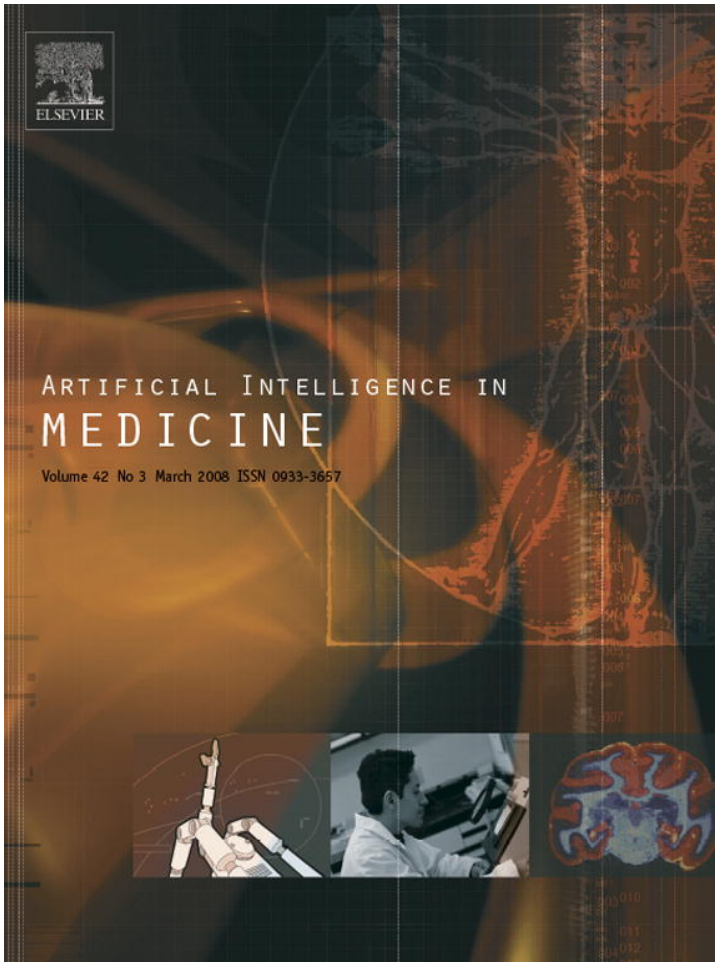


Artificial Intelligence in Medicine (2008) **42**, 229–245

ELSEVIER

ARTIFICIAL INTELLIGENCE IN MEDICINE

http://www.intl.elsevierhealth.com/journals/aiim

Latent tree models and diagnosis in traditional Chinese medicine

Nevin L. Zhang [a,*], Shihong Yuan [b], Tao Chen [a], Yi Wang [a]

[a] *Department of Computer Science & Engineering, The Hong Kong University of Science & Technology, Clear Water Bay Road, Kowloon, Hong Kong, China*
[b] *Department of Chinese Medicine Diagnostics, Beijing University of Traditional Chinese Medicine, 11 Beisanhuan Donglu, Beijing 100029, China*

Received 13 January 2007; received in revised form 17 October 2007; accepted 25 October 2007

KEYWORDS
Machine learning;
Latent structure models;
Multidimensional clustering;
Traditional Chinese medicine;
Syndrome differentiation

Summary

Objective: TCM (traditional Chinese medicine) is an important avenue for disease prevention and treatment for the Chinese people and is gaining popularity among others. However, many remain skeptical and even critical of TCM because of a number of its shortcomings. One key shortcoming is the lack of objective diagnosis standards. We endeavor to alleviate this shortcoming using machine learning techniques.
Method: TCM diagnosis consists of two steps, patient information gathering and syndrome differentiation. We focus on the latter. When viewed as a black box, syndrome differentiation is simply a classifier that classifies patients into different classes based on their symptoms. A fundamental question is: do those classes exist in reality? To seek an answer to the question from the machine learning perspective, one would naturally use cluster analysis. Previous clustering methods are unable to cope with the complexity of TCM. We have therefore developed a new clustering method in the form of latent tree models. We have conducted a case study where we first collected a data set about a TCM domain called KIDNEY DEFICIENCY and then used latent tree models to analyze the data set.
Results: Our analysis has found natural clusters in the data set that correspond well to TCM syndrome types. This is an important discovery because (1) it provides statistical validation to TCM syndrome types and (2) it suggests the possibility of establishing objective and quantitative diagnosis standards for syndrome differentiation. In this paper, we provide a summary of research work on latent tree models and report the aforementioned case study.

* Corresponding author. Tel.: +852 2358 7015; fax: +852 2358 1477.
E-mail address: lzhang@cse.ust.hk (N.L. Zhang).

0933-3657/$ – see front matter
doi:10.1016/j.artmed.2007.10.004

1. Introduction

We present a work that applies machine-learning techniques to solve a problem in TCM (traditional Chinese medicine). To motivate the work, we will first briefly explain the differences between TCM and western medicine, discuss the increasing importance of TCM in health care, and point out one key shortcoming of TCM (Section 1.1). We will then motivate our approach to overcome the shortcoming (Section 1.2), and describe the content and organization of this paper (Section 1.3).

1.1. TCM and western medicine

TCM and western medicine represent two different paradigms. Western medicine approaches the human body from an anatomic and biochemical standpoint. It emphasizes specific disease entities and focuses on pathophysiological mechanisms. TCM approaches the human body from an energetic and functional standpoint. It believes that disease is a result of disharmony within the body and between the body and the environment. TCM diagnosis and treatment involve identifying the factors that are out of balance and attempting to bring them back into harmony. Instead of micro level laboratory tests, TCM diagnosis is primarily based on overall observation of patient through inspection, auscultation and olfaction, interrogation and palpation.

Western medicine is often complicated with undesirable side effects and does not have effective treatments for illness such as irritable bowel syndrome and menopause [1]. An increasing number of patients in developed countries are turning to traditional medicine for alternative treatment. According to a report by World Health Organization [2], the global market for herbal medicines currently stands at over US$60 billion annually and is growing steadily. As a matter of fact, 80% of Africans use traditional medicine; 90% of Canadians, 49% of French people, 48% of Australians, 42% of Americans, and 31% of Belgians have received traditional medicine treatment at least once in their life times; and 77% of German clinics recommend acupuncture for pain relief.

TCM is a systematic traditional medicine system and it has been clinically observed to have dramatic performance in treating many chronic and systematic diseases. However, several shortcomings are hindering its wide acceptance. One key shortcoming is the lack of objective diagnosis standards. In practice, this results in variability of diagnosis conclusions among TCM practitioners [1], which in turn raises doubts about TCM in the minds of many.

To alleviate the situation, researchers in China have been seeking for laboratory tests that can be used as gold standards for TCM diagnosis [3,4]. Despite extensive research work for more than half a century, little has been achieved [5]. In this paper, we propose and investigate a different approach.

1.2. Syndrome differentiation, clustering, and latent structures

TCM diagnosis consists of two steps. In the first step, a doctor collects patient information through inspection, auscultation and olfaction, interrogation and palpation. In the second step, he reaches diagnosis conclusions by analyzing patient information based on TCM theories and his experiences. The second step is known as *syndrome differentiation*, while the first step can be called *patient information gathering*.

Subjectivity is an issue in both the patient information gathering step and the syndrome differentiation step. We focus on the latter. Our long-term goal is to establish objective and quantitative standards for syndrome differentiation using machine learning techniques. How could this goal be possibly achieved?

When viewed as a black box, syndrome differentiation is simply a classifier that classifies patients into different classes based on their symptoms. Do those classes exist in reality? Can we characterize them mathematically? To seek answers to those questions from the machine learning perspective, one would naturally use cluster analysis. The idea is to: (1) collect patient data systematically; (2) perform cluster analysis to identify natural clusters of patients in these data; (3) compare the natural clusters with the classes mentioned in TCM. If some of the natural clusters match the TCM classes, then we would have provided statistical validation for TCM classes. Moreover, we can use the natural clusters as a basis for syndrome differentiation. The TCM classes are described in natural language and are often vague and prone to subjectivity. The natural clusters, however, are described in the language of mathematics and are objective (in a relative sense). Therefore, they can serve as the foundation for objective and quantitative syndrome differentiation standards.

Which clustering method should we use? To answer this question, we need to take a closer look at syndrome differentiation. There are several syndrome differentiation systems, each focusing on a different perspective of the human body and with its own theory. The theories describe relationships

between syndrome factors and symptoms, as illustrated by this excerpt:

KIDNEY YANG[1] is the basis of all YANG in the body. When KIDNEY YANG is in deficiency, it cannot warm the body and the patient feels cold, resulting in intolerance to cold, cold limbs, and cold lumbus and back [6].

The syndrome factor mentioned here is KIDNEY YANG DEFICIENCY. It is not directly observed. Rather, it is similar in nature to concepts such as 'intelligence' and is indirectly measured through its manifestations. Hence, we say that it is a *latent variable*. In contrast, symptom variables such as 'cold limbs' are directly observed and we call them *manifest variables*. TCM theories involve a large number of latent and manifest variables. Abstractly speaking, they describe relationships among latent variables, and between latent variables and manifest variables. They can be viewed as *latent structure models* specified in natural language. To study syndrome differentiation, therefore, we need statistical models that involve multiple interrelated latent variables, i.e. *latent structure models* described in the language of mathematics.

How are latent structure models related to clustering? A latent structure model represents a *multidimensional classification* of objects. Imagine a latent structure model about people that contains two latent variables 'intelligence' and 'personality'. Suppose the first latent variable has three possible values 'high', 'medium' and 'low', while the second latent variable has two possible values 'outgoing' and 'withdrawn'. Then, the model represents one classification of people along the dimension of 'intelligence' into three classes, and at the same time another classification along the dimension of 'personality' into two classes. By considering the two dimensions at the same time, one gets a cross classification where one can speak of classes such as 'outgoing people with high intelligence' and 'withdrawn people with medium intelligence'.

