frac{Pr(D|m,\widetilde{\Omega})Pr(m|\widetilde{\Omega})Pr(\widetilde{\Omega})}{Pr(D,\widetilde{\Omega})} \tag{3.7}$$

The first term in the numerator of Eq. 3.7 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:

$$\begin{aligned} &Pr(X|D,E,\widetilde{\Omega}) \\ &\quad = \int Pr(X|E,m)Pr(D|m,\widetilde{\Omega})Pr(m|\widetilde{\Omega})dm \end{aligned} \tag{3.8}$$

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

$$Pr(X|E,\widetilde{\Omega}) = \int Pr(X|E,m)Pr(m|\widetilde{\Omega})dm \tag{3.9}$$

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

$$Pr(M_k|\widetilde{\Omega}) = \sum_{i=1}^{K} Pr(M_k|\Omega_i)Pr(\Omega_i|\widetilde{\Omega}) = Pr(\Omega_k) \tag{3.10}$$

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

$$Pr(m \in M_k|\widetilde{\Omega}) = Pr(m|M_k)Pr(M_k|\widetilde{\Omega}) \tag{3.11}$$

Thus, the inference on X given evidence E and inconsistent knowledge set $\widetilde{\Omega}$ in Eq. 3.9 can be written as

$$Pr(X|E,\widetilde{\Omega}) = \sum_k \int_m dm Pr(X|m,E)Pr(m|M_k)Pr(\Omega_k) \tag{3.12}$$

where $Pr(\widetilde{m|\Omega}) = \sum_k Pr(m \in M_k|\widetilde{\Omega})$. 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 perform such integration iteratively over all possible model classes with the prior distribution. The integration in Eq. 3.12 is non-trivial to compute, however, Monte Carlo methods can be used to approximate the inference.

3.2.2 Modeling with Dynamic Bayesian Networks

As discussed in Section 2.3.2, we use Dynamic Bayesian network to model the recurrent structure. An example of DBN is shown in Fig. 2.6(a) and it can be defined by a vector of 2-Time-Slice Bayesian Networks (2TBN) over time as shown in Figure 2.6(b). As Eq. 2.50, the posterior probability distribution of each node at time t, i.e. $P(X_{n,t})$ can be calculated as

$$\begin{aligned} P(X_{n,t}) &= \int_{\pi(X_n)} P(X_n|\pi(X_n))P(\pi(X_n))d\pi(X_n) \\ &= \sum_{j=1}^{J} \theta_j P_j^{t-1}(\pi(X_n)) \end{aligned} \tag{3.13}$$

where θ_j denotes the *j-th* entry in the conditional probability table of node X_n given its parents. $P_j^{t-1}(\pi(X_n))$ represent the joint probability of *j-th* configuration of the parents states at time $(t-1)$. The posterior probability distribution of X_n can be used as the priori probability for the next time step. Thus the posterior probability $P(X_{n,t})$ can be calculated iteratively over time $t = \{0, \ldots, T\}$.

In this section, we extend the knowledge-driven Bayesian inference approach with a set of inconsistent hypotheses to Dynamic Bayesian network. As demonstrated in the last section, if there is a set of inconsistent hypotheses retrieved from a publication, each independent qualitative knowledge component, $\Omega_k \in \widetilde{\Omega}$, uniquely defines a model class M_k with a vector of features. The inference on the marginal probability of X_n at time t given evidence E and the inconsistent qualitative knowledge $\widetilde{\Omega}$ with full Bayesian approach is calculated as follows, by substituting the Eq. 3.13 into Eq. 3.12, we have

$$\begin{aligned} P(X_n|E,\widetilde{\Omega}) &= \sum_k \int_m dmPr(X|m,E)Pr(m|M_k)Pr(\Omega_k) \\ &= \sum_{k=1}^{K}\sum_{w=1}^{W}\int_\Theta \sum_{j=1}^{J} \theta_{k,w,j}P_j^{t-1}(\pi(X_n),E)P(s_{k,w},\theta_{k,w,j}|\Omega_k)Pr(\Omega_k)d\Theta \end{aligned} \tag{3.14}$$

where $\theta_{k,w,j}$ represents the *j-th* entry of the CPT in *w-th* DBN model in the *k-th* DBN model class. E denotes the evidence of the observed nodes and $P_j^{t-1}(\pi(X_n),E)$ denotes the joint probability distribution of the *j-th* configuration of the parent nodes $\pi(X_n)$ at time $(t-1)$ given the observation E. In each model class, the structure which is consistent with the hypotheses component is assigned with non-zero probability $P(s_{k,w}|\Omega)$. Likewise, only parameter values on that structure, which are consistent with the contents of the hypotheses, are assigned a nonzero probability $P(\theta_{k,w,j}|s_{k,w},\Omega)$. If no further information is available, the distribution is constant in the space of consistent models.

According to Section 2.4.1, the inference in Eq. 3.14 can be approximated by performing inference in the mean model of each model class $M_k \in \widetilde{M}$ as

$$\begin{aligned} P(X_n|E,\widetilde{\Omega}) &= \sum_{k=1}^{K}\sum_{w=1}^{W}\int_\Theta \sum_{j=1}^{J} \theta_{k,w,j}P_j^{t-1}(\pi(X_n),E)P(s_{k,w},\theta_{k,w,j}|\Omega_k)Pr(\Omega_k)d\Theta \\ &= \sum_{k=1}^{K} Pr(\Omega_k)\sum_{w=1}^{W}\sum_{j=1}^{J} P_j^{t-1}(\pi(X_n),E)\int_\Theta \theta_{k,w,j}P(s_{k,w},\theta_{k,w,j}|\Omega_k)d\Theta \end{aligned}$$

(3.15)

3.2.3 ASIA Network

As Section 2.3.3, we demonstrate the Bayesian modeling scheme based on inconsistent qualitative knowledge with ASIA network in Figure 2.3.2. The parameter of ASIA network is given in Table 2.2.

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.* (www.netdoctor.co.uk)
2. *The lifetime risk of developing lung cancer in smokers is approximately 10%.* (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.* (www.quit-smoking-stop.com/lung-cancer.html)
4. *Lifetime smoker has a lung cancer risk 20 to 30 times that of a nonsmoker.* (www.cdc.gov/genomics/hugenet/ejournal/OGGSmoke.htm)
5. *15% of smokers ultimately develop lung cancer.* (www.cdc.gov/genomics/hugenet/ejournal/O GGSmoke.htm)
6. *The mechanisms of cancer are not known. It is NOT possible to attribute a cause to effects whose mechanisms are not fully understood.* (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.* (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.* [31]
9. *At least 60% of stage 4 lung cancer victims report dyspnea.* (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.* [33]
11. *95% of patients with chronic bronchitis and emphysema reported Dyspnea.* [63]

Each statement is analyzed by the hierarchical knowledge model in Figure 3.1(a) and the extracted features are summarized in Table 3.2. 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 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 [48], 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

$$\begin{array}{ccc} \mathrm{Pr(Dp)}{=}5/6 & & \mathrm{Pr}(\overline{Dp}){=}1/6 \\ \mathrm{Pr(I|Dp)}{=}1 & \mathrm{Pr}(\overline{I}|\overline{Dp}){=}1 & \mathrm{Pr(SP|I)}{=}1 \\ \mathrm{Pr}(r_1|\mathrm{SP}){=}1/5 & \mathrm{Pr}(r_2|\mathrm{SP}){=}1/5 & \mathrm{Pr}(r_3|\mathrm{SP}){=}2/5 \\ \mathrm{Pr}(r_4|\mathrm{SP}){=}1/5 & \mathrm{Pr}(b_1|\mathrm{SP}){=}1/2 & \mathrm{Pr}(b_2|\mathrm{SP}){=}1/2 \end{array} \tag{3.16}$$

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. For the collider structure of lung cancer, bronchitis, and dyspnea, the conditional probability of every knowledge features can be calculated as

$$\mathrm{Pr(Dp)}{=}1 \quad \mathrm{Pr(I|Dp)}{=}1 \quad \mathrm{Pr(PlSyn|I)}{=}1 \tag{3.17}$$

and the marginal conditional probability of knowledge features for the structure of lung cancer and dyspnea can be calculated as

$$\begin{array}{cccc} \mathrm{Pr(Dp)}{=}1 & \mathrm{Pr(I|Dp)}{=}1 & \mathrm{Pr(SP|Dp)}{=}1 & \\ \mathrm{Pr}(b_3|\mathrm{SP}){=}3/11 & \mathrm{Pr}(b_4|\mathrm{SP}){=}4/11 & \mathrm{Pr}(b_5|\mathrm{SP}){=}3/11 & \mathrm{Pr}(b_6|\mathrm{SP}){=}1/11 \end{array} \tag{3.18}$$

where $b_3 = [60\%, 69\%]$, $b_4 = [69\%, 78\%]$, $b_5 = [78\%, 90\%]$ and $b_6 = [90\%, 100\%]$. The continuous-valued feature Bd are discretized into $|Bd| = 4$ discrete-value intervals respectively. While the marginal conditional probability of knowledge features for the structure of bronchitis and dyspnea can be calculated as

$$\begin{array}{ccc} \mathrm{Pr(Dp)}{=}1 & & \mathrm{Pr(I|Dp)}{=}1 \\ \mathrm{Pr(SP|Dp)}{=}1 & \mathrm{Pr}(b_7|\mathrm{SP}){=}1/2 & \mathrm{Pr}(b_8|\mathrm{SP}){=}1/2 \end{array} \tag{3.19}$$

where $b_7 = [91\%, 94\%]$ and $b_8 = [94\%, 96\%]$. Based on the features and their prior belief, a set of qualitative knowledge $\widetilde{\Omega} = \{\Omega_1, \ldots, \Omega_{16}\}$ is formed in Table 3.2.

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 Table 2.2.

(a) Feature-vector of Statements

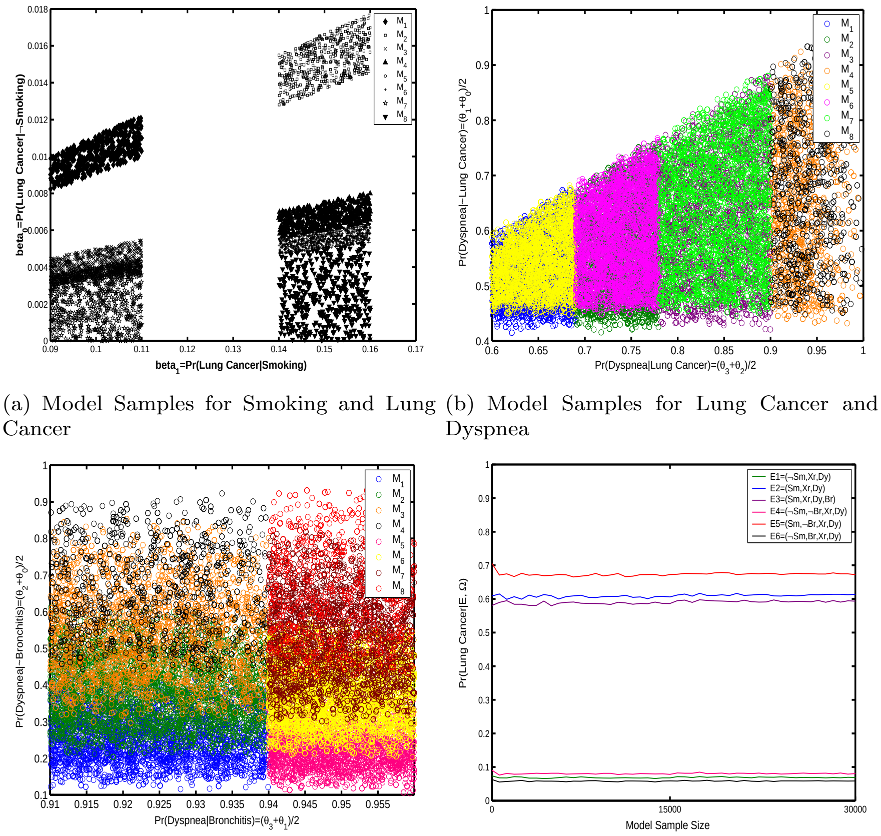
Stat.	Dp	I	Σ	R	Δ	Bd	Weight
S_1	1	1	SP	[9, 11]	null	null	w_1
S_2	1	1	SP	null	null	[9%, 11%]	w_2
S_3	1	1	SP	[25, ∞]	null	null	w_3
S_4	1	1	SP	[20, 30]	null	null	w_4
S_5	1	1	SP	null	null	[14%, 16%]	w_5
S_6	0	0	SP	null	null	null	w_6
S_7	1	1	SP	null	null	[60%, 90%]	w_7
S_8	1	1	SP	null	null	[60%, 90%]	w_8
S_9	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}

(b) Prior Probability over Inconsistent Qualitative Knowledge

$\widetilde{\Omega}$	Dp	I	Σ	R	Δ	Bd	Pr(Ω_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

Table 3.2: Qualitative Statements and Knowledge in ASIA network

Therefore the uncertainty of ASIA model space is restricted to the uncertainty of the local structure and parameter space on *Smoking* and *Lung Cancer* 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



(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

Figure 3.2: ASIA Model Sampling and Inference

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.

$$Pr(M_k) = Pr(\Omega_{10})Pr(\Omega_i)Pr(\Omega_j) \tag{3.20}$$

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 3.1(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 Table 3.2. The ground model samples are shown in Figure 3.2(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. 3.20 based on

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

Table 3.3: Inference Results on ASIA Network

$\{\Omega_k | k = 10, \dots, 16\}$ in Table 3.2. The marginal conditional probability samples are shown in Figure 3.2(b) and 3.2(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 3.2(a) to 3.2(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.

ASIA Model Inference

For each of the model sample, according to Eq. 3.12, 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 3.2(d). The true prediction values with parameters in Table 3.3 under the evidence set $\widetilde{E}$ are listed below in Table 3.3.