Now suppose we start from a data set and by analyzing the data set we obtain a latent structure model. That would mean that we simultaneously cluster data in multiple ways. Each latent variable in the resultant model represents one dimension along which to cluster the data set, and different latent variables represent different ways to cluster the data set. So, to analyze a data set using latent structure models is to perform *multidimensional clustering* on the data set.

Is the aforementioned *latent structure approach* to the study of syndrome differentiation feasible? To answer this question, we have studied a special class of latent structure models called latent tree models, and we have used latent tree models to investigate a subdomain of TCM diagnosis called KIDNEY DEFICIENCY. This case study shows that (1) there exist natural clusters in data that correspond to TCM syndrome types, (2) we can find those classes using latent tree models. Those indicate that the latent structure approach is indeed feasible.

In summary, we propose a novel approach to the study of syndrome differentiation. The long-term goal is to establish objective and quantitative standards for syndrome differentiation. This paper presents the tools that have been developed for achieving the goal and demonstrates the feasibility of the approach through a case study.

1.3. Content and organization

In the first half of the paper, we provide a summary of research work on latent tree models. We will first introduce latent tree models as a generalization of latent class models [7,8](Section 2). Then we will discuss two representational issues, namely model equivalence (Section 3) and identifiability (Section 4). In Section 5, we will survey algorithms for learning latent tree models. In the second half of the paper, we report the aforementioned case study. We will start by describing the data set and the data analysis process (Section 6). In Sections 7 and 8, we interpret the latent variables and clusters in the resultant model. The reader will see that the latent variables correspond to syndrome factors and the latent clusters can be interpreted as syndrome types. The paper will end in Section 9 with a summary and remarks about future directions.

2. Latent tree models

The concept 'latent variable' is defined with respect to some given data set. A variable is *observed* with respect to a data set if its value can be found in at least one of the records. A variable is *latent* with respect to a data set if its value is missing from all the records. For simplicity, we will often talk about latent variables without explicitly referring to data sets.

A *latent class model* [7,8] is a Bayesian network that consists of a latent variable X and a number of observed variables $Y_1, Y_2, \ldots, Y_n$. All the variables are categorical and the relationships among them are as shown in Fig. 1. In applications, the latent variable stands for some latent factor and the observed variables stand for manifestations of the latent factor. The observed variables are also called *manifest variables*.

[1] Words in small capital letters are reserved for TCM terms.

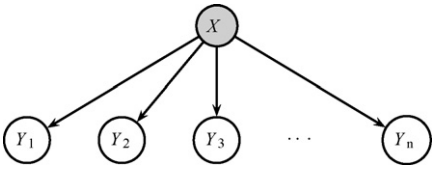


Figure 1 The structure of a latent class model.

To learn a latent class model is to: (1) determine the *cardinality*, i.e. the number of states for the latent variable X, and (2) estimate the model parameters $P(X)$ and $P(Y_i|X)$. A state of the latent variable X corresponds to a class of individuals in a population. It is called a *latent class*. Thus, to determine the cardinality of X is to determine the number of latent classes, and to estimate $P(Y_i|X)$ is to reveal the statistical characteristics of the latent classes. So, latent class analysis is a type of *model-based cluster analysis*.

Underlying a latent class model is the assumption that the manifest variables are mutually independent given the latent variable. A serious problem with the use of latent class models, known as *local dependence*, is that this assumption is often violated. If one does not deal with local dependence explicitly, one implicitly attributes it to the latent variable. This can lead to spurious latent classes and poor model fit. It can also degenerate the accuracy of classification because locally dependent manifest variables contain overlapping information [9].

The local dependence problem is acknowledged in the latent class analysis literature [10]. Methods for detecting and modeling local dependence have been proposed. To detect local dependence, one typically compares observed and expected cross-classification frequencies for pairs of manifest variables. To model local dependence, one can join manifest variables, introduce multiple latent variables, or reformulate latent class models as log-linear models and then impose constraints on them.

Previous work for dealing with local dependence is not sufficient for a number of reasons. First, there are no criteria for making the trade-off between increasing the cardinalities of existing latent variables and introducing additional latent variables. In [10,11], cardinalities of all latent variables are fixed at 2 while the number of latent variables is allowed to change. In most other work, the standard one-latent-variable structure is assumed and fixed, while the cardinality of the latent variable is allowed to change. Second, the search for the best model is carried out manually. Typically only a few simple models are considered [11,12]. Finally, when there are multiple pairs of locally dependent manifest variables, it is not clear which pair should be tackled first, or if all pairs should be handled simultaneously.

Latent tree models are previously known as HLC (hierarchical latent class) models [13,14]. They are a generalization of latent class models. A *latent tree model* is a Bayesian network where

- the network structure is a rooted tree;
- the internal nodes represent latent variables and the leaf nodes represent manifest variables; and
- all the variables are categorical.

Fig. 2 shows an example latent tree model. In this paper, we do not distinguish between variables and nodes. So we sometimes speak also of *manifest nodes* and *latent nodes*.

The class of latent tree models is clearly much larger than the class of latent class models. In the meantime, latent tree models remain computationally attractive because their structures are restricted to trees. Latent tree models can alleviate the local dependence problem that latent class models face. We will later present search-based algorithms for learning latent tree models. When there is no local dependence, the algorithms return latent class models. When local dependence is present, they return latent tree models with local dependence appropriately modeled.

In Section 1.2, we have explained why we need latent tree models in TCM research. There are two other reasons why latent tree models are interesting in general. First, latent tree models represent complex dependencies among observed variables and yet are computationally simple to work with. It was for the reason that Pearl [15] first identified latent tree models as a potentially useful class of Bayesian networks. Second, the endeavor of learning latent tree models can reveal interesting latent structures. Researchers have already been inferring latent structures from observed data. One example is the reconstruction of phylogenetic trees [16], which can be viewed as special latent tree models. Latent structure discovery is also interesting for TCM. After all, TCM theories themselves are latent structure models described in natural language.

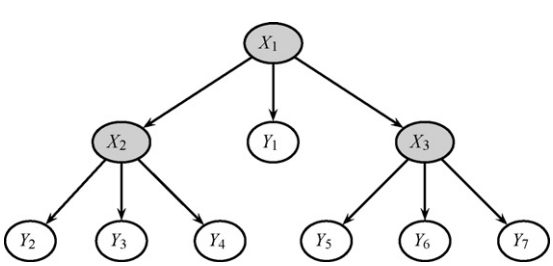


Figure 2 An example latent tree model. The X_i's are latent variables and the Y_j's are manifest variables.

3. Model equivalence

We write a latent tree model as a pair $\mathcal{M} = (\mathcal{G}, \theta)$, where $\mathcal{G}$ stands for the model structure plus cardinalities of variables, and θ stands for the vector of probability parameters. We will sometimes refer to $\mathcal{G}$ also as a latent tree model. Two latent tree models $(\mathcal{G}, \theta)$ and $(\mathcal{G}', \theta')$ are *marginally equivalent* if they share the same manifest variables $Y_1, Y_2, \ldots, Y_n$ and

$$P(Y_1, \ldots, Y_n|\mathcal{G}, \theta) = P(Y_1, \ldots, Y_n|\mathcal{G}', \theta'). \quad (1)$$

A latent tree model $\mathcal{G}$ *includes* another $\mathcal{G}'$ if for any parameter value vector θ' of $\mathcal{G}'$, there exists parameter value vector θ of $\mathcal{G}$ such that $(\mathcal{G}, \theta)$ and $(\mathcal{G}', \theta')$ are marginally equivalent. If $\mathcal{G}$ includes $\mathcal{G}'$, then $\mathcal{G}$ can represent any distributions over the manifest variables that $\mathcal{G}'$ can. If $\mathcal{G}$ includes $\mathcal{G}'$ and vice versa, we say that $\mathcal{G}$ and $\mathcal{G}'$ are *marginally equivalent*. Marginally equivalent models are *equivalent* if they have the same number of independent parameters. It is impossible to distinguish between equivalent models based on data if penalized likelihood score [17] is used for model selection.

Let X_1 be the root of a latent tree model $\mathcal{G}$. Suppose X_2 is a child of X_1 and it is also a latent node. Define another latent tree model $\mathcal{G}'$ by reversing the arrow $X_1 \rightarrow X_2$. Variable X_2 becomes the root in the new model. The operation is called *root walking*; the root has walked from X_1 to X_2.

It has been proved that root walking leads to equivalent models [14]. Therefore, the root and edge orientations of a latent tree model cannot be determined from data. What we learn from data are *unrooted latent tree models*, which are latent tree models with all directions on the edges dropped. Fig. 3 shows the unrooted latent tree model that corresponds to the latent tree model in Fig. 2.

An unrooted latent tree model represents a class of latent tree models. Members of the class are obtained by rooting the model at various latent nodes. Semantically it is a Markov random field on an undirected tree. The concepts of marginal equivalence and equivalence can be defined for unrooted latent tree models in the same way as for rooted models. From now on when we speak of latent tree models, we always mean unrooted latent tree models unless explicitly stated otherwise.

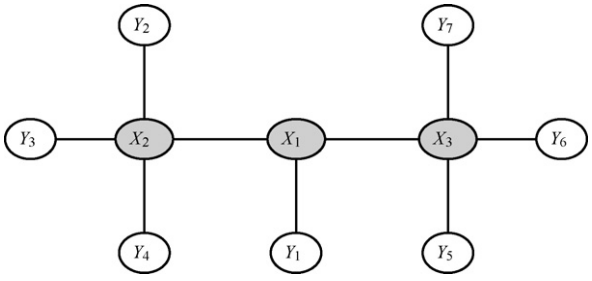


Figure 3 The unrooted latent tree model that corresponds to the latent tree model in Fig. 2.

4. Identifiability and regularity

Let Y_1, Y_2, ..., Y_n be the manifest variables in a latent tree model $\mathcal{G}$. If there exist two different parameter value vectors θ and θ' such that

$$P(Y_1, \ldots, Y_n|\mathcal{G}, \theta) = P(Y_1, \ldots, Y_n|\mathcal{G}, \theta'), \quad (2)$$

then model $\mathcal{G}$ is said to be *unidentifiable*. In such a case, it is impossible to distinguish between the two parameter value vectors θ and θ' based on data.

Identifiability is closely related to three other concepts, namely effective dimensions, standard dimensions, and parsimonious models. Let k be the product of the cardinalities of the manifest variables. For a given parameter value vector $\theta, P(Y_1, \ldots, Y_n|\mathcal{G}, \theta)$ is a point in the $\mathbb{R}^{k-1}$ space. If we let θ run through all its possible values, we would get a set of points in the $\mathbb{R}^{k-1}$ space. This set is a manifold [18]. We call it the *manifold image* of model $\mathcal{G}$. The *effective dimension* of $\mathcal{G}$ is defined to be the dimension of its manifold image. The number of independent parameters for $\mathcal{G}$ is called its *standard dimension*. The standard dimension of a latent tree model is always greater than or equal to its effective dimension.

A latent tree model $\mathcal{G}$ is *parsimonious* if there does not exist another model that is marginally equivalent to $\mathcal{G}$ and that has fewer independent parameters than $\mathcal{G}$. A latent tree model is *strongly parsimonious* if its standard dimension is the same as its effective dimension. Clearly, if a model is not parsimonious, then it is not strongly parsimonious. If a model is not strongly parsimonious, then it is unidentifiable.

Are strongly parsimonious models identifiable? The answer is negative. Let $\mathcal{G}$ be an arbitrary latent tree model, strongly parsimonious or not, and let θ be a parameter value vector for $\mathcal{G}$. Further let X be a latent variable in $\mathcal{G}$ and s_1 and s_2 be two states of X. Let θ' be obtained from θ by swapping the parameter values pertaining to s_1 and those pertaining to s_2. Then the two pairs $(\mathcal{G}, \theta)$ and $(\mathcal{G}, \theta')$ satisfy Eq. (2). Hence, $\mathcal{G}$ is unidentifiable. This means that one cannot identify the identities of the states of latent variables based on data. We call this the *unidentifiability of latent state*. It is an intrinsic property of all models with latent variables. It should be noted that, in applications, one can usually determine the meanings of the latent states from domain knowledge.

If a model is not parsimonious, then it contains redundancies. The concept of regularity was introduced to identify some of the redundances [14]. Let $|X|$ stand for the cardinality of a variable X. For a latent variable Z in a latent tree model, enumerate its neighbors as $X_1, X_2, \ldots, X_k$. A latent tree model is *regular* if for any latent variable Z,

$$|Z| \leq \frac{\prod_{i=1}^{k}|X_i|}{\max_{i=1}^{k}|X_i|}, \tag{3}$$

and when Z has only two neighbors, one of the neighbors must be a latent node and the inequality (3) holds strictly.

If a latent tree model $\mathcal{G}$ is not regular, then there exists a regular model $\mathcal{G}'$ that is marginally equivalent to $\mathcal{G}$ and has fewer independent parameters [14]. The model $\mathcal{G}'$ can be obtained from $\mathcal{G}$ through the following *regularization* process:

(1) For each latent variable Z in $\mathcal{G}$,
   (a) If it violates inequality (3), reduce the cardinality of Z to $\prod_{i=1}^{k}|X_i|/\max_{i=1}^{k}|X_i|$.
   (b) If it has only two neighbors with one being a latent node and it violates the strict version of inequality (3), remove Z from $\mathcal{G}$ and connect the two neighbors of Z.
(2) Repeat Step 1 until no further changes.

Regularization can remove redundancies in latent tree models. Can it remove all the redundancies? In other words, are regular models parsimonious? This is an open question. What we do know is that some regular models are not strongly parsimonious. Examples of this can be found in Table 1, which is borrowed from [19]. The table shows the effective and standard dimensions of several latent class models. Latent class models are identified using the cardinalities of their variables. For example, "2:3,3" refers to a latent class model with two manifest variables, where the cardinality of the latent variable is 2 while those of the two manifest variables are both 3. All the models in the table are regular. However, their effective dimensions are smaller than their standard dimensions. Hence, they are not strongly parsimonious.

Table 1 Standard and effective dimensions of several latent class models

Model	Effective dimension	Standard dimension
2:2,2	3	5
3:2,2,2	7	11
4:2,2,2	7	15
2:3,3	7	9
3:4,5	17	23
4:3,3,3	25	27

In addition to helping us remove redundancies, the concept of regularity also provides a finite search space for the algorithms to be described in the next section. Let, $\mathbf{Y}$ be a set of variables. There are infinitely many possible latent tree models with $\mathbf{Y}$ as manifest variables. However, only a finite number of them are regular [14].

5. Learning latent tree models

Assume that there is a collection $\mathcal{D}$ of i.i.d. samples that were generated by an unknown regular latent tree model. Each sample contains values for some or all the manifest variables. By *learning latent tree model* we mean the effort to reconstruct, from $\mathcal{D}$, the regular unrooted latent tree model that corresponds to the generative model.

Three search-based algorithms have been proposed for this task, namely DHC (double hill-climbing) [13,14], SHC (single hill-climbing) [20], and HSHC (heuristic single hill-climbing) [20]. All those algorithms aim at finding the model with the highest BIC score [21]. The BIC score of a model $\mathcal{G}$ given data set $\mathcal{D}$ is given below:

$$BIC(\mathcal{G}|\mathcal{D}) = \log P(\mathcal{D}|\mathcal{G}, \theta^*) - \frac{d(\mathcal{G})}{2}\log N, \tag{4}$$

where θ^* is the maximum likelihood estimate of model parameters, $d(\mathcal{G})$ the standard dimension of $\mathcal{G}$, and N is the sample size.

The BIC score is a large sample approximation of the marginal likelihood $P(\mathcal{D}|\mathcal{G})$ derived in a setting where all variables observed. Geiger et al. [18] have re-done the derivation for latent variable models and arrived at another scoring function called the BICe score. The BICe score is the same as the BIC score except that the standard dimension $d(\mathcal{G})$ is replaced by the effective dimension of the model. Theoretically, BICe is advantageous over BIC. Why BIC is still used to guide the search algorithms? There are three reasons. First, effective model dimensions are difficult to compute, despite recent decomposition results [22]. Second, there is no substantial empirical evidence showing that BICe is advantageous over BIC in practice. Third, our experiences with about one dozen data sets indicate that one can find good models with BIC.

5.1. Double hill-climbing

The DHC algorithm searches in the space of regular latent tree models. It starts with the simplest latent tree model, i.e. the model with one binary latent

variable. At each step of search, it first generates a number of candidate model structures by modifying the structure of the current model using three search operators, namely node introduction, node deletion, and node relocation. It then optimizes the cardinalities of the latent variables in each of the candidate model structures, resulting in candidate models. Finally, it evaluates the candidate models and picks the best one to seed the next step of search. Search terminates when the best candidate model is no better than the current model. To optimize the cardinalities of the latent variables in a candidate model structure, the algorithm employs another hill-climbing routine. This is why it is called *DHC (double hill-climbing)*.

Node introduction is conceptually the most important among the three search operators. To motivate the operator, consider the latent tree model $\mathcal{G}_1$ shown in Fig. 4. It assumes that the manifest variables are mutually independent given the latent variable. What if this assumption is not true? To be more specific, what if Y_1 and Y_2 are correlated given X? A natural thing to do in this case is to introduce a new latent node X_1 and make it a parent for Y_1 and Y_2, as shown in $\mathcal{G}_2$.

The *NI (node introduction)* operator is the result of applying the idea to unrooted latent tree model structures. Let X be a latent node in such a structure. Suppose X has more than two neighbors. For any two neighbors Z_1 and Z_2 of X, we can *introduce a new latent node Z to mediate X and its neighbors* Z_1 *and* Z_2. Afterwards X is no longer connected to Z_1 and Z_2. Instead X is connected to Z and Z is connected to Z_1 and Z_2. Consider the model structure $\mathcal{G}_1'$ in Fig. 4. Introducing a new latent node X_1 to mediate X and its neighbors Y_1 and Y_2 results in the model structure $\mathcal{G}_2'$. For the sake of computational efficiency, node introduction is not allowed to involve three or more neighbors of a latent node. This implies that we cannot reach $\mathcal{G}_3'$ from $\mathcal{G}_1'$ using the NI operator only.

ND (node deletion) is the opposite of node introduction. Let X be a latent node and let $Z_1, Z_2, \ldots, Z_k$ be the neighbors of X. If one of these neighbors, say Z_1, is also a latent node, then we can *delete X with respect to* Z_1. This means to remove X from the model and connect Z_1 to each of $Z_2, Z_3, \ldots, Z_k$. In Fig. 4, deleting X_1 from $\mathcal{G}_2'$ or $\mathcal{G}_3'$ with respect to X results in $\mathcal{G}_1'$.

The third search operator is called *node relocation*. Let X be a latent node and Z be a neighbor of X. Suppose X has another neighbor Z' that is also a latent node. Then we can *relocate Z from X to Z′*. This means to disconnect Z from X and reconnect it to Z'. In Fig. 4, relocating Y_3 from X to X_1 in $\mathcal{G}_2'$ results in $\mathcal{G}_3'$. For the sake of computational efficiency, it is not allowed to relocate Z to a latent node that is not a neighbor of X. In Fig. 3, for example, we cannot relocate Y_2 from X_2 to X_3.

Given a latent tree model structure, how do we determine the cardinalities of the latent variables? The answer is to search in the space of all the regular latent tree models with the given model structure. The search starts with the model where the cardinalities of all the latent variables are set at 2. At each step, we obtain a number of candidate models by increasing the cardinality of a single latent variable by one. Irregular candidate models are discarded.