3.3 Empirical Study

3.3.1 Smad7 in TGFβ-Smad Pathway

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 [59, 60, 74, 100, 114], a set of qualitative statements on molecular interactions in the breast cancer bone metastasis network can be extracted and a Dynamic Bayesian model can be constructed based on this set of statements as shown in Fig. 3.3 [8, 12]. We consider the inconsistent qualitative statements with regard to the mechanism of Smad7 in blockade of the TGFβ signals. In [100], 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 [114], 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.*

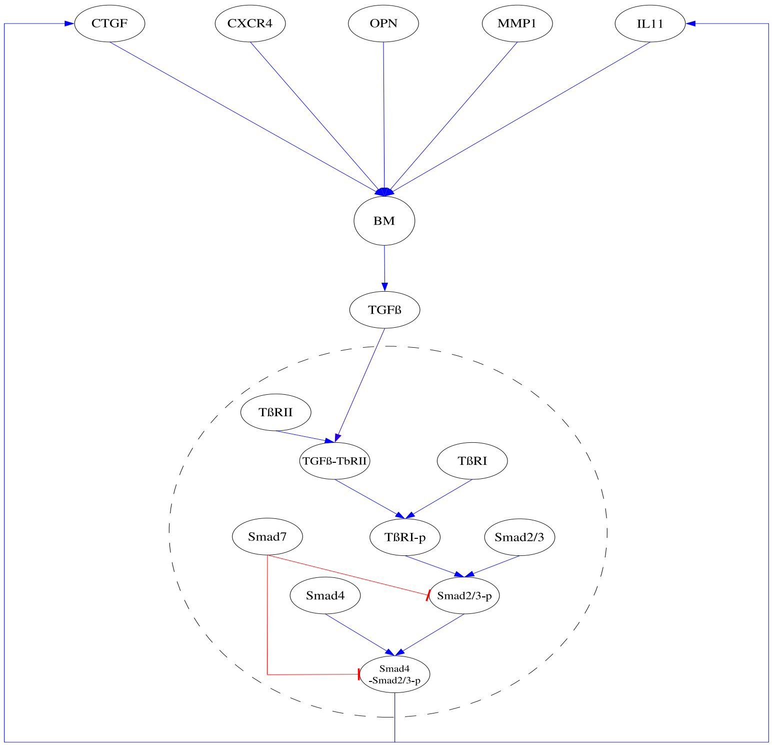


Figure 3.3: Integrated TGFβ-Smad BCBM Network and Prediction

Similar statements can be extracted from [74] 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 3.1(a) and the extracted features are summarized in Table 3.4. 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 [48], 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, $Pr(\overline{I}|\overline{Dp})$=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, $Pr(\overline{I}|\overline{Dp})$=1. Based on the features and their prior belief, a set of qualitative knowledge $\widetilde{\Omega}$ is formed in Table 3.4. In this experiment, the extended features of the inconsistent knowledge is not available and the integration with inconsistent extended

(a) Feature-vector of Statements

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

(b) Prior Probability over Inconsistent Qualitative Knowledge

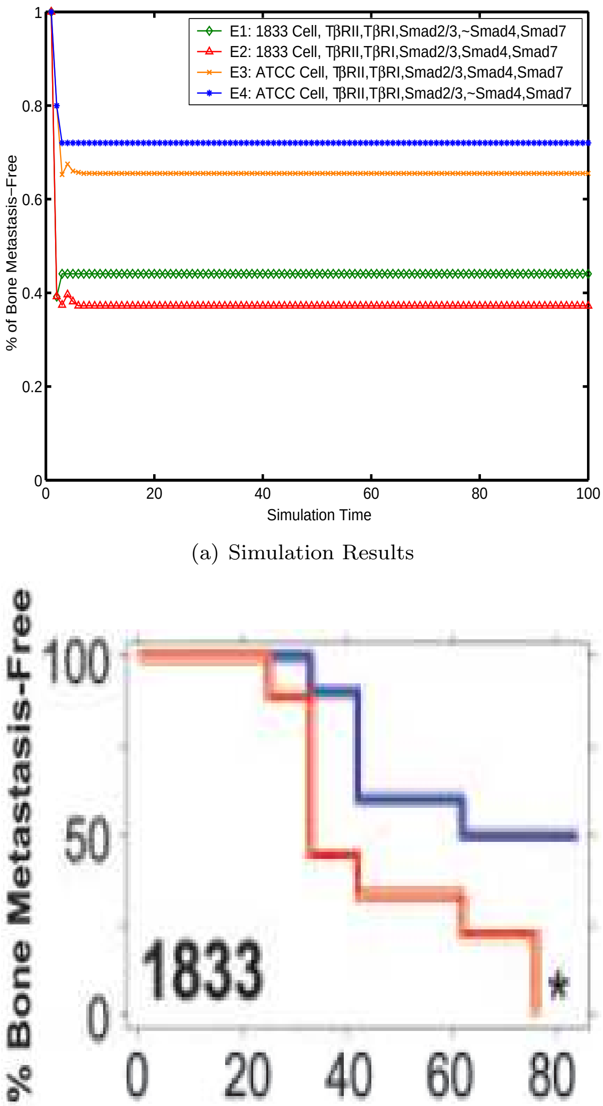
$\widetilde{\Omega}$	Dp	I	Σ	R	Δ	Bd	$\Pr(\Omega_i)$
Ω_1	0	0	null	null	null	null	1/3
Ω_2	1	1	MxSyn	null	null	null	2/3
Ω_3	1	1	MxSyn	null	null	null	1

Table 3.4: Qualitative Statements and Knowledge in TGFβ-Smad BCBM Network

knowledge features are studied in [9, 14].

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 $\{\Omega_1,\Omega_2\}$ in Table 3.4. 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. $\Pr(M_k)=\Pr(\Omega_k)$. 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. 2.1 to Eq. 2.17. 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 [78]. 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. 3.12, 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 [12]. The simulation results and the observed bone metastasis probability in [59] are shown in Fig. 3.4(a) and Fig. 3.4(b).



(a) Simulation Results

(b) Observation

Figure 3.4: Prediction in TGFβ-Smad BCBM Network

3.4 Summary

In this chapter, we extend the knowledge-driven probabilistic network modeling framework to a set of inconsistent knowledge. We investigate the method to reconcile the contradicting qualitative information and utilize these inconsistent information to make coherent quantitative reasoning. We do so by transforming the qualitative knowledge model into a hierarchical model in which knowledge features are encoded and quantified by their mutual (in)dependences and the associated conditional probability. Given the hierarchical knowledge model, a set of inconsistent knowledge are dissected and the conditional probabilities of the knowledge features are calculated by evaluating their statistics. Expert belief can be integrated into the method as a prior belief on the inconsistent knowledge components. Each knowledge component uniquely define a class of constrained Bayesian networks as in the last chapter. The conditional probability of knowledge features are used to compute the knowledge prior, i.e. the joint probability over the feature space. Thus, multiple classes of Bayesian networks are inferred from the inconsistent knowledge which are weighted by their corresponding knowledge prior distribution. The incoherent Bayesian network classes are reconciled in this way consistently into uniform representation and the averaged quantitative prediction can be calculated over all ground models in one class and over all classes.

Chapter 4

Bayesian Modeling with Incomplete Qualitative Knowledge

It is well known that knowledge are often incomplete just like the data. The incompleteness of knowledge can be best described by the assertion that the information represented in the knowledge on a set of variables in a domain at a particular time point may be complemented by the new discoveries in the same domain. The incomplete knowledge distinguish from the inconsistent knowledge by providing information on a newly discovered set of variables and/or the connections of the new variable set to the existing variables in the same domain whereas the inconsistent set of knowledge only declares contradicting structural and/or parameter information on the same set of variables. Therefore, when we construct probabilistic networks based on a set of qualitative knowledge as introduced in Section 2 and if this set of qualitative knowledge are composed by complementary components, we shall obtain a set of supplementing Bayesian networks, each of which consists of graphs over (partial) different set of variables in a domain with quantified edges. In this section, we propose several methods to fuse such Bayesian networks which are built on the complementary knowledge set.

4.1 Incomplete Qualitative Knowledge and Bayesian Network Fusion

In this case, we study the computational aspects of the knowledge-based framework [12,13] for utilizing the spatial and temporal knowledge properties in modeling and integrating Bayesian networks based on incomplete sets of knowledge. We apply our approach to model the TGFβ-Smad signaling pathway in the breast cancer bone metastasis network. Firstly, We integrate two complementary sets of knowledge in time order and form a Smad-dependent breast cancer bone metastasis network*(Smad-BCBM)*. Secondly, we integrate the TGFβ-Smad signaling pathway into the Smad-BCBM network spatially by replacing the aggregate type in Smad-BCBM with TGFβ-Smad pathway. Quantitative inference on the probability of the bone metastasis are performed by model averaging based on the qualitative knowledge model [13]. We show that our method is able to consistently integrate the sets of complementary knowledge in space-time dimension and produce reasonable quantitative predictions.

4.1.1 Incomplete Qualitative Knowledge

Like most physical systems, the knowledge has spatial and temporal properties, i.e. knowledge exist in space-time dimension. The spatial property describes that the knowledge represents information on a specific local structure of a domain and the temporal property of the knowledge states that knowledge represents human understandings of a domain at a particular time point. Therefore, incompleteness of the qualitative knowledge are two-fold: 1)At the same time, different knowledge can be drawn from independent studies and experiments to describe various local structures with distinct set of variables and connections in the same domain. 2)At the same location of a domain, identified variables and/or their connections to the existing variables can be provided by the new discoveries.

For example, one piece of knowledge Ω_1 concluded that "*Transforming Growth Factorβ (TGFβ) activates gene CTGF*" and in a later study, another piece of knowledge Ω_2 has identified that "*Smad-family proteins participate in the signaling pathway and mediate the TGFβ signals on gene CTGF*". Ω_1 and Ω_2 are incomplete knowledge and complementary to each other. The newly identified variable Smad-pathway in Ω_2, is supplementing to the existing knowledge on $TGF\beta$ and $CTGF$ in Ω_1, i.e. the influence from $TGF\beta$ on CTGF identified in Ω_1 is mediated by the Smad-pathway in stated Ω_2. Thus, this set of knowledge can be integrated and the resulting structure is a feed-forward chain TGFβ→Smad-pathway→CTGF which can be explicitly derived from Ω_1 and Ω_2. This scenario demonstrate that knowledge's temporal property.

A third piece of knowledge Ω_3 states that: CXCR4 and IL11 are found to promote breast cancer bone metastasis which in turn increase the level of TGFβ. Ω_3 provide different local information to Ω_1 and Ω_2 in that Ω_3 describes a different local structure than the forward chain represented by Ω_1 and Ω_2. However, these different local structures exist in the same domain related to breast cancer bone metastasis. This scenario demonstrate that knowledge's spatial property.

4.2 Bayesian Inference with Incomplete Qualitative Knowledge

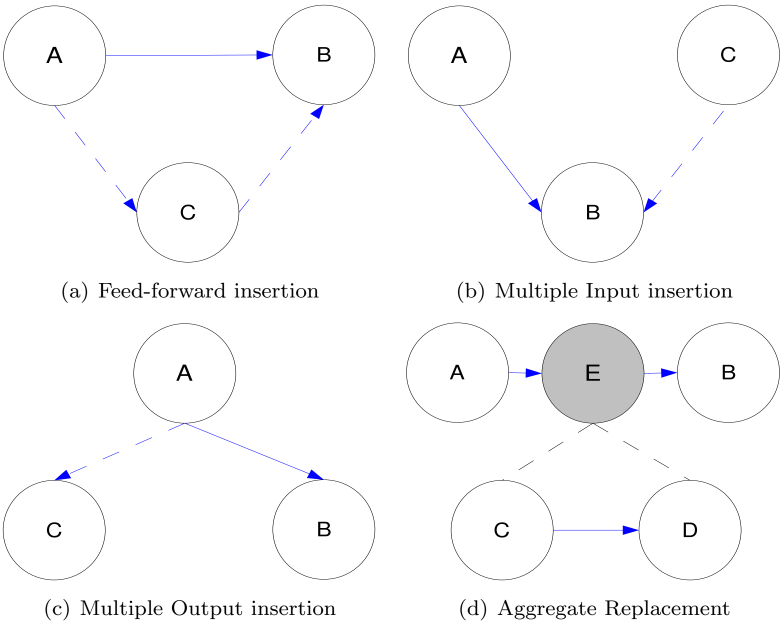
4.2.1 Bayesian Network Fusion

Bayesian networks built on the incomplete knowledge may represent only a specific local structure of an underlying system at a particular time, thus, the Bayesian networks are incomplete as well. However, in many cases, the network is required to be extended by including external nodes in the original network and/or replace single node in the original network with a set of nodes. The extended network has a different skeleton and set of variables comparing to the original network. For example, in molecular interaction network, the nodes represent biological entities at different scales, such as protein, mRNA, DNA, aggregate type and phenotype. Previously unobserved proteins may be identified to participate in the interactions established by earlier studies, thus need to be included in the molecular interaction network or aggregate type node shall be replaced by a network with only protein and DNA at higher resolution. Bayesian model fusion in space-time dimension can be described as the inclusion of external nodes into an existing Bayesian network as well as the hierarchical replacement of the aggregate type in the existing model by an external Bayesian network. The model integration represents the procedure where knowledge-based computational framework makes use of the spatial and temporal properties of the incomplete knowledge to integrate a vector of complementary Bayesian networks and form a uniform representation. Therefore, it is crucial for knowledge-based framework to deal with these knowledge properties and consistently integrate the Bayesian networks based on the set of knowledge in space-time dimension to form an up-to-date and complete representation of the underlying system.

According to [8,10], fusion of multiple Bayesian networks based on the incomplete set of qualitative knowledge can be performed under two scenarios: i)The structure of the integrated model can be explicitly derived from the knowledge; ii)The structure of the integrated model is ambiguous and can not be inferred directly. In the first case, [8] has shown that the integration problem boils down to the determination of the parameters given the integrated model structure. This method provides a basis for Bayesian model integration based on the precondition that integrated model structure can be explicitly derived and the joint distribution of the integrated model given the structure can be determined by setting an equality constraint on the (marginal) joint distributions of the variables in the integrated models. This method is based on the assumption that a physical system should remain constant regardless to the degree of human knowledge on it. If we use Bayesian model to represent the physical system, the marginal joint probability distribution of the original models should be invariant during the model integration. However, this presumption is not always satisfied in many cases, i.e. the structure of the integrated network is not always directly derivable. In the second scenario, the set of complementary knowledge impose uncertainty over the structure space of integrated model. Therefore, it is required to learn the integrated model structure from the mixture of *local statistics* represented by every incomplete model. The set of local statistics are combined together to form the *global statistics* and the complete model are learned from the global statistics. This approach assumes that the set of complementary qualitative hypotheses represent the complete set of variables in the domain, thus, the hidden variables of a specific network can be identified as those which are unobserved in this network but observed by others.

4.2.2 Bayesian Network Fusion in Parameter Space

If the structure of the integrated model can be explicitly derived from the set of incomplete knowledge, the Bayesian fusion problem boils down to determination of integrated network parameters based on the constraints in Section 2 and the equality constraints on the (marginal) joint space. For example, the structure of the integrated network with TGFβ, Smad-pathway and CTGF can be obtained directly by combining the body of knowledge in Ω_1 and Ω_2. The pa-



(a) Feed-forward insertion (b) Multiple Input insertion

(c) Multiple Output insertion (d) Aggregate Replacement

Figure 4.1: Model Integration with Single Node

rameters in the integrated model consists of the interactions between TGFβ and Smad-pathway as well as the interactions from Smad-pathway to CTGF. These interactions can be described by *positive single influence* in Eq. 2.2. Meanwhile, the strength of interaction between TGFβ and CTGF is consistent during the model integration, i.e. P_{Ω_1}(CTGF,TGFβ)=$P_{\Omega_{1,2}}$(CTGF,TGFβ). This method provides an basic solution to the incomplete knowledge integration problem since it assumes that the structure of the integrated model, i.e. (in)dependency between the variables in the model, can be explicitly derived from the complementary set of knowledge. In [8], it is proposed that the Bayesian networks built on the incomplete knowledge can be integrated by firstly constructing the structure of the integrated model based on the combination of the knowledge and then sampling the joint probability distributions over the integrated parameter space based on the qualitative knowledge model and the equality constraints on the (marginal) joint probability distribution over the variables in the individual networks.

Assume that a Bayesian model is composed by node A, B and a directed edge from A to B. The integration of a node C can be shown in Figure 4.1 which changes the structure according to the position of its insertion. In Figure 4.1(a), the node is inserted along the edge between A and B to form a feed-forward chain, e.g. a molecular(C) is identified in a signal transduction pathway between A and B. In Figure 4.1(b), the node is inserted as a parent, e.g. a protein(C) is identified to be a co-transcript factor of A to activate a gene(B). In Figure 4.1(c), the node is inserted as a child, e.g. a disease(C) is verified to be caused by the mutation of a gene(A). If the existing (dynamic) Bayesian model consists aggregate types [44], the model integration in Figure 4.1(d) can be performed by replacing the aggregate type(E) in the model with the alternative networks of C and D, e.g. a signaling pathway is replaced by a set of protein-protein interactions.