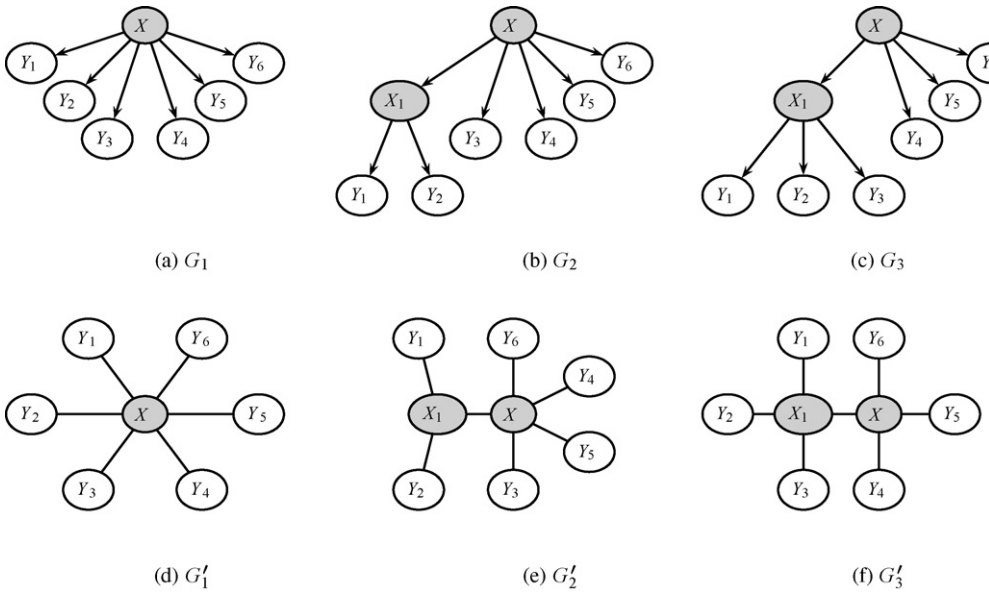


Figure 4 Illustration of structural search operators.

Each of the candidate models is then evaluated and the best one is picked to seed the next search step. To evaluate a model, one needs to estimate its parameters. The EM algorithm [23,24] is used for this task.

Zhang [14] tested the DHC algorithm on both synthetic data and real-world data. The synthetic data sets were generated from a latent tree model whose structure is shown in Fig. 2. The sample size ranges from 5000 to 100,000. The models obtained by DHC are of high quality because: (1) their structures are either identical to that of the generative model or only one step away from it, and (2) their logarithmic scores on a testing data set are very close to that of the generative model. In another setting where strong dependence between variables were enforced, DHC correctly recovered the structure of the generative model for all sample sizes. The real-world data sets were taken from the latent class analysis literature. In all cases but one, DHC found the models that are considered to be the best in the literature. The exception happened with a data set for which latent tree models are apparently inappropriate.

The DHC algorithm has one serious drawback, namely its extremely high computational complexity. Although the synthetic data sets involve only seven manifest variables, DHC took around 100 h on a 1 GHz Pentium III machine to analyze each of them.[2] As such, DHC is a concept testing algorithm.

5.2. Single hill-climbing

A latent tree model $\mathcal{G}$ consists of a network structure, the cardinalities of the manifest variables, and the cardinalities of the latent variables. If the cardinalities of the latent variables are unknown, then $\mathcal{G}$ is called a *pre-LT (pre-latent-tree) model*. The outputs of the three search operators used by DHC are pre-LT models and the cardinalities of the latent variables are optimized using a separate hill-climbing routine. We have been vague about this point until now for readability.

The SHC algorithm is the same as the DHC algorithm in that they both search in the space of regular latent tree models. However, there are two major differences. First, the search operators in SHC produces latent tree models rather than pre-LT models. This implies attention must be paid to the cardinalities of the latent variables when designing the operators. Second, SHC does not have a separate routine to optimize the cardinalities of latent variables. They are optimized at the same time as the model structure.

Five search operators are used in SHC, namely node introduction, node deletion, node relocation, state introduction, and state deletion. The *NI (node introduction)* operator of SHC is the same as the NI operator of DHC, except that it specifies a cardinality for the newly introduced latent node. When introducing a new latent node Z to mediate a latent node X and two of its neighbors, we set $|Z| = |X|$. Let $\mathcal{G}$ and $\mathcal{G}'$ the model before and after the operation. We have the nice property that $\mathcal{G}'$ includes $\mathcal{G}$.

The *ND (node deletion)* and *NR (node relocation)* operators of SHC are the same as the ND and NR operators of DHC, except that cardinalities of latent variables are no longer ignored. The SI (*state introduction*) operator increases the cardinality of a latent variable by one, while the SD (*state deletion*) operator decreases the cardinality of a latent node by one. Applying the operators to a regular latent tree model might result in irregular models. For this reason, they are always followed immediately by the regularization process described in Section 4. The operators can sometimes make a latent node a leaf node. When this happens, the node is simply deleted.

The SHC algorithm begins with the simplest latent tree model, i.e. the model with one binary latent node. It works in two phases. In Phase I, SHC expands models by applying the NI and SI operators. The aim is to improve the likelihood term of the BIC score. The NR operator is also used in Phase I. In Phase II, SHC simplifies models by applying the ND and SD operators. The aim is to reduce the penalty term of the BIC score, while keeping the likelihood term more or less the same. If model quality is improved in Phase II, SHC goes back to Phase I and the process repeats itself.

In Phase I, SHC needs to choose between NI and SI operations. There is an issue of *operation granularity* here. Suppose there are 100 manifest variables. At the first step of search, the current model contains only one binary latent variable and it has only two states. Denote the latent variable by X_0. Applying the SI operator to X_0 would introduce 101 additional model parameters. Introducing a new latent node to mediate X_0 and two of its neighbors, on the other hand, would increase the number of model parameters by only 2. The latter operation is clearly of much finer-grain than the former. The former is like a bulldozer if the latter is compared to a shovel.

To deal with operation granularity, SHC adopts the so-called cost-effectiveness principle when choosing among candidate models in Phase I. Let $\mathcal{G}$ be the current model and $\mathcal{G}'$ be a candidate model. Define

[2] The running times for different data sets are more or less the same because the data sets contain the same number of distinct samples although their sample sizes are different.

the *unit score improvement of* $\mathcal{G}'$ *over* $\mathcal{G}$ *given data* $\mathcal{D}$ to be

$$U(\mathcal{G}', \mathcal{G}|\mathcal{D}) = \frac{BIC(\mathcal{G}'|\mathcal{D}) - BIC(\mathcal{G}|\mathcal{D})}{d(\mathcal{G}') - d(\mathcal{G})} \quad (5)$$

It is the increase in model score per unit increase in model complexity. The *cost-effectiveness* principle states that, among all candidate models, choose the one that has the highest unit score improvement over the current model.

Phase I of SHC involves not only the NI and SI operators, but also the NR operator. A technical difficulty might arise when computing the quantity $U(\mathcal{G}', \mathcal{G}|\mathcal{D})$ for a candidate model generated by NR. The NR operator does not necessarily increase the number of model parameters. The denominator $d(\mathcal{G}') - d(\mathcal{G})$ could be zero or negative. One way to overcome the difficulty is to replaced the denominator with $max\{1, d(\mathcal{G}') - d(\mathcal{G})\}$.

5.3. Heuristic single hill-climbing

SHC is more efficient than DHC, but it still does not scale up well. Here is the reason. At each search step, SHC generates a list of candidate models, evaluates each of them, and selects the best one. To evaluate a model, it first runs the EM algorithm to obtain the maximum likelihood estimation of the parameters and then computes the BIC score. EM is known to be computationally expensive. SHC runs EM on each candidate model and hence has high computational complexity.

One way to reduce the complexity of SHC is to apply the technique of structural EM [25]. The idea is to complete the data set $\mathcal{D}$ based on the current model and evaluate the candidate models using the completed data set $\mathcal{D}_c$. Parameter optimization based on $\mathcal{D}_c$ does not require EM and hence EM is avoided during model evaluation and selection. The *HSHC (Heuristic SHC)* algorithm is the result of incorporating this strategy into SHC.

There is an important issue that one needs to address when incorporating structural EM into SHC. Structural EM cannot be directly applied to the candidate models generated by the NI, SI, SD operators. The latent nodes in those models are different from those in the current model. NI introduces a latent node that is not in the current model, while SI and SD alter the cardinality of one latent variable in the current model. $\mathcal{D}_c$ is complete with respect to the variables in the current model, but it is incomplete with respect to the variables in the candidate models generated by the NI, SI and SD operators.

To solve the problem, HSHC divides the candidate models into groups, with one group for each operator. Models in a group are compared with each other using heuristics based on the completed data set $\mathcal{D}_c$, and the best one is selected. Thereafter a second model selection process is invoked to choose the overall best model among the best models of the groups. This second process is the same as the model selection process in SHC, except that there is only one candidate model for each operator. Here the model parameters are optimized using EM.

The heuristics for ranking candidate models are different for different operators. Here we explain the heuristic for NI. Let $\mathcal{G}$ be the current model and $\mathcal{G}'$ be a candidate model obtained from $\mathcal{G}$ by introducing a new latent node Z to mediate a latent node X and two of its neighbors Z_1 and Z_2. As explained in Section 5.1, the new node would be necessary if Z_1 and Z_2 are not independent of each other given X. This hypothesis can be tested based on the completed data $\mathcal{D}_c$. The variables X, Z_1 and Z_2 are all observed in $\mathcal{D}_c$. Hence, one can compute the G-squared statistic for testing the hypothesis that Z_1 and Z_2 are independent given X. The larger the statistic, the further away Z_1 and Z_2 are from being independent given X, and hence the more necessary it is to introduce a new node. Therefore, one can rank all the candidate models generated by NI using the G-squared statistics and pick the model with the largest value.

There is one more issue to clarify before the description of the HSHC algorithm is complete. For a given operator, instead of choosing the best model, one can choose the best K, for some integer K, models based on heuristic. This top-K scheme reduces the chance of local maxima. In general, the larger the K, the lower the probability of encountering local maxima. On the other hand, larger K also implies running EM on more models and hence longer computation time.

For the sake of computational efficiency, HSHC replaces EM with the so-called local EM in the top-K scheme. Let $\mathcal{G}'$ be a candidate model obtained from the current model $\mathcal{G}$ by applying one of the search operators. The parameters for $\mathcal{G}$ have been optimized and $\mathcal{G}'$ differs from $\mathcal{G}$ only in one or two nodes. *Local EM* optimizes only the parameters pertaining to those one or two nodes in $\mathcal{G}'$ while freezing the values of all the other parameters. Obviously, model parameters obtained by local EM deviate from those obtained by EM. To avoid accumulation of deviations, HSHC runs EM once at the end of each search step on the best model.

5.4. Empirical results with SHC and HSHC

How much more efficient is SHC than DHC? To answer this question, SHC was tested on one of the synthetic data sets mentioned in Section 5.1[20]. SHC turned

out to be 22 times faster than DHC, and it obtained the same model as DHC. No more experiments were run to compare DHC and SHC because DHC is extremely slow.

How much more efficient is HSHC than SHC? Can HSHC find high quality models? To answer those questions, experiments were conducted on five synthetic data sets with 6, 9, 12, 15 and 18 manifest variables, respectively and 10,000 records [20]. For the top-K scheme in HSHC, three values were used for K, namely 1, 2 and 3. The experiments were run on a 2.26 GHz Pentium 4 machine and the time limit was set at 100 h. The results indicate that HSHC is much more efficient than SHC and scales up much better. In particular, SHC was not able to finish analyzing the two data sets with 15 and 18 manifest variables. On the data set with 12 manifest variables, it was 10 times slower than HSHC when K was set at 3. When $K = 1$, the models found by HSHC for the data sets with 15 and 18 manifest variables are of poor quality. When $K = 2$ or 3, however, all the models reconstructed by HSHC match the generative models extremely well in terms of the joint distribution of the manifest variables. The structures of these models are either identical or very similar to the structures of the generative models.

How does HSHC perform on real-world data? Can it discover interesting latent structures and learn good probabilistic models? To answer those questions, the algorithm was tested on the CoIL Challenge 2000 data set [26]. This data set consists of 5,822 customer records of an insurance company. Each record contains socio-demographic information and information about insurance product ownerships. There are 86 attributes. The task is to learn a model and use it to predict who would buy mobile home insurance policies.

Before the data set was fed to HSHC, it was first preprocessed and the resulting data set contains 42 manifest variables. Four different values were used for K in the top-K scheme, namely 1, 5, 10 and 20. The best model was found in the case of $K = 10$. The analysis took 121 h. The structure of the best model is shown in Fig. 5.

The latent structure discovered by HSHC is interesting. This can be appreciated from several angles. First of all, the data set contains two variables for each type of insurance. For example, the two variables for bicycle insurance are 'contribution to bicycle insurance policies (Y_{62})' and 'number of bicycle insurance policies (Y_{83})'. HSHC introduced a latent variable for each of such pairs. The latent variable introduced for Y_{62} and Y_{83} is X_{11}. Obviously, X_{11} can be interpreted as 'attitude toward bicycle risks'. Similarly, X_{10} can be interpreted as 'attitude toward motorcycle risks', X_9 as 'attitude toward moped risks', and so on.

Next consider the manifest variables below the latent variable X_8. Besides 'social security', they are all related to private vehicles. HSHC concluded that people's decisions to buy insurance for private vehicles are influenced by one common latent variable. This is clearly reasonable and X_8 can be interpreted as 'attitude toward private vehicle risks'. The variables about heavy vehicles such as car, mobile home, and boat are placed under X_{12}. This is also reasonable and X_{12} can be interpreted as 'attitude toward heavy private vehicle risks'.

Similarly, we see that X_1 corresponds to 'attitude toward firm risks' and X_{15} means 'attitude toward agriculture risks'. The two latent variables X_3 and X_6 capture attitudes toward risks in daily life; X_{21} summarizes socio-demographic information contained in the manifest variables Y_{04}, Y_{05} and Y_{43};

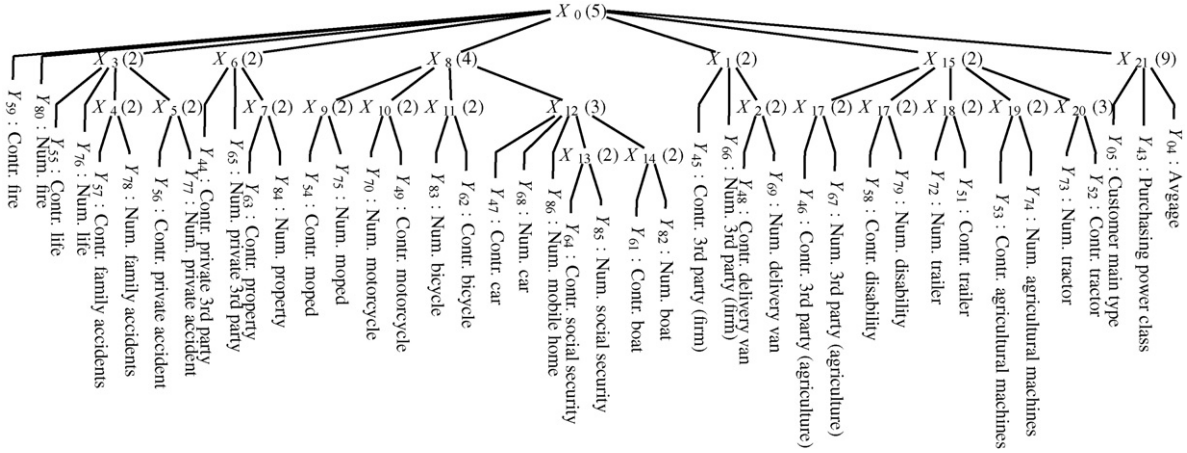


Figure 5 Latent tree model found by HSHC for the CoIL Challenge 2000 data. The number next to a latent variable is the cardinality of that variable.

and X_0 can be interpreted as 'general attitude toward risks'.

It is particularly interesting to note that, although delivery vans (Y_{48} and Y_{69}) and tractors (Y_{37} and Y_{52}) are vehicles, HSHC did not conclude that the decisions to buy insurance for delivery vans and tractors are influenced by 'attitude toward vehicle risks' (X_8). Rather, it correctly concluded that the decision to buy insurance for delivery vans is influenced by 'attitude toward firm risks' and the decision to buy insurance for tractors is influenced by 'attitude toward agriculture risks'.

Is the latent tree model found by HSHC a good probabilistic model for the domain? The prediction task of CoIL Challenge 2000 provides one criterion for answering this question. In CoIL Challenge 2000, there is a test set of 4000 records that contains 238 mobile home policy owners. The prediction task requires the participants to identify a subset of 800 records that contains as many mobile home policy owners as possible. The random selection by the contest organizers resulted in 42 policy owners, while the best entry identified 121 policy owners. The selection based on the model found by HSHC includes 110 policy owners. This is good performance considering that HSHC aims at optimizing BIC score rather than classification error.

For the purpose of building a probabilistic model for the domain, one can choose to learn a Bayesian network without latent variables. How would such a network compare with the latent tree model found by HSHC? To answer this question, Zhang and Kocka also analyzed the data set using the GES algorithm [27] for learning Bayesian networks. The structure of the resulting model is shown in Fig. 6. We see that the network structure is less interpretable than the structure of the latent tree model found by HSHC.

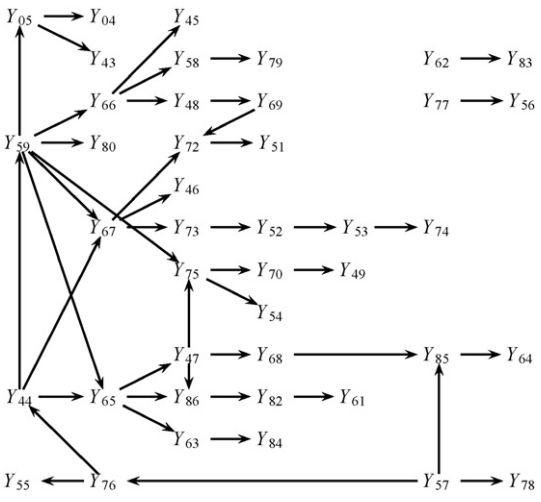


Figure 6 Bayesian network model found by GES for the CoIL Challenge 2000 data set.

When the model is used to guide the CoIL prediction task, only 83 policy owners were identified. This is significantly better than the random selection, but significantly worse than the selection based on the latent tree model.

6. Analysis of kidney deficiency data

We propose a new machine learning approach to the study of TCM syndrome differentiation. It is called the *latent structure approach*. The idea is to first collect patient data systematically and then perform multidimensional cluster analysis using latent structure models. The goal is to find natural clusters in data that correspond well to TCM syndrome types and hence can be used to establish objective and quantitative standards for syndrome differentiation. To investigate the feasibility of the latent structure approach, we have used latent tree models to study a subdomain of TCM diagnosis, namely KIDNEY DEFICIENCY. We report this case study in the rest of the paper. In this section, we describe the domain, the data set, and the data analysis process.