The possible model integration patterns with single node are shown in Figure 4.1 which may have different structures due to the distinct insert position of the new set of nodes. The continuous lines in the figure denote the original graph with initial set of variables as well as the dashed lines stand for the insertion of new set of nodes into the mode during model integration. However, the joint probability of the nodes in the original network are constant. In Fig. 4.1(a), node C is inserted in the middle of the pathway between node A and B as to d-separate them. In this case, the equality constraint on the marginal probability of nodes (A, B) can be written as

$$\begin{aligned} P^0(A, B) &= \int_C P(A, B, C) dC \\ &= \int_C P(B|C)P(C|A)P(A) dC \end{aligned} \tag{4.1}$$

In Fig. 4.1(b), node C is inserted as the parent node of B. In this case, the equality constraint on the marginal probability of nodes (A, B) can be written as

$$\begin{aligned} P^0(A, B) &= \int_C P(A, B, C) dC \\ &= \int_C P(B|A, C)P(C)P(A) dC \end{aligned} \tag{4.2}$$

In Fig. 4.1(c), node C is inserted as the child node of B. In this case, the marginal probability of nodes (A, B) is not affected by the node C, i.e.

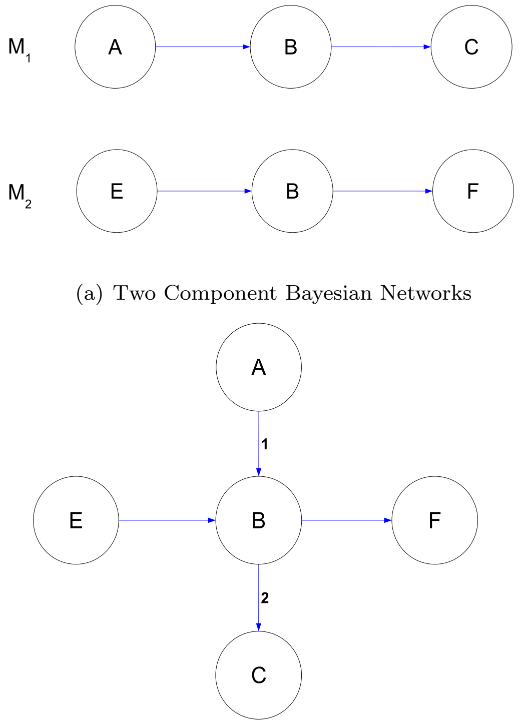
$$P^0(A, B) = P(A, B) \tag{4.3}$$

In Fig. 4.1(d), node C and node D are inserted to replace the aggregate type node E. In this case, the inserted network with multiple nodes preserve its d-separate property. But the original networks with node A and B changes its d-separation properties. For example, in the original network, given node E, node B are independent from node A whereas, in the integrated network, node E is replaced by node C and D and d-separation between A and B are described as given node C or D or nodes (C,D) renders the independence of node A and B. This means that the set of parents of node B is changed during the model integration. By comparison, in the inserted network, the parent of node D during the integration is constant. Thus, in this case, the marginal probability of nodes (A, B) can be calculated by marginalizing out the nodes by which node A and B are d-separated

$$\begin{aligned} P^0(A, B) &= \int_{C,D} P(A, B, C, D) dC dD \\ &= \int_{C,D} P(B|D)P(D|C)P(C|A)P(A) dC dD \end{aligned} \tag{4.4}$$

Since the parent of node D is stable during the integration. The joint probability over C and D is constant as:

$$P^0(C, D) = P(C, D) \tag{4.5}$$



(a) Two Component Bayesian Networks

(b) Integrated Bayesian Networks

Figure 4.2: Model Integration with Multiple Nodes

The model integrations with a set of multiple nodes can be decomposed into a series of integration steps with single node as shown in Fig. 4.1. The constancy of the marginal joint probability distribution is satisfied in each integration step. Note that the equality constraint on the marginal probability is symmetric to both Bayesian network components which are fused together. For example, two component Bayesian networks in Fig. 4.2(a) are fused to form the integrated Bayesian network in Fig. 4.2(b). This integration procedure involves multiple nodes which can be decomposed into a multiple steps of single node insertion. The number on the edge in Fig. 4.2(b) denote the step order of the integration process.

In the first step, node A in the M_1 is inserted as a parent onto node B in M_2. In this case, only the node B's parent(s) changed during the integration, therefore, the equality constraints on the marginal joint probability of (A,B) and (E,B) shall be imposed.

$$\begin{aligned} P^0(A,B) &= \int_E P(B|A,E)P(E)dD \\ P^0(E,B) &= \int_A P(B|A,E)P(A)dA \end{aligned} \tag{4.6}$$

In the second step, node C in the M_1 is inserted as a child node of B in M_2. In this case, given node B, node C and F are independent to each other and

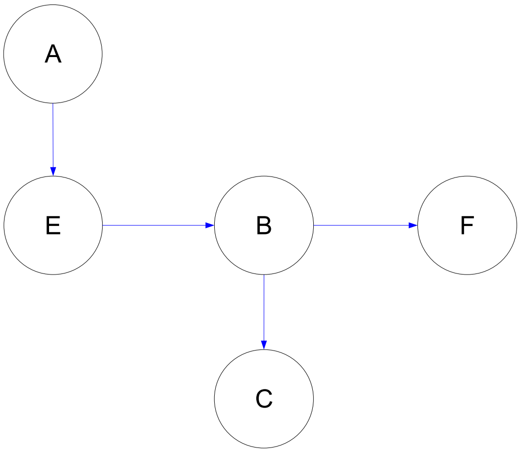


Figure 4.3: Structure Uncertainty in Model Fusion

are d-separated from the rest of nodes in the network. Therefore, the joint probability of (B,C) and (B,F) are constant as

$$\begin{aligned} P^0(B,C) &= P(B,C) \\ P^0(B,F) &= P(B,F) \end{aligned} \tag{4.7}$$

In Eq. 4.1 to Eq. 4.7, $P^0(*)$ indicate the probability in the original networks before model integration and $P(*)$ represents the probability in the integrated Bayesian network. The equality constraints describes the fact that a physical system will remain constant regardless to the degree of human knowledge on this system.

4.2.3 Bayesian Network Fusion in Structure Space

The discussion in the above section is based on the assumption that structure of the integrated Bayesian network is known which is not always valid since the (in)dependency between variables in the integrated model may not be exactly inferred based on a combination of the incomplete knowledge. In this case, the integration of the individual models involves i)determination of the model structure and ii)computation of the (marginal) joint probability distributions over the variables in the integrated model. The set of incomplete knowledge impose uncertainty over the structure space of integrated model. For example, in the Bayesian network fusion with component networks M_1 and M_2 in Fig. 4.2(a), if we do not know the structure of the integrated network as in Figure 4.2(b), the structure of the integrated network can not be inferred directly from these knowledge. In these scenarios, the interaction between the nodes in M_1 and M_2 can be indirect connections, such as the edge between A and B can be mediated through D as shown in Fig. 4.3. Therefore, it is required to fit a "good" model into the uncertainty by learning the integrated network from the mixture of *local statistics* sampled by every component Bayesian networks, i.e. learn complete network structure from *global statistics*. Moreover, due to the uncertainty in the knowledge, there may be a large set of possible structures all of which can explain the set of incomplete qualitative hypotheses equally well. Without further information, our method take each possible structure for granted and adopt full Bayesian approach in making belief inference.

Local Statistics and Global Statistics

The component Bayesian networks are constructed based on the qualitative knowledge as introduced in Section 2 and the statistics of the individual hypothesis can be sampled by the artificially generating data samples from each component Bayesian networks. The set of samples from each component Bayesian network are combined to form a global representation of the statistics over the variables. The unobserved nodes in a model are treated as hidden variables. The combined data are used as training data to learn the structure of the fused Bayesian network and the set of learned structure candidates are formalized to perform quantitative inference by calculating the average of the inference from each candidate model.

Assume a set of incomplete qualitative hypotheses $\overline{\Omega} = \{\Omega_1, \ldots, \Omega_H\}$, each qualitative hypothesis Ω_h defines the structural and parameter space of a class of Bayesian models, $\overline{M}_h = \{s_h, \Theta_h\}$, over a set of variables $\overline{X}_h = \{X_{n,h} | n = 1, \ldots, N\}$.

Local Statistics The local statistics of each Bayesian model class are described by artificially drawing samples out of the models in the class. The data samples of *k-th* Bayesian model in *h-th* class, $D_{k,h}$, can be drawn based on the local statistics of *h-th* model class $\overline{D}_h$ as

$$\overline{D}_h = \{D_{k,h} | k = 1, \ldots, K\} \tag{4.8}$$

Global Statistics The complete set of variables in the domain can be described as $\overline{X} = \{\overline{X}_h | h = 1, \ldots, H\}$, i.e. a non-redundant combination of the variables in each model class across $\overline{\Omega}$. The hidden variables for *h-th* model class can be described as

$$\overline{Y}_h = \{X | X \in \overline{X}, X \notin \overline{X}_h\} \tag{4.9}$$

Thus, the combined training data $\overline{D}_F = \{\overline{D}_h | h = 1, \ldots, H\}$ contains missing values at $\overline{Y}_h$ in $\overline{M}_h$. In this manner, we fuse the local statistics of individual Bayesian model into the global statistics to represent a complete uncertainty over the structure and parameter space of a domain which is consistent with the semantics of the complementary qualitative hypotheses $\overline{\Omega}$. Thus, the complete Bayesian network can be constructed through learning algorithm based on the global statistics. The global statistics over the set of incomplete knowledge can be described as

$$\overline{D}_F = \{\overline{D}_h | h = 1, \ldots, H\} \tag{4.10}$$

Bayesian Model Fusion by SEM based on Global Statistics

As introduced in Section 1.2.3, given $\overline{D}_F$ and $\overline{\Omega}$, the posterior probability of network structure s can be formulated as

$$P(s|\overline{D}_F, \overline{\Omega}) = \frac{P(\overline{D}_F|s, \overline{\Omega})P(s|\overline{\Omega})}{P(D_1, \ldots, D_H)} \tag{4.11}$$

The first term of numerator is the data marginal likelihood and the second term denote the prior distribution over the network structure space given the set of qualitative knowledge. In general, if the training data has full observation, we can learn the Bayesian network structure with BIC score(in Eq. 1.11). Further, if we assume the domain to satisfy five assumptions, namely, i)Multinomial distribution of the training data; ii)Parameter independency; iii) Parameter modularity; iv)Dirichlet prior; v)Full observation, we can use BD score(in Eq. 1.16) to learn the Bayesian model structure.

However, our combined dataset $\overline{D}_F$ contains missing variables $\overline{Y}_h$ for *h-th* component. The difficulty with learning from missing values is that the data likelihood is no longer decomposable [37, 39]. In Section 1.2.3, *Structural Expectation-Maximization (SEM)* algorithm [37] is proposed to solve this problem by iterative steps. In E-step, the missing values are "filled-in" by computing the *expected counts* based on the MAP parameter estimation from current structure and parameters; In M-step, the *expected counts* are used to calculate the *expected BD score*, i.e. *E(BDs)* of each candidate model produced by structure searching algorithm, e.g. greedy hill climbing, and *E(BDs)* is decomposable as the full observation case. Finally the "best" structure with the maximum score is selected as the learned structure.

Here we give an example to fuse two component Bayesian networks shown in Fig. 4.2(a) by assuming the structure of integrated network is unknown and can not be inferred from the set of incomplete knowledge. The structure of these component Bayesian networks can be described as s_1 and s_2. Two qualitative knowledge can be used to exemplify the interactions between the variables.

1. Ω_1: Protein A activates gene B to cause disease C;
2. Ω_2: Protein E activates gene B to cause disease F;

The parameters together with s_1 and s_2 define two Bayesian model classes, $M_1(s_1, \Theta_1)$ and $M_2(s_2, \Theta_2)$. These Bayesian model classes are used to sample local statistics. According to Eq. 4.8, local statistics $\overline{D}_1$ and $\overline{D}_2$ can be generated by each ground Bayesian model in each class. We sample K=10,000 ground Bayesian models in each class and sample 3 data points from ground model as

$$\begin{aligned} \overline{D}_1 &= \{D_{1,1}, \ldots, D_{1,K}\} \\ &= \{A_{1,1}, \ldots, A_{1,3K}, B_{1,1}, \ldots, B_{1,3K}, C_{1,1}, \ldots, C_{1,3K}\} \end{aligned} \tag{4.12}$$

$$\begin{aligned} \overline{D}_2 &= \{D_{2,1}, \ldots, D_{2,K}\} \\ &= \{E_{2,1}, \ldots, E_{2,3K}, B_{2,1}, \ldots, B_{2,3K}, F_{2,1}, \ldots, F_{2,3K}\} \end{aligned} \tag{4.13}$$

The local statistics of the two classes of Bayesian models can be combined into a global statistics as

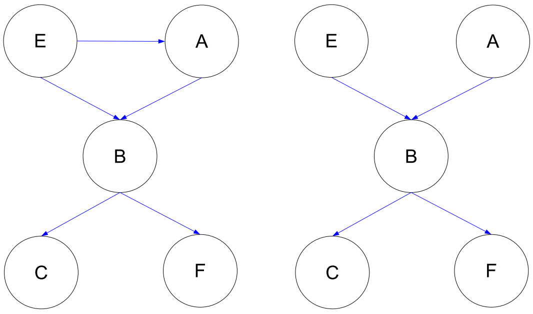
$$\overline{D}_F = \{\overline{D}_1, \overline{D}_2\}$$

A	B	C	E	F
0	1	1	×	×
1	1	0	×	×
⋮	⋮	⋮	⋮	⋮
×	0	×	1	1
×	1	×	0	1
⋮	⋮	⋮	⋮	⋮

Table 4.1: Example of Global Statistics

$$\begin{aligned} = \quad & \{A_{1,1}, \ldots, A_{1,3K}, A_{2,1}, \ldots, A_{2,3K}, \\ & B_{1,1}, \ldots, B_{1,3K}, B_{2,1}, \ldots, B_{2,3K}, \\ & C_{1,1}, \ldots, C_{1,3K}, C_{2,1}, \ldots, C_{2,3K}, \\ & E_{1,1}, \ldots, E_{1,3K}, E_{2,1}, \ldots, E_{2,3K}, \\ & F_{1,1}, \ldots, F_{1,3K}, F_{2,1}, \ldots, F_{2,3K}\} \end{aligned} \tag{4.14}$$

Now we wish to learn a complete Bayesian network with two component models in Fig. 4.2(a). For simplicity, we assume that we keep the prior information on the edge direction to differentiate the Bayesian models in an equivalent class in Section 1.2.1. We apply SEM algorithm to learn a set of candidate models from various initial graph with Greedy-hill climbing method. The results are shown in Fig. 4.4. Although it may seems reasonable to integrate the two models by combining the edge from A to B in M_1 in the Figure 4.2(a) and the edge from E to B in M_2 in the Figure 4.2(a) to form a collider structure from *(A,E)* to B as shown in Figure 4.4(b), this intuitive method may simply neglect the uncertainty of the qualitative hypotheses and ignores other possible structures which could be the true model. We applied SEM algorithm with BD score to learn the integrated model and the result shows that the structure in Fig. 4.4(a) can be learned to represent the uncertainty encoded in the set of complementary qualitative hypotheses. Structure $\hat{S}_1$ indicates the scenario where protein A interacts with protein E to activate gene B and cause disease D.



(a) $\hat{S}_1$ (b) $\hat{S}_2$

Figure 4.4: Incomplete Bayesian Network Structure Fusion Example

4.3 Empirical Study

We apply the proposed computational framework to fuse a set of incomplete Bayesian networks which model the molecular signaling transduction pathways in breast cancer based on a set of incomplete qualitative knowledge. We conduct our empirical study in two scenarios: i)The structure of the integrated Bayesian network can be explicitly inferred from the knowledge. ii)The structure of the integrated network is unknown and must be learned from the global statistics. The probability of bone metastasis is predicted by taking the average of the inferences in each Bayesian model candidates.