Diagnosis and treatment in TCM are based on several theories. One of the theories is called ZANG-FU theory. It explains the physiological functions, pathological changes, and mutual relationships of the TCM 'vital organs', which include HEART, LUNG, KIDNEY, and so on. Although minor similarities exist, the vital organs of the body in TCM are different from the anatomical organs of western medicine. They are classified according to their functions and functional entities. The main physiological functions of KIDNEY, for instance, are: (1) storing YIN-ESSENCE; (2) controlling water metabolism; (3) receiving QI; (4) controlling human reproduction, growth and development; (5) producing marrow, controlling the bones, manufacturing blood and influencing hair luster; (6) opening into the ear; (7) controlling the urinary bladder. Most KIDNEY DISEASES are due to the lack of essence, a condition known as KIDNEY DEFICIENCY. KIDNEY DEFICIENCY includes several subtypes, namely KIDNEY YANG DEFICIENCY, KIDNEY YIN DEFICIENCY, KIDNEY ESSENCE INSUFFICIENCY, and so on [6]. Common symptoms of KIDNEY DEFICIENCY include lumbago, sore and weak lumbus and knees, tinnitus, deafness, loss of hair, impotence, irregular menstruation, edema, abnormal defecation and urination, and so on.

The first step in the latent structure approach is data collection, and the first step in data collection is to determine its coverage in terms of symptom variables. In consultation with the China national standards on clinic terminology of TCM syndromes [28] and some textbooks on TCM diagnosis [6,29], we have selected 67 symptoms variables for our study.

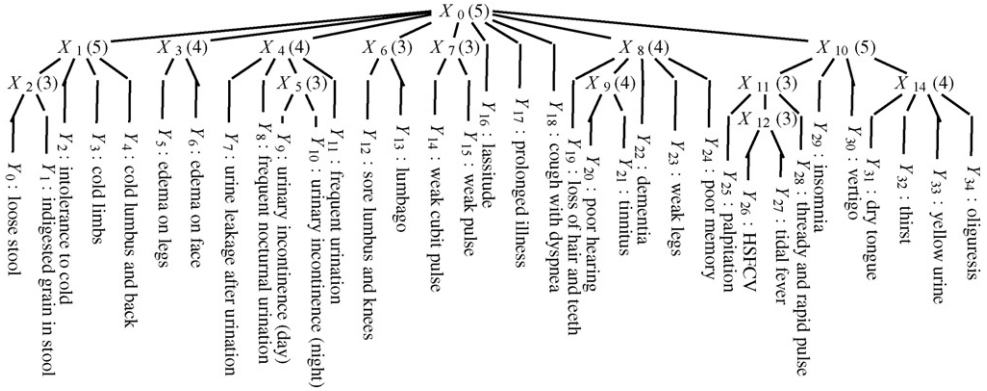


Figure 7 The structure of the best model $\mathcal{M}^*$ found for the KIDNEY data set. The symptom variables are at the leaf nodes and the latent variables are at the internal nodes. The numbers in parentheses are the numbers of states of the latent variables. The abbreviation HSFCV stands for hot sensation in the five centers with vexation, where the five centers refer to the centers of two palms, the centers of two feet, and the heart.

Those variables cover all aspects of the aforementioned physiological functions of KIDNEY. Each of the variables has four possible values, namely 'no', 'light', 'medium', and 'severe'.

TCM diagnosis consists of two steps, patient information gathering and syndrome differentiation. Subjectivity is an issue in both steps. The focus of our research is on the second step. To minimize the influence of subjectivity in the first step, we adopted, during data collection, the operational standards for determining the severity levels of KIDNEY DEFICIENCY symptoms by Yan et al. [30].

A total of 2600 data cases were collected. The data set was collected from communities in several regions in China and all the subjects were at or above the age of 60 years. The sample space was set this way because KIDNEY DEFICIENCY is more common among senior people. The entire data collection process was managed personally by the second co-author and a variety of measures were taken to ensure data quality.

We attempted to analyze the whole KIDNEY data using the HSHC algorithm. Despite being the most efficient algorithm for learning latent tree models[3], HSHC was not able to handle all 67 symptom variables. We were forced to reduce the number of variables and only 35 variables were included in the analysis. This limits the applicability of the results in practice. However, it does not prevent us from drawing conclusions about the feasibility of the latent structure approach. The approach is possible if (1) there exist natural clusters in data that correspond to TCM syndrome types, and (2) it can discover such clusters from data. We can demonstrate those two points using a data set of 35 variables as well as using a data set of 67 variables.

The analysis was conducted on a 2.4 GHz Pentium 4 machine and it took 98.5 h to finish. The BIC score of the resultant model is $-73{,}947$. HSHC is a hill-climbing algorithm. Like most hill-climbing algorithms, it might get stuck at local maxima. To detect whether HSHC was trapped at a local maximum and, if this was the case, to help it escape from the local maximum, we modified the initial result based on domain knowledge and continued the search with HSHC from the modified models. This resulted in another model, denoted by $\mathcal{M}^*$, with BIC score $-73{,}860$. Further efforts to improve $\mathcal{M}^*$ did not result in better models. We hence regard $\mathcal{M}^*$ as the best model for the data set. The structure of $\mathcal{M}^*$ is given in Fig. 7.

7. Interpretation of latent variables

Model $\mathcal{M}^*$ consists of 14 latent variables, X_0 to X_{13}. This means that our analysis has identified 14 latent factors from the KIDNEY data set. Each of the latent variables has a specific number of states. For example, the variable X_1 has five states. This means that we have grouped the data set into five clusters according to the latent factor. The variable X_{13} has four states. This means that we have grouped the data set in another way into four clusters. Thus, we have simultaneously clustered the data set in multiple ways.

Do the clusters correspond to TCM syndrome types? This question will be answered in the next section. As a preparatory step, we examine the meanings of the latent variables. In this process, we need to compare the structure of model $\mathcal{M}^*$ with the TCM theory on KIDNEY, which is itself a latent

[3] The situation is likely to change soon due to some on-going work.

structure model described in natural language. We rely on textbooks [6,29] for description of the TCM theory.

7.1. Variable interpretation

We start from the lower left corner of $\mathcal{M}^*$. Here the model states that there is a latent variable X_1 that: (1) directly influences the three symptoms 'intolerance to cold', 'cold limbs', and 'cold lumbus and back' and (2) through another latent variable X_2 indirectly influences 'loose stool' and 'indigested grain in stool'. Do those match the TCM theory on KIDNEY? What are the meanings of the latent variables X_1 and X_2?

According to TCM theory, KIDNEY YANG is the basis of all YANG in the body. When KIDNEY YANG is in deficiency, it cannot warm the body and the patient feels cold, resulting in manifestations such as cold lumbus and knees, intolerance to cold, and cold limbs. Deficiency of KIDNEY YANG also leads to spleen disorders, resulting in symptoms such as loose stool and indigested grain in stool.

Both of the above two paragraphs speak of two latent factors and some symptoms. The relationships between the latent factors and the symptoms are almost identical in the two cases. Therefore, there is a good match between the lower left corner of $\mathcal{M}^*$ and the relevant part of the TCM KIDNEY theory. The latent variables X_1 can be interpreted as KIDNEY YANG deficiency, while X_2 can be interpreted as SPLEEN DISORDERS DUE TO KYD, where KYD stands for KIDNEY YANG DEFICIENCY.

To the right of X_1, model $\mathcal{M}^*$ states that there is a latent variable X_3 that influences the symptoms 'edema on legs' and 'edema on face'. On the other hand, TCM theory maintains that when KIDNEY YANG is in deficiency, water is out of control. It flows to the skin and brings about edema. We see a perfect match here. The latent variable X_3 can be interpreted EDEMA DUE TO KYD.

To the right of X_3, model $\mathcal{M}^*$ states that there is a latent variable X_4 that: (1) directly influences the symptoms 'urine leakage after urination', 'frequent urination', and 'frequent nocturnal urination', and (2) through another latent variable X_5 indirectly influences 'urinary incontinence (day)' and 'urinary incontinence (night)'. On the other hand, TCM theory maintains that when KIDNEY fails to control the urinary bladder, one would observe clinical manifestations such as frequent urination, urine leakage after urination, frequent nocturnal urination, and in severe cases urinary incontinence. Once again, there is a fairly good match between this part of $\mathcal{M}^*$ and the relevant part of the TCM KIDNEY theory. The latent variable X_4 can be interpreted as KIDNEY FAILING TO CONTROL UB, where UB stands the urinary bladder.

According to TCM theory, the clinical manifestations of KIDNEY ESSENCE INSUFFICIENCY include premature baldness, tinnitus, deafness, poor memory, trance, declination of intelligence, fatigue and weakness. Those match the symptom variables under X_8 fairly well and hence X_8 can be interpreted as KIDNEY ESSENCE INSUFFICIENCY. The clinical manifestations of KIDNEY YIN DEFICIENCY include dry throat, tidal fever or hectic fever, fidget, hot sensation in the five centers, and yellow urine. Those match the symptom variables under X_{10} fairly well and hence X_{10} can be interpreted as KIDNEY YIN DEFICIENCY. Finally, TCM theory maintains that there are several subtypes of KIDNEY DEFICIENCY, namely KIDNEY YANG DEFICIENCY, KIDNEY ESSENCE INSUFFICIENCY, KIDNEY YIN DEFICIENCY, and so on. They share common symptoms such as lumbago, sore and weak lumbus and knees, mental and physical fatigue. Moreover, KIDNEY DEFICIENCY can be caused by prolonged illness. Those and the topology of $\mathcal{M}^*$ suggest that X_0 should be interpreted as KIDNEY DEFICIENCY.

Table 2 summarizes the meanings of the latent variables.