4.3.1 Integration of TGFβ-Smad pathway into Smad-dependent Breast Cancer Bone Metastasis Network

In the first scenario, a set of incomplete qualitative hypotheses has been extracted from the publications [59,60,74,100,114]. In [59,60], the studies address the network of breast cancer bone metastasis and Smad-dependent pathway. In [74, 100, 114], the TGFβ-Smad signaling pathway is studied. We shall integrate the TGFβ-Smad signaling pathway into the Smad-dependent breast cancer bone metastasis network by assuming that we know exactly the connections between the molecules through the pathway and make predictions on the bone metastasis. In this scenario, we perform multiple-step model integration. In the first step, Smad-dependent pathway node are inserted into the breast cancer bone metastasis network to form Smad-dependent Breast Cancer Bone Metastasis (Smad-BCBM) network. In the second step, we replace the Smad-dependent pathway node(aggregate type node) in Smad-BCBM network with a network of Smad-family proteins. We compute the equality constraint in the marginal joint probability space during these model integration.

Integrating Smad-dependent Pathway into Breast Cancer Bone Metastasis Network

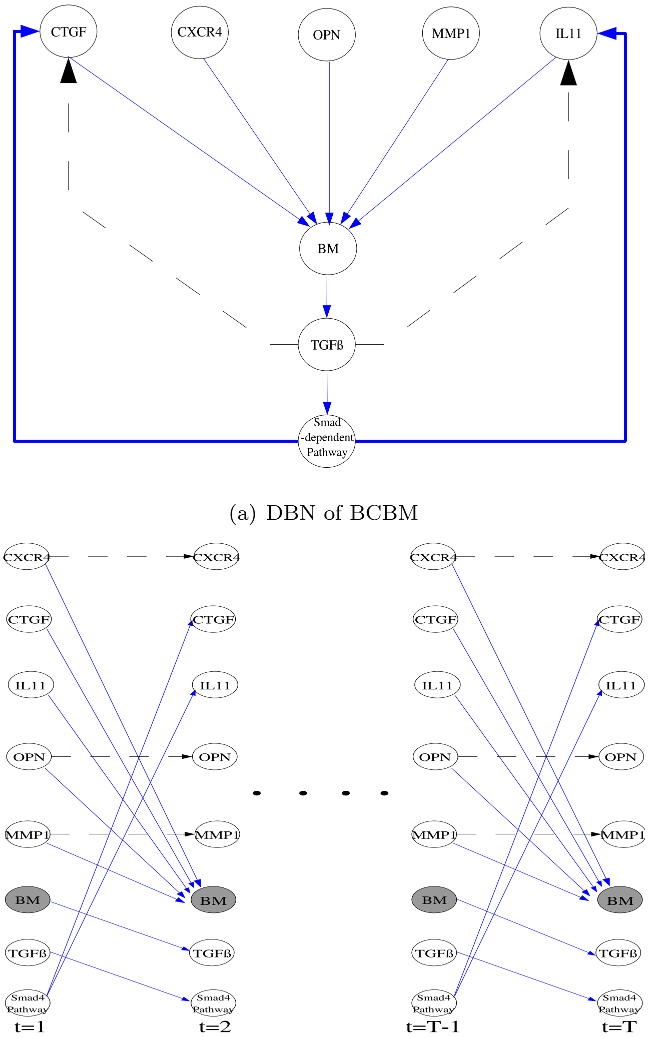
In [60], Kang identified a set of candidate genes which are responsible for promoting breast carcinoma metastasis to bone. This set of genes include the bone homing receptor CXCR4, the osteoclast-activating cytokine Interleukin-11 (IL11), the proteolytic factor MMP1 (matrix metalloprotease-1, or collagenase-1), the angiogenic factor CTGF (connective tissue growth factor) and Osteopontin (OPN). The protein TGFβ further enhances the activity of bone metastasis by increasing the expression level of CTGF and IL11. A set of qualitative hypotheses on the function of these genes in the formation of bone metastasis is extracted in $\overline{\Omega} = \{\Omega_1, \Omega_2, \Omega_3\}$ where

Ω_1*: These genes act cooperatively to cause osteolytic metastasis.*

Ω_2*: Two of these genes, interleukin-11 and CTGF, encode osteolytic and angiogenic factors whose expression is further increased by the prometastatic cytokine TGFβ.*

Ω_3*: It has been proposed that TGFβ released during osteolysis supports a cycle of metastatic breast cancer stimulation.*

In [12], the breast cancer bone metastasis network is generated based on this set of hypotheses. The bone metastasis ability of different cell lines, e.g. 1834, ATCC and 1833, as well as the bone metastasis probability of these cell lines



(a) DBN of BCBM

(b) 2TBN of BCBM

Figure 4.5: Smad-BCBM Network

with various genetic transfections has been successfully simulated. However, in a later study [59], Kang has verified that *Smad-dependent pathway* involves in the breast cancer bone metastasis. A set of qualitative hypotheses on the function of Smad-dependent pathway is extracted in $\overline{\Omega'} = \{\Omega_4, \Omega_13\}$ where Ω_4*: Results suggest that breast cancer cells undergo Smad-dependent transcriptional activation in the bone microenvironment.*

Ω_5*: The induction of CTGF and IL11 in breast cancer bone metastasis is a Smad-dependent process.*

The breast cancer bone metastasis (BCBM) network with Smad-dependent pathway form a recurrent structure as shown in Figure 2.21(a) [60] and the original Dynamic Bayesian network of breast cancer bone metastasis is shown in Fig. 2.21(b). Based on the hypotheses $\overline{\Omega}$, the interactions from TGFβ to

(a) CPT of IL11 given Smad-pathway

Smad-pathway	ρ_j=P(IL11|Smad-p.)
0	ρ_0
1	ρ_1

(b) CPT of CTGF given Smad-pathway

CTGF	α_j=P(CTGF|Smad-p.)
0	α_0
1	α_1

(c) CPT of Smad-p. given TGFβ

TGFβ	ϕ_j=P(Smad-p.|TGFβ)
0	ϕ_0
1	ϕ_1

Table 4.2: CPTs for Smad-BCBM Network

CTGF and IL11 can be modeled as direct edges by the dashed curves in Figure 4.5(a) [12]. However, according to the hypotheses $\overline{\Omega'}$, the external node *Smad-dependent pathway* which was unobserved previously in [60] has been verified for existence by a later study [59], thus, it can be integrated in the existing BCBM network by a feed-forward chain insertion(Figure 4.1(a)) between TGFβ, CTGF and IL11 as shown by the bold lines in Figure 4.5(a) which is a Dynamic Bayesian network. The integrated model forms a Smad-dependent breast cancer bone metastasis (Smad-BCBM) network and the *Smad-dependent pathway* represents an network of interactions between TGFβ and Smad-family proteins, thus, can be treated as an aggregate type. The dynamic Bayesian model can be unrolled over time into a vector of 2TBNs as shown in Figure 4.5(b). The parameters β, γ, λ and ζ can be descibed by the conditional probability tables in Table 2.5 and 2.6 and the parameters ρ, α and ϕ are described by the conditional probability tables(CPT) in Table 4.2.

According to [12,13], the parameters of the original BCBM network ζ can be constrained by *plain synergy with positive individual influence*, as well the parameters, β, γ and λ can be restrained by *single positive influence(SP)*. Similarly, the parameters in the integrated network, i.e. α, ρ and ϕ can be constrained by *SP* as

$$\alpha_0 \leq \alpha_1 \qquad \rho_0 \leq \rho_1 \qquad \phi_0 \leq \phi_1 \tag{4.15}$$

Given TGFβ, the joint probability distribution of (TGFβ, CTGF) and (TGFβ, IL11) in the original BCBM network are determined by the parameters λ and γ respectively [12]. The (marginal)joint probability distribution in the integrated model, Smad-BCBM, can be computed by marginalize the newly added *Smad-dependent Pathway(S)* as

$$Pr(C,T) = P_T(\textstyle\sum_S \alpha\phi) \quad Pr(I,T) = P_T(\textstyle\sum_S \rho\phi) \tag{4.16}$$

where C stands for CTGF, I represents IL11, T states TGFβ and P_T equals to the probability of TGFβ. Based on Section 4.2.2, to ensure the consistency of the model integration, the marginal joint probability of (TGFβ, CTGF) and (TGFβ, IL11) in Eq. 4.16 can be constrained by λ and γ approximately as

$$\lambda \approx \textstyle\sum_S \alpha\phi \quad \gamma \approx \textstyle\sum_S \rho\phi \tag{4.17}$$

where $\sum_S \alpha\phi \in [\lambda - \Delta, \lambda + \Delta]$ and $\sum_S \rho\phi \in [\gamma - \Delta, \gamma + \Delta]$. Δ is a small quantity.

Modeling TGFβ-Smad Signaling Pathway

Transforming Growth Factor β controls a diverse set of cellular processes, including cell proliferation, recognition, differentiation, apoptosis, by the activation of Smad proteins through ligand-receptor binding, phosphorylation and transcriptional regulation of target gene expression. The mechanism of Smad-mediated TGFβ signaling pathway is shown in Figure 4.6(a) [74]. A set of qualitative hypotheses $\overline{\Omega} = \{\Omega_6, \ldots, \Omega_{11}\}$ on the protein-protein interactions in this pathway can be extracted from the publications [74, 100, 114] and described below:

Ω_6*: TGFβ ligand binds to the type II receptor serine/threonine kinases*

Ω_7*: The TGFβ-bound type II receptor kinases phosphorylate the receptor I kinase*

Ω_8*: The phosphorylated type I receptor kinases phosphorylate the R-Smad, Smad2/3*

Ω_9*: The inhibitory Smad, Smad7, may inhibit the signaling transduction by competing with Smad2/3 for type I receptor kinases*

Ω_{10}*: The phosphorylated Smad2/3 and Co-Smad, Smad4, form complex to activate the target gene*

Ω_{11}*: The inhibitory Smad, Smad7, may inhibit the signaling transduction by competing with Smad4*

Based on the qualitative knowledge model in [13], the set of qualitative hypotheses can be used to generate a vector of constraints on the structure and parameter space of the TGFβ-Smad signaling pathway. The structure can be represented by a static Bayesian network as shown in Figure 4.6(b) and the parameters can be described by the conditional probability tables (CPT) in Table 4.3.

The interaction between TGFβ and type-II receptor (TβRII) in Ω_6 can be described as *ligand-receptor binding.* Since TGFβ binds to TβRII and form a complex TGFβ-TβRII, the probability of the complex with sufficient amount of all reactants (η_3) are higher than when there is insufficient amount on partial reactants ($\eta_{1,2}$) which in turn are higher than none of the reactants are of sufficient amount (η_0). In addition, $\eta_{1,2}$ may be small enough such that η_3 is larger than their sum. Therefore, the constraint rules in the parameter space of the *ligand-receptor binding* interaction can be described by *additive synergy* in [13] as follows:

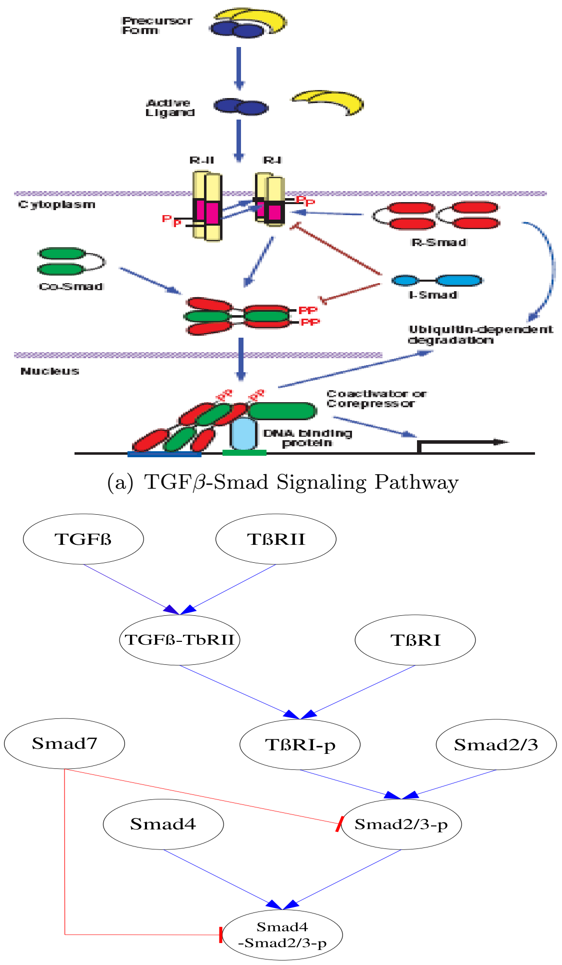
$$\eta_3 \geq (\eta_2 + \eta_1) \quad \eta_1 \geq \eta_0 \quad \eta_2 \geq \eta_0 \tag{4.18}$$

In hypothesis Ω_7, the complex TGFβ-TβRII and the type-I receptor (TβRI) form a phosphorylated type-I receptor. Similar to *ligand-receptor binding*, phosphorylation can be described by *additive synergy* as well, i.e.

$$\mu_3 \geq (\mu_1 + \mu_2) \quad \mu_1 \geq \mu_0 \quad \mu_2 \geq \mu_0 \tag{4.19}$$

In $\Omega_{8,9}$, the inhibitory Smad, Smad7 inhibits the signaling transduction by competing with the receptor-regulate protein Smad2/3 for TβRI. Thus, the probability of Smad2/3 being phosphorylated by the active form of TβRI is decreased with the presence of Smad7. The parameter of this interaction can be represented by a set of second-order constraints. In the first order, the phosphorylation between Smad2/3 and activated TβRI can be described by *additive synergy* regardless to Smad7.

$$\begin{array}{lll} \sigma_3 \geq (\sigma_1 + \sigma_2) & \sigma_1 \geq \sigma_0 & \sigma_2 \geq \sigma_0 \\ \sigma_7 \geq (\sigma_5 + \sigma_6) & \sigma_5 \geq \sigma_4 & \sigma_6 \geq \sigma_4 \end{array} \tag{4.20}$$



(a) TGFβ-Smad Signaling Pathway

(b) Bayesian Model of TGFβ-Smad Signaling Pathway

Figure 4.6: TGFβ-Smad Signaling Pathway

In the second order, the interaction between Smad7 and the first-order phosphorylation can be treated as *mixed joint influence* [13].

$$\sigma_3 \geq \sigma_7 \quad \sigma_1 \geq \sigma_5 \quad \sigma_0 \geq \sigma_4 \quad \sigma_2 \geq \sigma_6 \tag{4.21}$$

Similarly, the parameter of the interaction between Smad7, active Smad2/3 and Smad4, θ can be restrained by the set of second-order constraints as Eq. 4.20 to Eq. 4.21.

Integrating TGFβ-Smad Signaling Pathway into Smad-BCBM Network

Since the *Smad-dependent pathway* in Smad-BCBM network is an aggregate type, we can spatially integrate the network in Figure 4.6(b) into Figure 4.5(a) by replacing the *Smad-dependent pathway* with the TGFβ-Smad signaling pathway which consists a vector of Smad-family proteins($\overline{S}$), i.e. $\overline{S}$=\{$T\beta RII$, $T\beta RI$,

(a) TGF-TbRII

TGFβ	TβRII	η_j=Pr(TGFβ-TβRII|TGFβ,TβRII)
0	0	η_0
0	1	η_1
1	0	η_2
1	1	η_3

(b) TβRI

TGFβ-TβRII	TβRI	μ_j=Pr(TβRI-p|TβRI,TGFβ-TβRII)
0	0	μ_0
0	1	μ_1
1	0	μ_2
1	1	μ_3

(c) Smad23-p

Smad7	TβRI-p	Smad2/3	σ_j=Pr(Smad2/3-p|Smad7,TβRI-p,Smad2/3)
0	0	0	σ_0
0	0	1	σ_1
0	1	0	σ_2
0	1	1	σ_3
1	0	0	σ_4
1	0	1	σ_5
1	1	0	σ_6
1	1	1	σ_7

(d) Smad23-p4

Smad7	Smad4	Smad2/3-p	θ_j=Pr(Smad4-Smad2/3-p|Smad7,Smad4,Smad2/3-p)
0	0	0	θ_0
0	0	1	θ_1
0	1	0	θ_2
0	1	1	θ_3
1	0	0	θ_4
1	0	1	θ_5
1	1	0	θ_6
1	1	1	θ_7

Table 4.3: CPTs for TGFβ-Smad Pathway

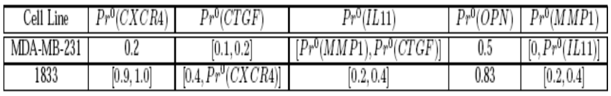
Smad2/3, *Smad4*, *Smad7*, *TGFβ-TβRII*, *TβRI-p*, *Smad2/3-p*, *Smad4-Smad2/3-p*}. The integrated network, TGFβ-Smad BCBM, is shown in Figure 3.3 and the joint probability distribution in Eq. 4.16 can be reformulated as

$$P_{(C,T)} = P_{T,\overline{G}} \textstyle\sum_{\overline{S}} \alpha\theta\sigma\mu\eta \quad P_{(I,T)} = P_{T,\overline{G}} \textstyle\sum_{\overline{S}} \rho\theta\sigma\mu\eta \tag{4.22}$$

where $\overline{G} = \{T\beta RII, T\beta RI, Smad2/3, Smad4, Smad7\}$ and $P_{T,G}$=Pr(TGFβ)Pr($\overline{G}$). Thus, the parameters of the integrated model can be constrained by the consistency of Bayesian model integration as Eq. 4.17.

Bone Metastasis Prediction

In [60], by in-vivo selection of MDA-MB-231 human breast cancer cell line, subpopulations with distinct bone metastatic ability are isolated. Cell line 1833 generates large osteolytic bone lesions while the populations of 1834 exhibit low metastatic activity towards the bone. The different metastasis ability of these populations are due to the various expression signatures of the bone metastasis related genes, i.e. $\overline{G'}$={CXCR4,CTGF,IL11,OPN,MMP1}. The qualitative hypotheses on gene expression profile in [7, 60] can be summarized in Figure 4.7.

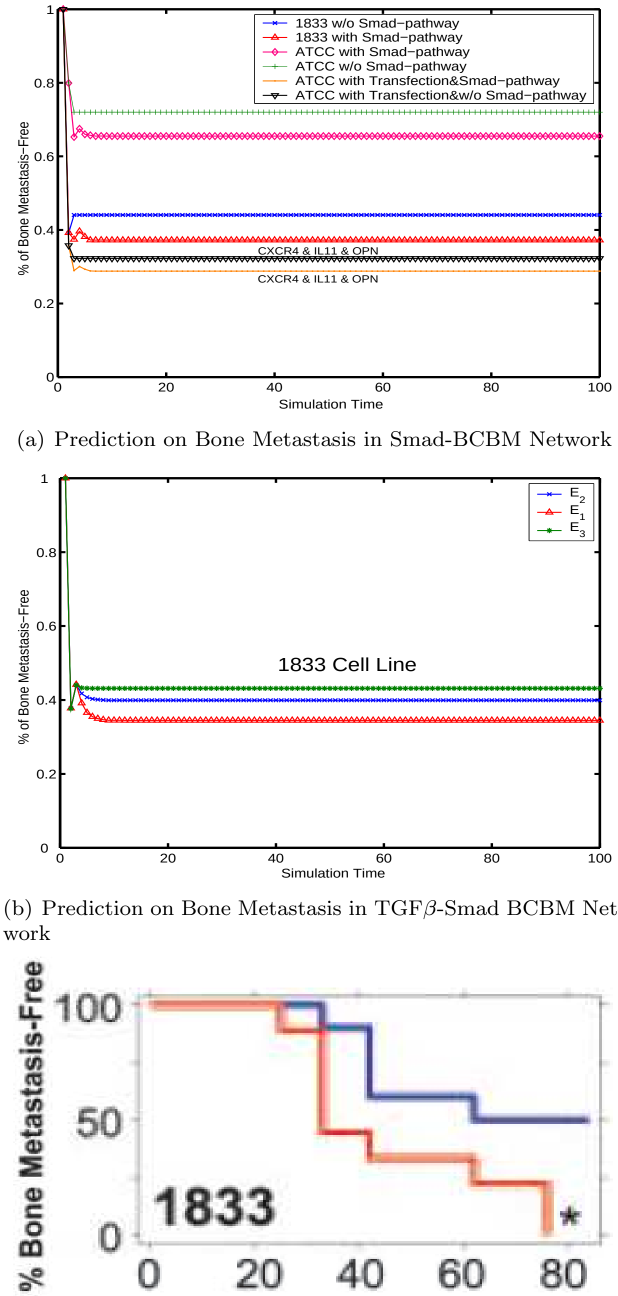


Cell Line	$Pr^0(CXCR4)$	$Pr^0(CTGF)$	$Pr^0(IL11)$	$Pr^0(OPN)$	$Pr^0(MMP1)$
MDA-MB-231	0.2	$[0.1, 0.2]$	$[Pr^0(MMP1), Pr^0(CTGF)]$	0.5	$[0, Pr^0(IL11)]$
1833	$[0.9, 1.0]$	$[0.4, Pr^0(CXCR4)]$	$[0.2, 0.4]$	0.83	$[0.2, 0.4]$

Figure 4.7: Gene Expression Constraints

According to Section 2.3.2, the inference can be performed by calculating the *equivalent mean model* with expect values of the model parameters. The vector of equivalent mean parameters of Smad-BCBM network can be described as $\{\overline{\alpha}, \overline{\beta}, \overline{\gamma}, \overline{\zeta}, \overline{\lambda}, \overline{\phi}, \overline{\rho}\}$ as well as of the integrated TGFβ-Smad BCBM network can be represented as $\{\overline{\alpha}, \overline{\rho}, \overline{\zeta}, \overline{\eta}, \overline{\mu}, \overline{\theta}, \overline{\sigma}\}$. Since these parameters are high-dimensional, the integration in Eq. 2.53 can be approximated by Monte Carlo simulation. For each parameter, we simulate K=50,000 samples constrained by Eq. 4.15 to Eq. 4.22.

We perform the bone metastasis inference in the Smad-BCBM and TGFβ-Smad BCBM network respectively. In the first simulation, we perform inference on the probability of bone metastasis formed by 1833 cells and ATCC cells in the Smad-BCBM network in Figure 4.5(a) . The inference is simulated with interference on *Smad-dependent pathway* by clamping it to 0(minima) and with transfection of ATCC by clamping CXCR4, IL11, OPN to 1(maxima). The inference results are shown in Figure 4.8(a). Secondly, the inference on the probability of bone metastasis are performed in the integrated TGFβ-Smad BCBM network in Figure 3.3. Since the prior probability of the variables in $\overline{G}$ are unknown, we perform the simulation based on three configurations on $\overline{G}$, i.e. $\overline{E} = \{E_1, E_2, E_3\}$. In E_1, the prior probabilities of the entities in $\overline{G}$ except Smad7 are set to 1 and the prior of Smad7 is set to 0; In E_2, the prior probabilities of the entities in $\overline{G}$ except Smad4 and Smad7 are set to 1. The prior of Smad4 and Smad7 are set to 0. In E_3, we set the prior probabilities of entities in $\overline{G}$ except Smad7 to 0 and the prior of Smad7 is set to 1. The simulation results of E_1 and E_3 provide the maximum and minimum boundary on the bone metastasis probability in the TGFβ-Smad BCBM network according to different settings on the variables in $\overline{G}$ and the simulation of E_2 comply with the experiment on Smad4 interference by Kang and are shown in Figure 4.8(b). The biological observation on bone metastasis formed in [59] is shown in Figure 4.8(c) for comparison.



(a) Prediction on Bone Metastasis in Smad-BCBM Network

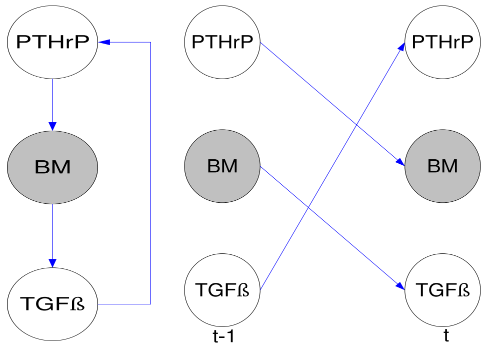
(b) Prediction on Bone Metastasis in TGFβ-Smad BCBM Network

(c) Observation on Bone Metastasis by Kang

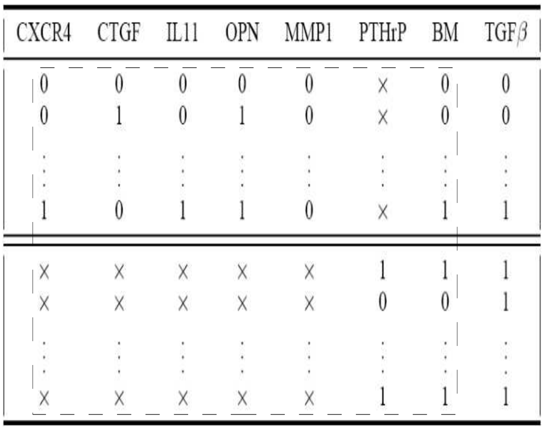
Figure 4.8: Integrated TGFβ-Smad BCBM Network and Prediction

4.3.2 Integration of TGFβ-PTHrP Pathway in Breast Cancer Bone Metastasis Network

In the second scenario, a set of incomplete qualitative hypotheses has been extracted from the publications [43, 60, 113]. In [60], the study reports a set of bone metastasis-related genes and their interaction with TGFβ signaling pathway in breast cancer bone metastasis network. In [43, 113], the causal role of PTHrP in breast cancer bone metastasis and its interaction with TGFβ are studied. We shall fuse the two TGFβ-mediated signaling pathways in the breast cancer bone metastasis network by fitting a integrated network to the global uncertainty and perform quantitative inference in this network.



(a) PTHrP-BCBM (b) 2TBN of PTHrP-BCBM



CXCR4	CTGF	IL11	OPN	MMP1	PTHrP	BM	TGFβ
0	0	0	0	0	x	0	0
0	1	0	1	0	x	0	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	0	1	1	0	x	1	1
x	x	x	x	x	1	1	1
x	x	x	x	x	0	0	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
x	x	x	x	x	1	1	1

(c) Example of Combined Global Data Samples

Figure 4.9: Breast Cancer Bone Metastasis Network with PTHrP

(a) BM given PTHrP

PTHrP	Pr(BM|PTHrP)
0	η_0
1	η_1

(b) TGFb given BM

BM	Pr(TGFβ|BM)
0	ρ_0
1	ρ_1

(c) PTHrP given TGFβ

TGFβ	Pr(PTHrP|TGFβ)
0	ϕ_0
1	ϕ_1

Table 4.4: CPTs for PTHrP-related Bayesian Network

Parathyroid Hormone-related Protein (PTHrP) in Breast Cancer Bone Metastasis Network

In [43], Parathyroid Hormone-related Protein (PTHrP) are proven to have a causal role in human breast cancer bone metastasis by stimulating the osteoclastic bone resorption. Moreover, in [113], it has been identified that TGFβ released during the bone destruction is responsible for enhancing PTHrP production by tumor cells in the bone microenvironment and the enhanced PTHrP in turn further promotes the bone destruction to form a vicious metastastic cycle. A set of qualitative hypotheses on the causal effect of PTHrP on breast cancer bone metastasis and the TGFβ-mediated PTHrP signaling pathway can be summarized in $\overline{\Omega'} = \{\Omega_{12}, \Omega_{13}\}$.

Ω_{12}: *Tumor cells in the bone microenvironment produce PTHrP and stimulate osteoclastic bone resorption, which in turn results in the release of active TGFβ.*

Ω_{13}: *TGFβ then acts on the tumor cells to endow them with metastatic capability and the ability to stimulate production of PTHrP.*

The structure of the PTHrP-TGFβ signaling pathway model in BCBM network is shown in Fig 4.9(a) and it can be unrolled over time as in Fig. 4.9(b). According to [13], the parameters of the PTHrP-TGFβ BCBM network η, ρ and ϕ in Table 4.4 can be constrained by *single positive influence* as

$$\eta_0 \leq \eta_1 \qquad \rho_0 \leq \rho_1 \qquad \phi_0 \leq \phi_1 \tag{4.23}$$

Integrating PTHrP and Related Genes into TGFβ Signaling Pathway in BCBM Network

The full set of qualitative hypotheses can be written as $\overline{\Omega}_F = \{\Omega_1, \Omega_2, \Omega_3, \Omega_{12}, \Omega_{13}\}$ and the complete set of variables includes eight entities at different levels, i.e. $\overline{X}$={CXCR4,CTGF,IL11,OPN,MMP1,PTHrP,BM,TGFβ}. Two classes of TGFβ signaling pathway models of breast cancer bone metastasis network can be constructed based on each set of hypotheses as

$$\begin{aligned} \overline{M}_{h=\overline{\Omega}} &= \{m_{k,h}(s,\zeta,\beta,\gamma,\lambda)|k=1,\ldots,K\} \\ \overline{M}_{h=\overline{\Omega'}} &= \{m_{k,h}(s',\eta,\rho,\phi)|k=1,\ldots,K\} \end{aligned} \tag{4.24}$$

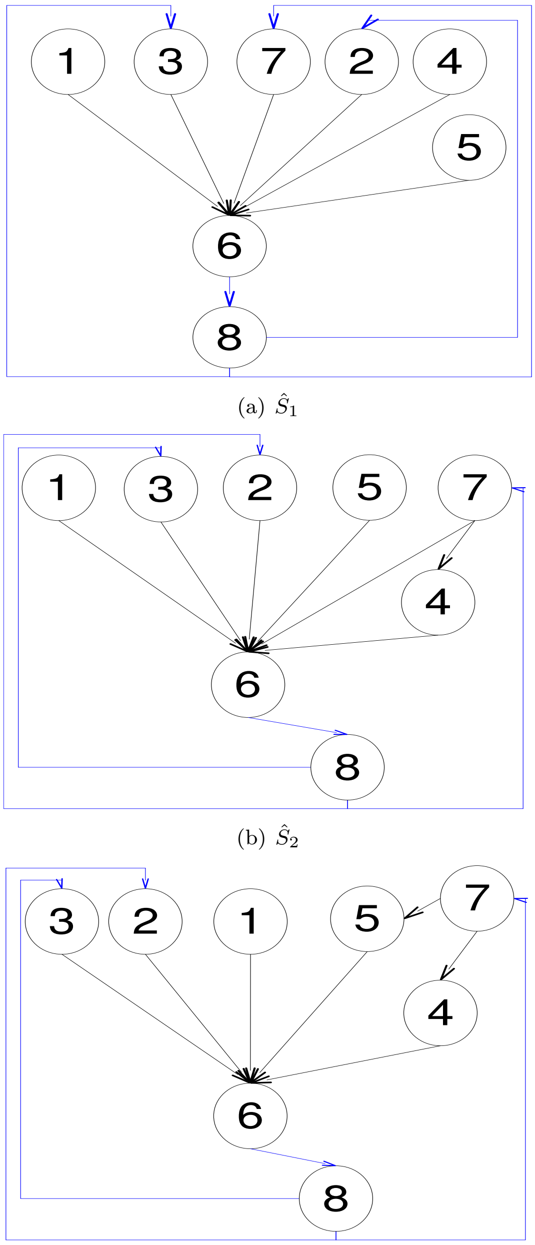
where s and s' can be shown in Fig. 2.21(b) and Fig. 4.9(a). Moreover, $\overline{\Omega}_F$ can be further decomposed into two sub-sets of hypotheses, i.e. a set of consistent hypotheses and a set of complementary hypotheses as

$$\overline{\Omega}_F = \{\overline{\Omega}_{con}(\Omega_2, \Omega_3, \Omega_{13}), \overline{\Omega}_{com}(\Omega_1, \Omega_{12})\} \tag{4.25}$$