7.2. Remarks

All symptom variables in our case study are those that a TCM doctor would consider when making diagnostic decisions about KIDNEY DEFICIENCY. There is hence no surprise that one of the latent variables in $\mathcal{M}^*$ can be interpreted as KIDNEY DEFICIENCY. However, it is very interesting that some of the latent variables correspond to syndrome factors such as KIDNEY YANG/YIN DEFICIENCY, as each of them is associated with only a subset of the symptom variables. Take KIDNEY YANG DEFICIENCY as an example. TCM claims that it can cause symptoms such as 'intolerance to cold', 'cold limbs', and 'cold lumbus and back'. On the other hand, the result of our analysis states that, judging from data, there should be a latent factor that directly influences those symptom variables. In this sense, our analysis has validated the TCM claim. This is important because TCM theories are sometimes discarded as being unscientific and even groundless. Our work has shown that there are

Table 2 Interpretations of some of latent variables in $\mathcal{M}^*$

X_0: KD (KIDNEY DEFICIENCY)
X_1: KYD (KIDNEY YANG DEFICIENCY)
X_3: EKDY (EDEMA DUE TO KYD)
X_4: KFCUB (KIDNEY FAILING TO CONTROL)
X_8: KEI (KIDNEY ESSENCE INSUFFICIENCY)
X_{10}: KYD (KIDNEY YIN DEFICIENCY)

scientific (more precisely, statistically valid) contents in TCM theories.

It should be noted that the scope of our case study is KIDNEY DEFICIENCY, which is determined by the selection of symptom variables. Hence, the results should be helpful in determining whether a patient is suffering from KIDNEY DEFICIENCY and if so which subtype. Table 2 shows that model $\mathcal{M}^*$ is indeed useful in this aspect. What should we do if we want to determine whether a disease is with KIDNEY, or LIVER, or other vital organs? The answer is to build another model with symptom variables that can help this higher level differentiation task.

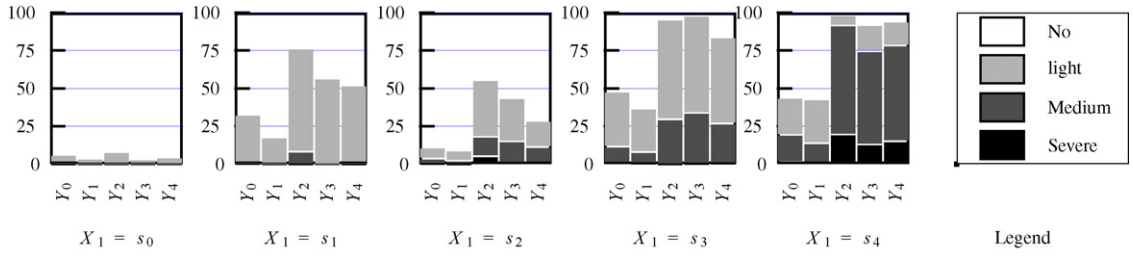
It should also be pointed out that there are some aspects of $\mathcal{M}^*$ that do not match TCM theory well. As one example, consider the symptoms 'tinnitus' and 'poor memory'. According to TCM theory, they can be caused both by KIDNEY YIN DEFICIENCY and KIDNEY ESSENCE INSUFFICIENCY. In $\mathcal{M}^*$, however, they are directly connected to only X_8, but not to X_{10}. This mismatch is due to the restriction that in latent tree models a manifest variable can be connected to only one latent variable. Another mismatch is in the scope: $\mathcal{M}^*$ involves fewer symptom variables than the TCM theory on KIDNEY. This is due to two reasons. First, we were not able to collect data about human reproduction because our subjects did not feel comfortable talking about it. Second, some symptom variables were excluded from data analysis for the sake of computational efficiency. The consequence is twofold. First, many symptoms, especially those about human reproduction, are absent from the discussions above. They are mentioned in TCM theory, but do not appear in the model $\mathcal{M}^*$. Second, two other common subtypes of KIDNEY DEFICIENCY, namely UNCONSOLIDATION OF KIDNEY QI and FAILURE OF KIDNEY TO RECEIVE QI are absent from $\mathcal{M}^*$.

Finally, we would like to make a remark about edge orientations in latent tree models. As explained in Section 3, edge orientations cannot be determined during data analysis. However, they sometimes can be determined during model interpretation. As an example, consider the edges between the latent variable X_1 and the symptom variables Y_2, Y_3, and Y_4. X_1 is interpreted as KIDNEY YANG DEFICIENCY, which according to TCM theory causes 'intolerance to cold', 'cold limbs', and 'cold lumbus and back'. Therefore, the edges can be oriented and they point from X_1 to the symptom variables. As a matter of fact, all edges in $\mathcal{M}^*$ can be oriented and they all point downward. There is only one exception, namely the edge between X_0 and Y_{17}.

8. Interpretation of latent clusters

Each of the latent variables in model $\mathcal{M}^*$ defines a number of clusters. In this section, we examine the meaning of some of the clusters and check to see whether they correspond to TCM syndrome types. We will focus primarily on X_1 and X_4.

The latent variable X_1 was interpreted as KYD (KIDNEY YANG DEFICIENCY) and it has five states. This means that the data has been grouped into five clusters according to KIDNEY YANG DEFICIENCY. Fig. 8 shows the probability distributions of the five symptom variables Y_0 to Y_4 in each of the clusters. There is one bar diagram for each cluster. In each diagram, there are five bars, each corresponding to one of the five symptom variables. The bars consist of up to



Cluster	Most Typical Members					$P(X_1 \mid Y_0, Y_1, Y_2, Y_3, Y_4)$				
	Y_0	Y_1	Y_2	Y_3	Y_4	$X_1 = s_0$	$X_1 = s_1$	$X_1 = s_2$	$X_1 = s_3$	$X_1 = s_4$
$X_1 = s_0$	0	0	0	0	0	.90	.03	.07	.00	.00
$X_1 = s_1$	1	0	1	0	1	.01	.94	.03	.02	.00
$X_1 = s_2$	0	0	3	0	0	.00	.00	.99	.00	.01
$X_1 = s_3$	1	0	1	2	1	.00	.00	.02	.97	.01
$X_1 = s_4$	3	3	3	3	3	.00	.00	.00	.00	1.0

Figure 8 Characteristics of the five clusters identified by the states of the latent variable X_1: the diagrams show the probability distribution of the symptom variables Y_0, Y_1, Y_2, Y_3, and Y_4 in each of the clusters. The table shows the most typical members of the clusters, together with the probability distribution of X_1 for those members. In the table, symptom severity levels are indicated using integers: 0—no; 1—light; 2—medium; 3—severe.

four segments, each corresponding to one state of the symptom variables. The clusters were ordered and named after an initial visual inspection of those bar diagrams.

The probability bar diagrams provide an overall characterization of the clusters, but they do not provide any information about individual members of the clusters. To compensate for this lack of specificity, we also provide information about the most typical members of the clusters. The *most typical member* of the cluster $X_1 = s_0$, for instance, is the configuration of the states of the variables Y_0, Y_1, Y_2, Y_3, and Y_4 that maximizes the probability $P(X_1 = s_0|Y_0, Y_1, Y_2, Y_3, Y_4)$. This configuration might not be unique. Nonetheless, we will use the term 'the most typical member' for simplicity.

We can now digest the meanings of the clusters defined by X_1. Five symptoms are involved here, namely LS (loose stool), (IGS) (indigested grain in stool), IC (intolerance to cold), CL (cold limbs), and CLB (cold lumbus and back). In the cluster $X_1 = s_0$, the five symptoms almost never occur. The most typical member of $X_1 = s_0$ has none of those symptoms. Hence, the clusters means '*no* KYD'.

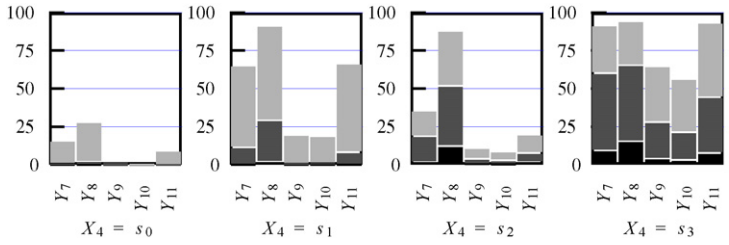
Next, consider the clusters $X_1 = s_1$ and $X_1 = s_2$. In both clusters, the five symptoms have substantial probability of occurring. The overall probability of the symptoms occurring is higher in $X_1 = s_1$, while the probabilities of the symptoms occurring at the medium or severe levels are higher in $X_1 = s_2$. For the most typical member of $X_1 = s_1$, three of the symptoms, namely LS, IC and CLB, occur only at the light level, while for the most typical member of $X_1 = s_2$, only one of the symptoms, namely IC, occurs at the severe level. Therefore, both clusters can be interpreted as 'light KYD'. To differentiate their characteristics, we label $X_1 = s_1$ as '*light KYD (1)*' and label $X_1 = s_2$ as '*light KYD (2)*'.

In the clusters $X_1 = s_3$ and $X_1 = s_4$, the overall probability of the five symptoms occurring is high, while the probabilities of their occurring at the medium and severe levels in $X_1 = s_4$ are much greater than those in $X_1 = s_3$. For the most typical member of $X_1 = s_3$, three of the symptoms occur at the light level and one occurs at the medium level. For the most typical member of $X_1 = s_4$, all the five symptoms occur at the severe level. Hence, $X_1 = s_3$ can be interpreted as '*medium*KYD' and $X_1 = s_4$ can be interpreted as '*severe* KYD'.

There is one interesting observation that we would like to make about the class probability distributions. In the last four of the clusters, the probabilities of the symptoms LS (Y_0) and IGS (Y_1) occurring are much lower than those of the other three symptoms. This is consistent with TCM theory. According to the latter, KYD would not affect other vital organs when at benign stage. As the severity increases, however, it might affect other vital organs. In particular, it might lead to spleen disorders, resulting in the symptoms LS and IGS. Hence, the symptoms IC, CL and CLB are more common than LS and IGS. IC, CL and CLB can occur as soon as KYD is present, while LS and IGS occur only when KYD becomes relatively more severe. Our analysis has validated this piece of TCM theory and has provided a quantification for it.