The consistent hypotheses in $\overline{\Omega}_{con}$ provide explicit and unique information on the relationship from the parent nodes to the child node, i.e. (BM;TGFβ), (TGFβ;CTGF), (TGFβ;IL11), (TGFβ;PTHrP) and the complementary hypotheses in $\overline{\Omega}_{com}$ describe all possible connections from the parent nodes to the child node, i.e. (PTHrP;BM) and (CXCR4,CTGF,IL11,OPN,MMP1;BM). Therefore, the model integration problem boils down to learn a set of "best" models for the variables in $\overline{\Omega}_{com}$ and then combine the "best" models with the qualitative hypotheses in $\overline{\Omega}_{con}$ to form the full model. Each hypothesis in $\overline{\Omega}_{com}$ can be used to define a model class as

$$\begin{aligned} \overline{M}_{h=\Omega_1} &= \{m_{k,\Omega_1}(s(\overline{G}_1; BM), \zeta) | k = 1, \ldots, K\} \\ \overline{M}_{h=\Omega_{12}} &= \{m_{k,\Omega_{12}}(s(\overline{G}_2; BM), \eta) | k = 1, \ldots, K\} \end{aligned} \tag{4.26}$$

where $\overline{G}_1$={CXCR4, CTGF, IL11, OPN, MMP1} and $\overline{G}_2$={PTHrP}. We start Bayesian learning by generating training data from each model class in Eq. 4.24. We adopt two schemes for data generation: i)We sampled 5000 training sequences from the *equivalent mean model* as Eq. 2.53 from each model class; ii)We firstly sampled K=500 ground Dynamic Bayesian models for each model class in Eq. 4.24 and generate 10 data sequences from each ground Bayesian model. Then the complete training data over $\overline{X}$ is formulated by mixing the training sequences from the two schemes with hidden variables from each dataset, i.e. $\overline{D}_F = \{D_{\overline{\Omega}}, D_{\overline{\Omega'}}\}$ as shown in Fig. 4.9(c). Since we are only interested in learning a structure on the variables in $\overline{\Omega}_{com}$, we collect the data in $\overline{D}_F$ corresponding to those variables in $\overline{G}_1$ and $\overline{G}_2$ to finalize the training data $\overline{D}_{com}$ with size $|\overline{D}_{com}|$=20,000. We applied the SEM algorithm with BD score to learn a set of models based on various starting graphs. The learned model structures represent the combined uncertainty of the set of complementary hypotheses $\overline{\Omega}_{com}$. Since we have some prior knowledge on the relationship between the molecules in $\overline{G}_1$ and $\overline{G}_2$ to some extent, thus we could set structural prior distribution to zero to ignore those structures which violate the knowledge, i.e. i)We shall not allow edges from *BM* to any of the molecules; ii)We shall assume that the molecules *PTHrP*, *CTGF* and *IL11* have no other parents than *TGF*β; iii) We shall not allow any interaction between the molecules in $\overline{G}_1$. Then, we incorporate this structural constraint into the learning. The learned model structures and the structures described in the consistent hypotheses, $\overline{\Omega}_{con}$, form a set of complete models as shown in Fig. 4.10. The edges of the learned structure are in black color and the edges of the structures in $\overline{\Omega}_{con}$ are in blue color. The node is indexed by the order in the set {CXCR4, CTGF, IL11, OPN, MMP1, BM, PTHrP, TGFβ}. Meanwhile, the complete vector of parameters is composed by the MAP parameter estimation for the learned structure and {β, γ, λ, ρ, ϕ} which can be defined by Eq. 4.23 and sampled by Monte Carlo method [12]. Thus, the quantitative predictions based on full Bayesian approach can be performed by averaging the inference from each of the complete model which can be calculated by Eq. 2.53.



(a) $\hat{S}_1$

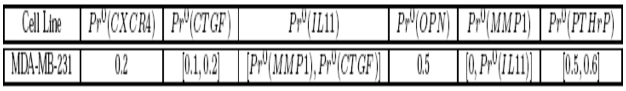
(b) $\hat{S}_2$

(c) $\hat{S}_3$

Figure 4.10: PTHrP-BCBM Model Candidates by SEM Learning

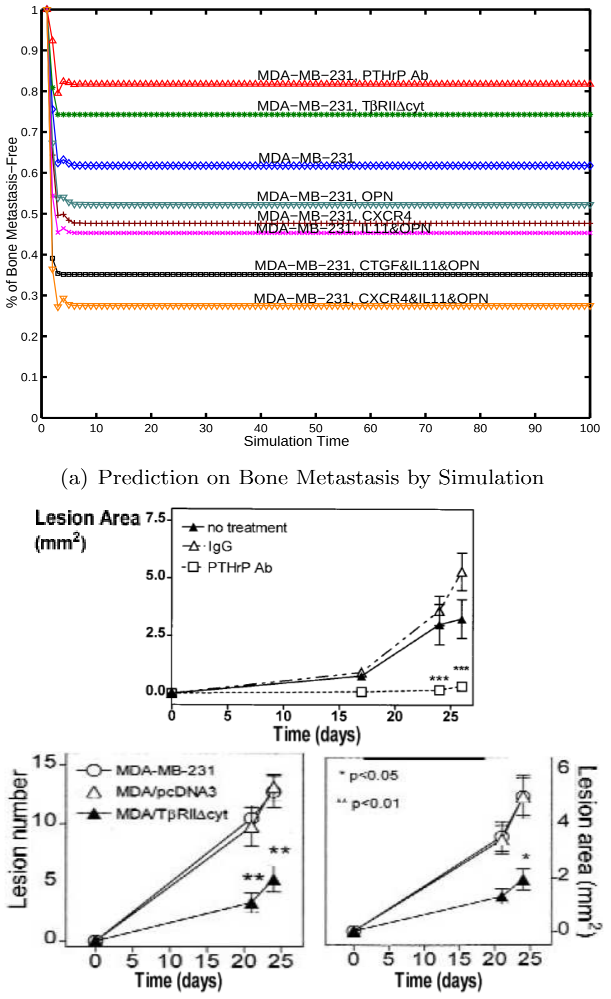
Bone Metastasis Prediction

In [43, 60, 113], the MDA-MB-231 human breast cancer cell line are used in the investigation of breast cancer bone metastasis. The metastasis ability of the population are due to the expression levels of the bone metastasis related molecules in $\overline{G}$=$\{\overline{G}_1$,$\overline{G}_2\}$. The qualitative hypotheses on the expression profile of $\overline{G}$ can be extracted from [7, 60, 113] and summarized in Fig. 4.11.



Cell Line	$Pr^U(CXCR4)$	$Pr^U(CTGF)$	$Pr^U(IL11)$	$Pr^U(OPN)$	$Pr^U(MMP1)$	$Pr^U(PTHrP)$
MDA-MB-231	0.2	[0.1, 0.2]	$[Pr^U(MMP1), Pr^U(CTGF)]$	0.5	$[0, Pr^U(IL11)]$	[0.5, 0.6]

Figure 4.11: Molecular Expression Constraints



(a) Prediction on Bone Metastasis by Simulation

(b) Prediction on Bone Metastasis by PTHrP and TGFβ Intervention

Figure 4.12: Integrated TGFβ-Smad BCBM Network and Prediction

We perform eight in-silico transfection experiments to predict the probability of bone metastasis in which single or combination of molecules in $\overline{G}$ are manually transfected to be overexpressed or underexpressed. Given the structure of each complete model in Fig. 4.10, the averaged prediction on bone metastasis can be calculated by Eq. 2.53 iteratively over time. The transfected molecule(s) in the experiments are treated as evidence of observed node(s) whose prior probabilities of being overexpressed are clamped to 1 (maxima) and of being underexpressed are clamped to 0 (minima). The mean estimate of bone metastasis activity evolves over time in the case of transfected ATCC cells are shown in Fig. 4.12(a). The time course of each simulation is labeled with the transfectant molecule names in the figure. Figure 4.12(b) show the experiment observations in [43, 60, 113].

Comparison between the simulation results and the experiment observations from Kang in Fig. 2.24 is straightforward since in both cases the bone metastasis probability is measured. However, the experiment observations in Fig. 4.12(b) and Fig. 2.24 are measured in bone lesion area/number than the probability, we could translate these results into probability by specifying the maximum probability of the bone metastasis in each figure. For example, in the top-central graph of Fig. 4.12(b), the bone lesion area (in mm^2) formulated by MDA-MB-231 cells with and without PTHrP are measured. If we assume the total bone area under concern is B_{total}, then it is obvious to infer from graph that $B_{total}{\geq}5.0$ and at 25th day, the lesion area formed by MDA-MB-231 cell without PTHrP is: $\mathrm{B}_P{\leq}1.0$, therefore, we could calculate the probability of bone metastasis in this case as $\frac{B_P}{B_{total}} \leq 0.2$. Similarly, the probability of bone metastasis formed by the MDA-MB-231 cells with mutated TGFβ receptor, TβRIIΔcyt, from the bottom graphs in Fig. 4.12(b) can be calculated as $\frac{B_T}{B_{total}} \leq 0.2$. Thus, we can conclude that our in-silico simulation produces reasonable quantitative predictions on the probability of bone metastasis in these experiments.

4.4 Summary

In this chapter, we investigate the methods to integrate the incomplete qualitative knowledge into our probabilistic modeling framework. Knowledge are often incomplete representation of an interested domain due to their spatial and temporal properties. For example, one knowledge component may only describe a local sub-structure of a domain which can be compensated by another knowledge component describing a different local sub-structure of the same domain with distinct set of domain variables. Also, even at the same location of a domain, new discoveries with a number of newly identified variables and/or connections might be used to update the existing knowledge and the associated Bayesian networks at this location. Thus, the incomplete knowledge integration problem can be transformed eventually to the problem of knowledge-based Bayesian network fusion.

We solve this problem in two scenarios. Firstly, we assume that the structure space of the fused Bayesian network is explicitly known. In this case, the integration problem boils down to modeling the integrated parameter space uncertainty. If a single node is inserted into one existing Bayesian network and if the d-separation properties changed in the network during the integration process, equality constraints on the (marginal) joint probability over the variables in the existing network are used to restrain the integrated model parameter space. If multiple nodes are inserted, the integration process can be decomposed into a series of steps with single node. As the above, the equality constraints on the parameter space can be imposed in case the d-separation criterion is changed. Especially, this criterion is symmetric to all Bayesian network components which are being fused together. In the second scenario, we need to model the uncertainty of the integrated network structure space besides the parameter space. We use Structural EM (SEM) algorithm to learn the integrated model structure from the artificial generated local and global statistics with missing values at some positions. We perform the SEM learning process several times with distinct initial graphs aiming to explore all the possible ”good” structures to explain the statistics. Quantitative probability configurations are learned associated with each structure candidate. Quantitative reasoning and predictions are calculated as an average of the quantity over all network structures.

Chapter 5

Discussion and Future Research

5.1 Discussion

The rapid growth of information in every scientific and industrial domain raises exciting challenge in handling vast amount of data and modeling underpinnings of a domain in a systematic and mathematic manner. In recent years, probabilistic network has become popular as practical representations of knowledge for reasoning under uncertainty. The probabilistic network computational framework uses a graphical model to capture random variables in a domain and relations between them, with probabilities that represent the uncertainties in the domain. The framework offers powerful algorithms of quantitative reasoning, such as predictive inference and diagnostic reasoning. Among these and other probabilistic graphical models, directed graphical models (also called Bayesian Networks or Belief Networks) are particular attractive for researches with the Artificial Intelligence (AI) and statistics communities.

The major data mining practice in inferring Bayesian network from the data, i.e. reverse-engineering approach is concentrated on inferring the structure and its associated parameters of a Bayesian network from data. It is known as structure learning and parameter learning of graphical models in machine learning. Structure learning of Bayesian network use the likelihood score, such as BIC score to find the "best" model fitting to the data or a second type score called Bayesian score, e.g. BD score, which incorporates prior belief on the model structure and parameter space in Bayesian network learning. These and other practices of computational modeling with Bayesian network, especially, the structure and parameter learning with Bayesian network, reveal a number of built-in problems of the data driven reverse-engineering approach. There are mainly three concerns with these approaches:

1. **Overfitting** Learning from sparse data might induce overfitting since the sparse data hardly provide sufficient entire "statistics" of an underlying system. Full Bayesian approach with Bayesian model averaging (BMA) can be used to avoid and/or alleviate overfitting by computing a posterior probability distribution over all possible models to reflect the true model uncertainty. The quantity of interest are calculated as an average under

each of the models. Prior knowledge over structural and parameter space is imperative in computation of posterior probability.

2. **Computational Complexity** When learning structure, with the number of variables increases the space of possible graph structures grows superexponentially. In case of high-dimensional data set, the learning problem become NP-hard. Heuristic search strategies, such as greedy hill-climbing, simulate annealing and MCMC algorithm are often used to search all possible models through the structure space. The computational complexity of these evaluations become inextricable when we learn from high-dimensional data.

3. **Multi-scale Integrative Learning** As more sources of data have become available, multi-scale integrative learning becomes imperative to raise the challenges to address fundamental understandings of a system as a whole by automatic integrating both homogeneous and heterogeneous types of data. Uniform, standardized data representations are seldomly adopted, which complicates cross-experiment comparisons as well as data quality, context and cross-lab variations represent another important hurdle. Statistical tests are employed in homogeneous data integration which combine single-level measurements from different platforms. However, prior knowledge are indispensable to automatically integrate heterogeneous data.

These and other discussions on the built-in problems and their solutions in data driven reverse-engineering approaches with Bayesian networks have invariably revealed one fact, i.e. the remarkable importance of prior information in reverse-engineering methods to prevent overfitting by providing structure and parameter prior distributions and to optimize computational complexity by reducing the heuristic search space, as well as to better recover the underlying network of a system by integrating homo- and heterogeneous multiple-scale and multiple-origin data sets.

5.2 Contribution

In this thesis, we have proposed unprecedented solutions to the challenges in Bayesian network learning, namely, how to construct prior distribution over structure and parameter space from prevalent amount of pre-existing qualitative information in science and industrial domain within an unified framework as well as to the tough question how qualitative statements about relationship between domain entities can be transformed to yield quantitative predictive models, able to perform probabilistic inference and reasoning.

In Chapter 2, we formally proposed the knowledge-driven probabilistic networks modeling framework which utilizes solely qualitative prior information to perform probabilistic network modeling and quantitative inference and reasoning. No quantitative data is available in our study to shield our insights in the function and effects of qualitative prior knowledge on quantitative modeling. We formulated the qualitative knowledge model and use it to capture the structural dependence and influence between the variables under concern and translate the cause-effect relationship described by the qualitative statements into a set of inequality constraints over the model parameter space. The structural and parameter constraints eventually forms the model prior distribution.

In this way, we can build a class of (Dynamic) Bayesian networks which are consistent with the body of the qualitative prior knowledge.

We employ full Bayesian approach to calculate the average quantity of interest, e.g. probabilistic inference and reasoning, over the class of Bayesian networks. The integral over model structure space and parameter space can be computed by Monte Carlo integration technique with Accept-Reject algorithm which induce computational complexity for high-dimensional model space. We proposed an simple and efficient method to approximate the averaged probabilistic inference and prediction by the mean model. We have analyzed the bias as a function of the message-propagation steps n and of the constraints over the parameter space. We concluded that for small n ($n > 1$) and relative restrictive constraints (relative small parameter sub-space), our approximation scheme is usually reasonable. For n=1, the approximation is exact. Also, we have studied the robustness of our approach for each possible knowledge feature in our knowledge model due to the noisy information in the statements, we studied the effect of such noise on our constructed model uncertainty and its consequent generalization accuracy.

In chapter 3, we extend the knowledge-driven probabilistic network modeling framework to a set of inconsistent knowledge. We investigate the method to reconcile the contradicting qualitative information and utilize these inconsistent information to make coherent quantitative reasoning. We do so by transforming the qualitative knowledge model into a hierarchical model in which knowledge features are encoded and quantified by their mutual (in)dependences and the associated conditional probability. Given the hierarchical knowledge model, a set of inconsistent knowledge are dissected and the conditional probabilities of the knowledge features are calculated by evaluating their statistics. Expert belief can be integrated into the method as a prior belief on the inconsistent knowledge components. Each knowledge component uniquely define a class of constrained Bayesian networks. The conditional probability of knowledge features are used to compute the knowledge prior, i.e. the joint probability over the feature space. Thus, multiple classes of Bayesian networks are inferred from the inconsistent knowledge which are weighted by their corresponding knowledge prior distribution. The incoherent Bayesian network classes are reconciled in this way consistently into uniform representation and the averaged quantitative prediction can be calculated over all ground models in one class and over all classes.

In chpater 4, we investigate the methods to integrate the incomplete qualitative knowledge into our probabilistic modeling framework. Knowledge are often incomplete representation of an interested domain due to their spatial and temporal properties. For example, one knowledge component may only describe a local sub-structure of a domain which can be compensated by another knowledge component describing a different local sub-structure of the same domain with distinct set of domain variables. Also, even at the same location of a domain, new discoveries with a number of newly identified variables and/or connections might be used to update the existing knowledge and the associated Bayesian networks at this location. Thus, the incomplete knowledge integration problem can be transformed eventually to the problem of knowledge-based Bayesian network fusion.

We solve this problem in two scenarios. Firstly, we assume that the structure space of the fused Bayesian network is explicitly known. In this case, the integration problem boils down to modeling the integrated parameter space uncertainty. If a single node is inserted into one existing Bayesian network and if the d-separation properties changed in the network during the integration process, equality constraints on the (marginal) joint probability over the variables in the existing network are used to restrain the integrated model parameter space. If multiple nodes are inserted, the integration process can be decomposed into a series of steps with single node. As the above, the equality constraints on the parameter space can be imposed in case the d-separation criterion is changed. Especially, this criterion is symmetric to all Bayesian network components which are being fused together. In the second scenario, we need to model the uncertainty of the integrated network structure space besides the parameter space. We use Structural EM (SEM) algorithm to learn the integrated model structure from the artificial generated local and global statistics with missing values at some positions. We perform the SEM learning process several times with distinct initial graphs aiming to explore all the possible "good" structures to explain the statistics. Quantitative probability configurations are learned associated with each structure candidate. Quantitative reasoning and predictions are calculated as an average of the quantity over all network structures.

In summary, our solutions to the tough question on how to make use of the qualitative information to yield quantitative predictive models for performing probabilistic predictions and reasoning will form an important link between usually validated but qualitative information and quantitative yet uncertain information derived from the data.

5.3 Future Researches

Further improvements on our methods could be

1. For high-dimensional parameter space and extreme restrictive constraints, it is computationally expensive to use Accept-Reject algorithm to sample in the model space. Therefore, more efficient sampling techniques can be used here to improve the efficiency of the algorithm, such as Monte Carlo Markov Chain (MCMC) algorithm or Gibbs sampling algorithm.

2. We could transform the uncertainty in the model space defined by the constraints derived from the qualitative knowledge model to a parametric presentation with a well-known probability density function, e.g. a Gaussian distribution function. In this way, we can build an interface between the inequality constraints derived from the qualitative information and the density function which supports parametric computations.

Also, it is very interesting and very important for future investigation of our methods on how to apply our methods to improve the Bayesian network learning from quantitative sparse data by avoiding overfitting, reducing computational complexity and enabling multi-scale integrative inference. One possible solution is to incorporate the prior uncertainty over Bayesian model space from the qualitative information to learning algorithms.

Appendix A

Bayesian Dirichlet Equivalent Score

A score $S(G)$ is assigned to the graph G to assess the fittness of a network G to the data set D. This score is given by the posterior probability as

$$S(G) = \frac{P(D|G)P(G)}{P(D)} \tag{A.1}$$

where $P(D|G)$ is the marginal data likelihood, $P(G)$ is the prior probability of structure G and $P(D)$ a normalizing constant. The marginal data likelihood can be calculated as

$$P(D|G) = \int_{\Theta} P(D|\Theta, G, \xi)P(\Theta|G, \xi)d\Theta \tag{A.2}$$

where Θ is the set of parameters and ξ indicates the prior background information. We assume that the data set D consists of N independent data samples d^l, then the data likelihood can be decomposed as

$$P(D|G) = \prod_{l=1}^{N} P(d^l|\Theta, G, \xi) \tag{A.3}$$

Thus, Eq. A.2 can be written as

$$P(D|G) = \prod_{l=1}^{N} \int_{\Theta} P(d^l|\Theta, G, \xi)P(\Theta|G, \xi)d\Theta \tag{A.4}$$

To solve the Eq. A.4 in closed form, five assumptions are made [21].

Assumption 1 Multinomial Distribution Let d_i^l and $d_{pa_i}^l$ denote the variable X_i and the parent set Pa_i in the *l-th* case of data set D, Then,

$$P(d^l = k|d_{pa_i}^l = j, \Theta, G, \xi) = \theta_{ijk}, \;\in [0, 1], \;\forall X_i, Pa_i \tag{A.5}$$

Assumption 2 Parameter Independence Given network structure G, the parameters associated with each variable are independent from each other such that $P(\Theta|G,\xi)$ decomposes into

$$P(\Theta|G,\xi) = \prod_{i=1}^{n} P(\Theta_i|G,\xi) \ \ \forall i = 1,\ldots,n \tag{A.6}$$

Since each instance of parents of a variable X_i are independent. $P(\Theta_i|G,\xi)$ decomposes into

$$P(\Theta_i|G,\xi) = \prod_{j=1}^{q_i} P(\Theta_{ij}|G,\xi) \ \ \forall i = 1,\ldots,n; j = 1,\ldots,q_i \tag{A.7}$$

where q_i is the number of configurations the set of parents pa_i can take.

Assumption 3 Parameter Modularity Given two network structures G_1 and G_2, if X_i has the same parents in G_1 and G_2, then

$$P(\Theta_{ij}|G_1,\xi) = P(\Theta_{ij}|G_2,\xi), \ \ \forall j = 1,\ldots,q_i \tag{A.8}$$

Assumption 4 Dirichlet Prior Given a network structure G, $P(\Theta_{ij}|G,\xi)$ is a priori Dirichlet distributed, $\theta_{ij} \sim D(N_{ij1},\ldots,N_{ijr_i})$, exist exponents N'_{ijk}, which depend on

$$P(\Theta_{ij}|G,\xi) = \frac{\Gamma(\sum_{k=1}^{r_i} N'_{ijk})}{\prod_{k=1}^{r_i}\Gamma(N'_{ijk})} \prod_k \theta_{ijk}^{N'_{ijk}-1} \tag{A.9}$$

where $\Gamma(*)$ denotes the Gamma function and r_i is the number of values of variable X_i. The hyper-parameters N'_{ijk} can be computed as

$$N'_{ijk} = N'P(X_i = k, Pa_i = j|\xi) \tag{A.10}$$

where N' is the equivalent sample size and $P(X_i = k, Pa_i = j|\xi)$ is the prior joint probability distribution over the variable X_i and its parents Pa_i.

Assumption 5 Complete Data The data set is complete. That is, D contains no missing values or hidden variables. From the multinomial sample assumption in Eq. A.5 and the assumption of complete data, $P(D|\Theta,G)$ can be factorized into

$$P(D|\Theta,G,\xi) = \prod_{l=1}^{N}\prod_{i=1}^{n} P(d_i^l = k|d_{Pa_i}^l = j,\Theta,G,\xi) = \prod_{i=1}^{n}\prod_{j=1}^{q_i}\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \tag{A.11}$$

where N_{ijk} equals to the number of the samples $(X_i = k, Pa_i = j)$ in D. Substituting Eq. A.9 and Eq. A.11 the marginal likelihood in Eq. A.2 can be re-formulated as

$$P(D|G) \ = \ \int_{\Theta} P(D|\Theta,G,\xi)P(\Theta|G,\xi)d\Theta$$

$$
\begin{aligned}
&= \int_{\Theta} \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}} \frac{\Gamma(\sum_{k=1}^{r_i} N'_{ijk})}{\prod_{k=1}^{r_i} \Gamma(N'_{ijk})} \prod_{k} \theta_{ijk}^{N'_{ijk}-1} \\
&= \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(\sum_{k=1}^{r_i} N'_{ijk})}{\prod_{k=1}^{r_i} \Gamma(N'_{ijk})} \int_{\Theta} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+N'_{ijk}-1} d\Theta \qquad \text{(A.12)}
\end{aligned}
$$

The posterior of each parameter remains in the conjugate family since the Dirichlet distribution is conjugate for this domain. The integral equals to

$$
\int_{\Theta} \prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}+N'_{ijk}-1} d\Theta = \frac{\prod_{k=1}^{r_i} \Gamma(N_{ijk} + N'_{ijk})}{\Gamma(N_{ij} + N'_{ij})} \qquad \text{(A.13)}
$$

where $N_{ij}{=}\sum_k N_{ijk}$ and $N'_{ij}{=}\sum_k N'_{ijk}$. Thus, the marginal data likelihood reads

$$
P(D|G,\xi) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \frac{\Gamma(N'_{ij})}{\Gamma(N_{ij} + N'_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(N_{ijk} + N'_{ijk})}{\Gamma(N'_{ijk})} \qquad \text{(A.14)}
$$

Appendix B

Textual Statements on Breast Cancer Bone Metastasis Network

In this appendix, we list the extracted qualitative textual statements and information for the breast cancer bone metastasis molecular interaction network(BCBM) in [60].

1. These genes act cooperatively to cause osteolytic metastasis, and most of them encode secreted and cell surface proteins.

2. Two of these genes, interleukin-11 and CTGF, encode osteolytic and angiogenic factors whose expression is further increased by the prometastatic cytokine TGFβ.

3. Most of the genes in this group that are overexpressed by more than 4-fold encode cell membrane or secretory products that may affect the host environment to favor metastasis (Figure3C). They include the bone-homing chemokine receptor CXCR4 pressed (Figure 3B).(Muller et al., 2001; Taichman et al., 2002); the angiogenesis factors fibroblast growth factor-5 (Giordano et al., 1996) and connective tissue-derived growth factor (Moussad and Brigstock,2000); the activator of osteoclast differentiation interleukin-11 (IL11, Manolagas, 1995); the matrix metalloproteinase collagenase MMP1, which promotes osteolysis by cleaving a specific peptide bond in the collagen of bone matrix (Egeblad and Werb, 2002; Holliday et al., 1997; Zhao et al., 1999); follistatin, which binds activin blocking its growth inhibitory effects (de Winter et al., 1996); the metalloproteinase-disintegrin family member ADAMTS1 (Kuno et al., 1999); and proteoglycan-1 (Timar et al., 2002).

4. A functionally diverse set of genes cooperatively promote bone metastasis.

5. IL11 is a potent inducer of osteoclast formation from progenitor cells in the bone marrow (Manolagas, 1995). Osteoclasts are direct mediators of bone resorption in osteolytic bone metastases (Boyce et al., 1999; Mundy, 2002).

6. Osteopontin (OPN) is consistently overexpressed in highly metastatic cells. OPN is a secretory protein with multiple functions, including the ability to stimulate osteoclast adhesion to bone matrix (Asou et al., 2001; Denhardt et al.,2001). OPN has been implicated in cancer aggressiveness metastasis to various organs (Furger et al., 2001; Hotte et al.,2002; Reinholz et al., 2002; Weber, 2001). As IL11 and OPN data suggest that overexpression of MMP1 alone or in combination play distinct roles in enhancing osteoclast function, we tested whether they could collaborate in promoting osteolytic metastasis. Indeed, the combined overexpression of IL11 and OPN in parental MDA-MB-231 cells significantly augmented the incidence of bone metastasis (Figure 4B).

7. When overexpressed alone in parental MDA-MB-231 cells, CXCR4 caused a limited but significant increase in bone metastasis formation, whereas CTGF did not (Figure 4D). However, triple transfectants overexpressing IL11, OPN, and either CXCR4 or CTGF (Figure 4C) showed a dramatic increase both in the rate and in the incidence of bone metastases (Figure 4E).

8. Preliminary data suggest that overexpression of MMP1 alone or in combination with IL11 and OPN also enhances bone metastasis. Thus, the combined activities of these genes specifically promote the growth of osteolytic bone metastases.

9. TGFβ activates bone metastasis genes IL11 and CTGF.

Bibliography

[1] Kyu-Baek Hwang ans Sek Won Kong, Steve A Greenberg, and Peter J. Park. Combining gene expression data from different generations of oligonucleotide arrays. *BMC Bioinformatics*, 2004.

[2] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 1999.

[3] Y. Bengio and P. Frasconi. Diffusion of context and credit information in markovian models. *J. of AI Research*, 1995.

[4] Christopher M. Bishop. *Neural Networks for Pattern Recognition.* Oxford University Press, Inc., New York, NY, USA, 1995.

[5] Wray L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994.

[6] C. Elizabeth Caldon, Roger J. Daly, Robert L. Sutherland, and Elizabeth A. Musgrove. *Journal of Cellular Biochemistry*, 2006.

[7] Gabriele Carlinfante, Daphne Vassiliou, Olle Svensson, Mikael Wendel, Dick Heinegrd, and Gran Andersson. Differential expression of osteopontin and bone sialoprotein in bone metastasis of breast and prostate carcinoma. *Clinical and Experimental Metastasis*, 2003.

[8] Rui Chang. Consistent modeling, integration and simulation of molecular interaction networks in space-time dimension. In *Proceedings of IEEE 7th International Symposium on Bioinformatics & Bioengineering (BIBE)*, 2007.

[9] Rui Chang and Wilfried Brauer. Hierarchical qualitative knowledge integration for quantitative bayesian inference. In *Proceedings of 2007 International Conference on Intelligent System and Knowledge Engineering(ISKE)*, 2007.

[10] Rui Chang and Wilfried Brauer. A novel computational framework towards multi-scale molecular interaction networks fusion based on artificial data and knowledge. In *The 12th Annual International Conference on Research in Computational Molecular Biology (RECOMB)*, 2008. In Processing.

[11] Rui Chang and Wilfried Brauer. A novel knowledge-driven system biology approach in quantifying tgf-mediated cytostasis in breast cancer. In

The 12th Annual International Conference on Research in Computational Molecular Biology(RECOMB), 2008. In Processing.

[12] Rui Chang and Martin Stetter. A knowledge-based dynamic bayesian framework towards molecular network modeling and quantitative prediction. In *Proceedings of 2007 International Conference on Bioinformatics & Computational Biology (BIOCOMP)*, 2007.

[13] Rui Chang and Martin Stetter. Quantitative bayesian inference by qualitative knowledge modeling. In *Proceedings of IEEE 20th International Joint Conference on Neural Networks (IJCNN)*, 2007.

[14] Rui Chang, Martin Stetter, and Wilfried Brauer. Modeling semantics of inconsistent qualitative knowledge for quantitative bayesian network inference. *Neural Networks*, 2007. Accepted for Revision.

[15] Rui Chang, Martin Stetter, and Wilfried Brauer. Quantitative inference by qualitative semantic knowledge mining with bayesian model averaging. *IEEE Transactions on Knowledge and Data Engineering*, 2007. Accepted for Revision.

[16] Chang-Rung Chen, Yibin Kang, Peter M. Siegel, and Joan Massague. *PNAS*, 2000.