We next consider the four clusters given by the latent variable X_4. The probability distributions of the symptom variables Y_7, Y_8, Y_9, Y_{11}, and Y_{11} in each of the clusters are given in Fig. 9. The typical members of the clusters are also shown. By using arguments similar to those used in the case of X_1, one can conclude that the clusters $X_4 = s_0$, $X_4 = s_1$, $X_4 = s_2$ and $X_4 = s_3$ can respectively interpreted as *no*KFCUB (KIDNEY FAILING TO CONTROL UB), '*light*KFCUB*(1)*', '*light*KFCUB*(2)*', and '*severe* KFCUB'.

As mentioned in the previous section, TCM theory maintains that when kidney fails to control the urinary bladder, one would observe clinical manifestations such as FU (frequent urination), ULU (urine leakage after urination), FNU (frequent nocturnal urination), and in severe cases UID (urinary incontinence (day)) and UIN (urinary incontinence (night)). Therefore, UID and UIN occur less frequently than the other three symptoms. This is



Cluster	Most Typical Members				
	Y_7	Y_8	Y_9	Y_{10}	Y_{11}
$X_4 = s_0$	0	0	0	0	0
$X_4 = s_1$	1	1	0	0	1
$X_4 = s_2$	0	3	0	0	0
$X_4 = s_3$	3	3	3	3	3

Figure 9 Characteristics of the four clusters identified by the states of the latent variable X_4: the diagrams show the probability distributions of the symptom variables Y_7, Y_8, Y_9, Y_{10}, and Y_{11} in each of the clusters. The table shows the most typical members of the clusters. In the table, symptom severity levels are indicated using integers: 0—no; 1—light; 2—medium; 3—severe.

Table 3 Interpretations of some of latent clusters in $\mathcal{M}^*$

	s_0	s_1	s_2	s_3	s_4
X_1	No KYD	Light KYD (1)	Light KYD (2)	Medium KYD	Severe KYD
X_3	No EKYD	Light EKYD (1)	Light EKYD (2)	Severe EKYD	—
X_4	No KFCUB	Light KFCUB (1)	Light KFCUB (2)	Severe KFCUB	—
X_8	No KEI	Light KEI (1)	Light KEI (2)	Severe KEI	—
X_{10}	No KID	Light KID	Medium KID (1)	Medium KID (2)	Severe KID

Refer to Table 2 for the meanings of the abbreviations.

consistent with the bar diagrams in Fig. 9. We see that the probabilities of UID (Y_9) and UIN (Y_{10}) occurring are always significantly lower than those of the other three symptoms. Once again, our analysis has validated a piece of TCM theory and has provided a quantification for it.

We have also examined the clusters identified by X_3, X_8, and X_{10}. The interpretations are given in Table 3.

9. Conclusions and future directions

TCM diagnosis consists of two steps, patient information gathering and syndrome differentiation. We propose a novel approach, namely the latent structure approach, to the study of syndrome differentiation. The long-term objective is to establish objective and quantitative standards for syndrome differentiation. The idea is to systematically collect patient data, perform multidimensional cluster analysis, and use the resultant clusters as the basis for syndrome differentiation.

To demonstrate the feasibility of the new approach, we have studied a special class of latent structure models called latent tree models. The study on latent tree models has spanned over several years. In this paper, we have provided a detailed survey of the work so far. It is intended to serve as a good entry point for those who want to use or further develop the methodology.

We have collected a data set about KIDNEY DEFICIENCY and have analyzed the data set using latent tree models. The resulting model matches relevant TCM theory well: the latent variables correspond to syndrome factors, and the latent clusters correspond to syndrome types. We have recently analyzed several other data sets from joint projects with Beijing University of Chinese Medicine and China Academy of Chinese Medicine. Equally good results were obtained.[4] Consequently, we have shown that the latent structure approach is indeed feasible.

For future work, there is obviously a need to develop more efficient algorithms for learning latent tree models. It would be also interesting to study latent structures beyond trees. In the TCM front, the immediate next step is to investigate how the results produced by the latent structure approach can actually be used to improve TCM diagnosis and treatment. One area is to start with the integration of TCM and west medicine. The idea here is to divide all patients suffering from a western medicine, e.g. irritable bowel syndrome, into several groups from the TCM perspective and apply different treatments to different groups. Doctors in China have been practicing this for decades. The latent structure approach can be applied to establish objective and quantitative standards for the grouping. TCM doctors and western medicine doctors can then come together, study the groups and form treatments accordingly.

[4] The results will be reported in forthcoming papers.

Acknowledgement

We thank Yang Weiyi and Wang Miqu for valuable discussions during the inception of this work in 2000. We also thank Yan Shilin, Wang Tianfang and many other TCM experts for helping with data collection in 2000 and model interpretation in 2004. Tomas Kocka, Finn Jensen, and Thomas Nielsen collaborated with us on developing algorithms for learning latent tree models and related models. We are grateful to Lai Shilong, Liu Baoyan, Wang Qingguo, Wang Jie, Wang Tianfang, and the anonymous reviewers for feedbacks on earlier versions of the paper. Research on this work was supported by Hong Kong Grants Council Grants #622105 and #622307, and The National Basic Research Program (aka the 973 Program) under project no. 2003CB517106.

References

[1] Sung JJY, Leung WK, Ching JYL, Lao L, Zhang G, Wu JCY, et al. Agreements among traditional Chinese medicine practitioners in the diagnosis and treatment of irritable bowel syndrome. Alimen Pharmacol Therapeut 2004;20:1205—10.

[2] World Health Organization, WHO traditional medicine strategy. http://www.who.int/medicines/publications/traditionalpolicy/en/index.html; 2002—2005 [accessed: 14.10.07].

[3] Wang HX, Xu YL. The current state and future of basic theoretical research on traditional Chinese medicine. Beijing: Military Medical Sciences Press; 1999.
[4] Feng Y, Wu Z, Zhou X, Zhou Z, Fan W. Knowledge discovery in traditional Chinese medicine: state of the art and perspectives. Aritif Intell Med 2006;38:219—36.
[5] Liang MX, Liu J, Hong ZP, Xu YY. Perplexity of TCM syndrome research and countermeasures. Beijing: People's Health Press; 1998.
[6] Yang W, Meng F, Jiang Y. Diagnostics of traditional Chinese medicine. Beijing: Academy Press; 1998.
[7] Lazarsfeld PF, Henry NW. Latent structure analysis. Boston: Houghton Mifflin; 1968.
[8] Bartholomew DJ, Knott M. Latent variable models and factor analysis, Kendall's Library of Statistics 7, 2nd ed., London: Arnold; 1999.
[9] Vermunt JK, Magidson J. Latent class cluster analysis. In: Hagenaars JA, McCutcheon AL, editors. Advances in latent class analysis. Cambridge University Press; 2002. p. 89—106.
[10] J. Uebersax, A practical guide to local dependence in latent class models, http://ourworld.compuserve.com/homepages/jsuebersax/condep.htm [accessed 14.10.07].
[11] Hagenaars JA. Latent structure models with direct effects between indicators: local dependence models. Sociol Meth Res 1988;16:379—405.
[12] Goodman LA. Exploratory latent structure analysis using both identifiable and unidentifiable models. Biometrika 1974;61:215—31.
[13] Zhang NL. Hierarchical latent class models for cluster analysis. In: Dechter R, Kearns M, Sutton AAAI R, editors. Proceedings of 18th National Conference on Artificial Intelligence. 2002. p. 230—7.
[14] Zhang NL. Hierarchical latent class models for cluster analysis. J Machine Learn Res 2004;5(6):697—723.
[15] Pearl J. Probabilistic reasoning in intelligent systems: Networks of plausible inference. Palo Alto: Morgan Kaufmann; 1988.
[16] Durbin R, Eddy S, Krogh A, Mitchison G. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. Cambridge University Press; 1998.
[17] Green P. Penalized likelihood. In: Encyclopedia of Statistical Sciences, Update Volume 3. John Wiley & Sons; 1999 . p. 578—86.
[18] Geiger D, Heckerman D, Meek C. Asymptotic model selection for directed networks with hidden variables. In: Horvitz E, Jensen FV, editors. Proceedings of the 12th Annual Conference on Uncertainty in Artificial Intelligence. 1996. p. 158—68.
[19] Kocka T, Zhang NL. Dimension correction for hierarchical latent class models. In: Breese J, Koller D, editors. Proceedings of the 18th Conference on Uncertainty in Artificial Intelligence (UAI-02). Morgan Kaufmann; 2002 . p. 267—74.
[20] Zhang NL, Kocka T. Efficient learning of hierarchical latent class models. In: Khoshgoftaar TM, editor. Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-2004). 2004. p. 585—93.
[21] Schwarz G. Estimating the dimension of a model. Ann Stat 1978;6(2):461—4.
[22] Zhang NL, Kocka T. Effective dimensions of hierarchical latent class models. J Artif Intell Res 2004;21:1—17.
[23] Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc B 1997;39:1—38.
[24] Lauritzen SL. The EM algorithm for graphical association models with missing data. Comput Stat Data Anal 1995;19: 191—201.
[25] Friedman N. Learning belief networks in the presence of missing values and hidden variables. In: Fisher DH, editor. Proceedings of the 14th International Conference on Machine Learning (ICML 1997). 1997. p. 125—33.
[26] van er Putten P, van Someren M, editors. CoIL Challenge 2000: The insurance company case. Amsterdam: Sentient Machine Research; 2002.
[27] Chickering DM. Learning equivalence classes of Bayesian-network structures. J Machine Learn Res 2002;2: 445—98.
[28] China State Bureau of Technical Supervision, National standards on clinic terminology of traditional Chinese Medical diagnosis and treatment—Syndromes, GB/T 16751.2—1997. Beijing: China Standards Press; 1997.
[29] Zhu WF. Diagnostics of traditional Chinese medicine. Shanghai Science Press; 1995.
[30] Yan SL, Zhang LW, Wang MH, Yuan SH. Operational standards for determining the severity levels of kidney deficiency symptoms. J Chengdu Univ Chin Med 2001;24(1):56—9.