[17] Chang-Rung Chen, Yibin Kang, Peter M. Siegel, and Joan Massague. *Cell*, 2002.

[18] D. Chickering. Learning bayesian networks is np-complete. In *Proceedings of AI and Statistics*, pages 85–96, 1995.

[19] David Maxwell Chickering Christopher Meek and David Heckerman. Autoregressive tree models for time-series analysis. In *Proceedings of 2006 Proceedings of the Second International SIAM Conference on Data Mining*, 2002.

[20] G. F. Cooper and E. Herskovits. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 09(4):309–347, October 1992.

[21] Gregory F. Cooper and Edward Herskovits. A bayesian method for constructing bayesian belief networks from databases. In *Proceedings of the seventh conference (1991) on Uncertainty in artificial intelligence*, pages 86–94, San Francisco, CA, USA, 1991. Morgan Kaufmann Publishers Inc.

[22] Michael B. Datto, Yan Li, Joanne Panus, David J. Howe, Yue Xiong, and Xiao-Fan Wang. *PNAS*, 1995.

[23] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Artificial Intelligence*, 1989.

[24] M. Dejori and M. Stetter. Identifying interventional and pathogenic mechanisms by generative inverse modeling of gene expression profiles. *J. Comput. Biology*, 11:1135–1148, 2004.

[25] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(Series B):1–38, 1977.

[26] Joseph L. DeRisi, Vishwanath R. Iyer, and Patrick O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 1997.

[27] D. Draper. Assessment and propagation of model uncertainty. *Journal of the Royal Statistic Society*, 1995.

[28] M. Druzdzel and M. Henrion. Belief propagation in qualitative probabilistic networks, 1993.

[29] Marek J. Druzdzel and Linda C. van der Gaag. Elicitation of probabilities for belief networks: Combining qualitative and quantitative information. In *Proceedings of Eleventh Conference on Uncertainty in Artificial Intelligence*, 1995.

[30] M.J. Druzdzel and L.C. van der Gaag. Building probabilistic networks: Where do the numbers come from? In *IEEE Transactions on Knowledge and Data Engineering*, 2000.

[31] D. J. Dudgeon and M. Lertzman. Dyspnea in the advanced cancer patient. *J. of Pain and Symptom Management*, 16(4):212–219, 1998.

[32] Hinton G. E., Sejnowski T. J., and Ackley D. H. Boltzman machines: Constraint satisfaction networks that learn. Technical report, Carnegie-Mellon University, 1984.

[33] P. Edmonds, S. Karlsen, S. Khan, and J. Addington-Hall. A comparison of the palliative care needs of patients dying from chronic respiratory diseases and lung cancer. *Palliative Medicine*, 15(4):287–295, 2001.

[34] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 1979.

[35] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci.*, 1998.

[36] N. Friedman, M. Linial, I. Nachman, and D. Peer. sing bayesian network to analyze expression data. *Journal of Computational Biology*, 2000.

[37] Nir Friedman. The Bayesian structural EM algorithm. In *UAI*, pages 129–138.

[38] Nir Friedman, Moises Goldszmidt, and Abraham Wyner. Data analysis with bayesian networks: A bootstrap approach. pages 196–205.

[39] Nir Friedman, Kevin Murphy, and Stuart Russell. Learning the structure of dynamic probabilistic networks. In *UAI*, pages 139–147, 1998.

[40] Giampaolo Gavelli and Emanuela Giampalma. Sensitivity and specificity of chest x-ray screening for lung cancer. *Cancer*, 89,S11:2453–2456, 1998.

[41] S. Geman and D. Geman. Stochastic relaxation, gibbs distribution and bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1984.

[42] Schwarz Gideon. *The Annals of Statistics*, (2):461–464.

[43] Theresa A. Guise, Juan Juan Yin, Suzanne D. Taylor, Yoshinari Kumagai, Mark Dallas, Brendan F. Boyce, Toshiyuki Yoneda, , and Gregory R. Mundy. Evidence for a causal role of parathyroid hormone-related protein in the pathogenesis of human breast cancer-mediated osteolysis. *Journal of Clinic Investigation*, 98, October 1996.

[44] Elias Gyftodimos and Peter Flach. Hierarchical bayesian networks: A probabilistic reasoning model for structured domains. In *Proceedings of the ICML-2002 Workshop on Development of Representations*, pages 23–30, 2002.

[45] J. Hamilton. *Time Series Analysis.* Wiley, 1994.

[46] Gregory J. Hannon and David Beach. p15ink4b is a potential effector of tgf-β-induced cell cycle arrest. *Nature*, 1994.

[47] Alexander. J. Hartemink, David K. Gifford, Tommi S. Jaakkola, and Richard A. Young. Using graphical models and genomic expression data to statistically validate models of genetic regulatory networks. *Pacific Symposium on Biocomputing*, 6:422–433, 2001.

[48] David Heckerman. A Tutorial on Learning with Bayesian Networks. Technical report, Microsoft Research, 1996.

[49] David Heckerman, Dan Geiger, and David Maxwell Chickering. Learning bayesian networks: The combination of knowledge and statistical data. In *KDD Workshop*, pages 85–96, 1994.

[50] Jennifer A. Hoeting, David Madigan, Adrian E. Raftery, and Chris T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 1999.

[51] Weimin Hu, Clifford J. Bellone, and Joseph J. Baldassare. *J. of Biological Chemistry*, 1999.

[52] Curtis Huttenhower and Olga G. Troyanskaya. Bayesian data integration: A functional perspective. *Computational Systems Bioinformatics*, 2006.

[53] Jordan M. I. *Learning in Graphical Models.* MIT Press, 1998.

[54] Seiya Imoto, Tomoyuki Higuchi, Takao Goto, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Combining microarrays and biological knowledge for estimating gene networks via bayesian networks. *Computational Systems Bioinformatics*, 2003.

[55] Spiegelhalter D. J. Probabilistic reasoning in predictive expert system. In *Uncertainty in Artificial Intelligence*, 1986.

[56] F. Jelinek. *Statistical methods for speech recognition.* MIT Press, 1997.

[57] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. In *Proceedings of the NATO Advanced Study Institute on Learning in graphical models*, pages 105–161, Norwell, MA, USA, 1998. Kluwer Academic Publishers.

[58] Pearl Judea. A constraint-propagation approach to probabilistic reasoning. In *Proceedings Workshop on Uncertainty in Artificial Intelligence*, 1986.

[59] Yibin Kang, Wei He, Shaun Tulley, Gaorav P. Gupta, Inna Serganova, Chang Rung Chen, Katia Manova-Todorova, Ronald Blasberg, William L. Gerald, and Joan Massague. Breast cancer bone metastasis mediated by the smad tumor suppressor pathway. In *Proceedings of the National Academy of Sciences of the USA*, 2005.

[60] Yibin Kang, Peter M. Siegel, Weiping Shu, Maria Drobnjak, Sanna M. Kakonen, Carlos Cordón-Cardo, Theresa A. Guise, and Joan Massagué. A multigenic program mediating breast cancer metastasis to bone. *Cell*, 3(6):537–549, June 2003.

[61] Khandan Keyomarsi and Arthur B. Pardee. *PNAS*, 1993.

[62] S. Kim, S. Imoto, and S. Miyano. Dynamic bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, 2003.

[63] RA Kinsman, RA Yaroush, E Fernandez, JF Dirks, M Schocket, and J Fukuhara. Symptoms and experiences in chronic bronchitis and emphysema. *Chest*, 83:755–761, 1983.

[64] W. Lam and F. Bacchus. Learning bayesian belief networks: An approach based on the mdl principle. *Computational Intelligence*, 98, July 1994.

[65] S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J. Royal Statistics Society B*, 50(2):157–194, 1988.

[66] Steffen L. Lauritzen. The em algorithm for graphical association models with missing data. Technical report, Department of Statistics, Aalborg University, 1991.

[67] Steffen L. Lauritzen. The em algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 1995.

[68] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Want, M. Kobayashi, H. Horton, and E. L. Brown. Dna expressionmonitoring by hybridization of high density oligonucleotide arrays. *Nature Biotechnology*, 1996.

[69] D. Madigan and A.E. Raftery. Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of American Statistical Association*, 1994.

[70] A. McCallum. *Reinforcement Learning with Selective Perception and Hidden State.* PhD thesis, Univ. Rochester, 1995.

[71] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *Computational Intelligence*, 21, 1953.

[72] B. Middleton, M. Shwe, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part II. Evaluation of diagnostic performance. *SIAM Journal on Computing*, 30:256–267, 1991.

[73] Tom M. Mitchell. *Machine Learning.* Morgan Kaufmann Publishers, Inc., San Mateo, CA, USA, 1988.

[74] Kohei Miyazono. Positive and negative regulation of tgf-β signaling. *Journal of Cell Science*, 113(7):1101–1109, April 2000.

[75] Mundy.G.R. Metastasis to bone:causes, consequences and therapeutic opportunities. *Nature Rev. Cancer*, 2002.

[76] K. Murphy. Learning bayes net structure from sparse data sets. Technical report, Comp. Sci. Div., UC Berkeley, 2001.

[77] K. Murphy and S. Mian. Modelling gene expression data using dynamic bayesian networks, 1999.

[78] Kevin Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning.* PhD thesis, University of California, Berkeley, 2002.

[79] Eric Neufeld. A probabilistic commonsense reasoner. *International Journal of Intelligent Systems*, 1990.

[80] Peter J. Park, Yun Anna Cao, Sun Yong Lee, Jong Woo Kim, Mi Sook Chang, Rebecca Hart, and Sangdun Choi. Current issues for dna microarrays: platform comparison, double linear amplification, and universal rna reference. *Journal of Biotechnology*, 2004.

[81] Judea Pearl. Fusion, propagation, and structuring in belief networks. *Artif. Intell.*, 29(3):241–288, 1986.

[82] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., San Mateo, CA, USA, 1988.

[83] Judea Pearl and Tom S. Verma. A theory of inferred causation. In James F. Allen, Richard Fikes, and Erik Sandewall, editors, *KR'91: Principles of Knowledge Representation and Reasoning*, pages 441–452, San Mateo, California, 1991. Morgan Kaufmann.

[84] D. Piperno, F. Bart, P. Serrier, M. Zureik, and L. Finkielsztejn. General practice patients at risk of chronic obstructive pulmonary disease: epidemiologic survey of 3 411 patients. *La Press Medicale*, 34(21):1612–1614, 2005.

[85] Kornelia Polyak, Mong-Hong Lee, Hedlye Erdjument-Bromage, Andrew Koff, James M. Roberts, Paul Tempst, and Joan Massague. *Cell*, 1994.

[86] Christian P.Robert. *Monte Carlo Statistical Methods.* Springer-Verlag, New York, USA, 2004.

[87] Silja Renooij. *Qualitative Approaches to Quantifying Probabilistic Networks.* PhD thesis, Universiteit of Utrecht, Holland, 2001.

[88] Silja Renooij, Simon Parsons, and Linda C. van der Gaag. Context-specific sign-propagation in qualitative probabilistic networks. In *IJCAI*, pages 667–672, 2001.

[89] Silja Renooij and Linda C. van der Gaag. Enhancing QPNs for trade-off resolution. pages 559–566.

[90] Inga Reynisdottir, Kornelia Polyak, Antonio Iavarone, and Joan Massague. Kip/cip and ihk4 cdk inhibitors cooperate to induce cell cycle arrest in response to tgf-β. *Genes and Development*, 1995.

[91] Hans L. Rieder. Risk of travel-associated tuberculosis. *Clinical Infectious Diseases*, 33:1393–1396, 2001.

[92] Roberts.A.B and Sporn.M.B. The transforming growth factor-betas. *In Peptide Growth Factors and Their Receptors*, 1990.

[93] C N Robson, V. Gnanapragasam, R L Byrne, and A T Collins. *Journal of Endocrinology*, 1999.

[94] Dana Ron, Yoram Singer, and Naftali Tishby. The power of amnesia: Learning probabilistic automata with variable memory length. *Machine Learning*, 1996.

[95] Stuart J. Russell, John Binder, Daphne Koller, and Keiji Kanazawa. Local learning in probabilistic networks with hidden variables. In *IJCAI*, pages 1146–1152, 1995.

[96] E. Segal, Shapira M., A. Regev, D. Peer, D. Botstein, D. Koller, and N. Friedman. Module networks: identifying regulatory modules and their condition-specific regulators from gene-expression data. *Nature Genetics*, 2003.

[97] Joan Seoane, Hong-Van Le, and Joan Massague. *Nature*, 2002.

[98] Joan Seoane, Celio Pouponnot, Peter Staller, Manuela Schader, Martin Eilers, and Joan Massague. *Nature*, 2001.

[99] Sarah E. Seton-Rogers, Yu Lu, Lisa M. Hines, M. Koundinya, J. LaBaer, S. K. Muthuswamy, and Joan S. Brugge. *PNAS*, 2003.

[100] Yigong Shi and Joan Massagué. Mechanisms of tgf-β signaling from cell membrane to the nucleus. *Cell*, 113, June 2003.

[101] Bill Shipley. *Cause and Correlation in Biology: A User's Guide to Path Analysis, Structural Equations and Causal Inference.* Cambridge, 2000.

[102] Peter M. Siegel and Joan Massague. Cytostatic and apoptotic actions of tgf-β in homeostasis and cancer. *Nature Reviews*, 2003.

[103] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search.* MIT, 2000.

[104] Yoshinori Tamada, SunYong Kim, Hideo Bannai, Seiya Imoto, Kousuke Tashiro, Satoru Kuhara, and Satoru Miyano. Estimating gene networks from gene expression data by combining bayesian network model with promoter element detection. *Bioinformatics*, 2003.

[105] R. J. Troisi, F. E. Speizer, Rosner B, Trichopoulos D, and Willett WC. Cigarette smoking and incidence of chronic bronchitis and asthma in women. *Chest*, 108(6):1557–1561, 1995.

[106] Tim Van Allen and Russ Greiner. Model selection criteria for learning belief nets: An empirical comparison. In *Proc. 17th International Conf. on Machine Learning*, pages 1047–1054. Morgan Kaufmann, San Francisco, CA, 2000.

[107] Lalage M. Wakefield and Anita B. Roberts. *Genetics & Development*, 2002.

[108] M. P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artif. Intell.*, 44(3):257–303, 1990.

[109] Michael P. Wellman. Fundamental concepts of qualitative probabilistic networks. *Artificial Intelligence*, 1990.

[110] X. Wen, S. Furhmann, G. S. Micheals, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi. Large-scale temporal gene expressionmapping of central nervous systemdevelopment. *Proc. Nat. Acad. Sci.*, 1998.

[111] J. Whittaker. *Graphical Models in Applied Multivariate Statistics.* Wiley Series in Probability & Statistics, 1990.

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

[113] Juan Juan Yin, Katri Selander, John M. Chirgwin, Mark Dallas, Barry G. Grubbs, Rotraud Wieser, Joan Massagu, Gregory R. Mundy, , and Theresa A. Guise. Tgf-β signaling blockade inhibits PTHrP secretion by breast cancer cells and bone metastases development. *Journal of Clinic Investigation*, 103, January 1999.

[114] Ying Zhang and Rik Derynck. Regulaiton of smad signalling by protein associations and signalling crosstalk. *Trends in Cell Biology*, 9(7):274–279, July 1999.

[115] Yu Zhang, Zhidong Deng, Hongshan Jiang, and Peifa Jia. Dynamic bayesian network (dbn) with structure expectation maximization (sem) for modeling of gene network from time series gene expression data. In *Proceedings of 2006 International Conference on Bioinformatics & Computational Biology (BIOCOMP06)*, 2006.

[116] Qun Zhou, Maryalice Stetler-Stevenson, and Patricia S Steeg. *Oncogene*, 1997.