An Evidence-Based Approach to Patient Classification in Traditional Chinese Medicine based on Latent Tree Analysis

Nevin L. Zhang, Ph.D.: The Hong Kong University of Science and Technology; E-mail:lzhang@cse.ust.hk

Chen Fu, M.D.: Dongfang Hospital, Beijing University of Traditional Chinese Medicine; E-mail: fuchen.bucm@gmail.com

Teng Fei Liu, B.S.: The Hong Kong University of Science and Technology; E-mail: liutf@cse.ust.hk

Kin Man Poon, Ph.D.: The Hong Kong Institute of Education; E-mail: kmpoon@ied.edu.hk

Pei Xian Chen, B.S.: The Hong Kong University of Science and Technology; E-mail: pchenac@cse.ust.hk

Bao Xin Chen, M.D.: Dongfang Hospital, Beijing University of Traditional Chinese Medicine; E-mail: chenbaoxin2008@163.com

Yun Ling Zhang, M.D.: Dongfang Hospital, Beijing University of Traditional Chinese Medicine; E-mail: yunlingzhang2004@163.com

Corresponding Author: Nevin L. Zhang, Department of Computer Science and Engineering, The Hong Kong University of Science and Technology; E-mail:lzhang@cse.ust.hk
Abstract

Objective: The efficacy of traditional Chinese medicine (TCM) treatments of western medicine diseases relies heavily on the proper sub-classification of the patients from the TCM perspective in a process known as syndrome differentiation. We develop an evidence-based method, called the latent tree analysis, for solving the sub-classification problem where definitions of patient subclasses and classification rules are established based on patterns detected in clinic symptom data.

Methods: The approach starts with a survey of patients with a western medicine disease where information about symptoms and signs of interest to TCM is collected. The data are analyzed using latent tree models to reveal symptom co-occurrence/mutual-exclusion patterns, which are represented by latent variables. The patterns are then used to perform clustering analysis of the patients. The resulting patient clusters are used to define patient subclasses and to establish classification rules.

Results: The approach is illustrated using a data set about vascular mild cognitive impairment that involves 803 patients and 93 symptoms. A latent tree model with 31 latent variables is obtained. The patients are clustered based on a combination of eight of latent variables that are related to qi deficiency. A quantitative definition of the qi deficiency subclass and the associated classification rule are established.

Conclusions: An evidence-based approach to TCM syndrome classification is presented. The approach can be used to answer the following questions about a western medicine disease: What TCM syndrome subclasses are there among the patients with the disease? What are the sizes of the subclasses? What are the statistical characteristics of each subclass? How can we determine whether a particular patient belongs to a specific subclass?

Keywords:

Traditional Chinese medicine, syndrome differentiation, symptom co-occurrence patterns, latent tree analysis, patient clustering, patient subclass definition, classification rules
1. Introduction

Traditional Chinese Medicine (TCM) is increasingly used in human healthcare as complementary or alternative to Western Medicine (WM). A common practice is to divide the patients of a WM disease into several subclasses based on symptoms and signs (both referred as symptoms henceforth for simplicity) that TCM is concerned with, and to apply different TCM treatments to patients in different subclasses. The efficacy of TCM treatment depends heavily on whether the patient sub-classification is done properly.

The patient sub-classification step is called syndrome differentiation and the results are called TCM syndrome classes or Zheng in Chinese. The TCM syndrome classification problem associated with a WM disease consists of four sub-problems: (1) What TCM syndrome subclasses are there among the patients with the disease? (2) What are the sizes of the subclasses? (3) What are the characteristics of each subclass in terms of symptom occurrence probabilities? (4) How do we determine to which subclass or subclasses a particular patient belongs to based on his symptoms? The four sub-problems will be referred to as syndrome composition, syndrome prevalence, subclass definition, and classification rule respectively.

Syndrome classification standards have been published for various WM diseases [1,2]. Such standards contain information about the syndrome compositions of WM diseases, provide for each patient subclass a list the symptoms that are likely to occur, and highlight the symptoms that are the most important for patient sub-classification. There are typically no symptom occurrence probabilities, no quantitative information about syndrome prevalence, and no clearly stated classification rules. In addition, those standards were all set up by panels of experts.

Clinic research on syndrome classification has also been conducted for various WM diseases [3,4,5]. In such a study, the patients of a WM disease are surveyed and information about symptom occurrence is collected. The patients are examined by TCM doctors and their syndrome types are determined. In other words, the data have class labels. Statistics are then
calculated to determine syndrome prevalence and symptom occurrence probabilities for each syndrome subclass. Classification rules are established using statistical and machine learning techniques such as regression, neural networks and support vector machines [6]. Abstractly speaking, the data analysis method used this kind of research is supervised learning and the conclusions are summaries of the behaviors of the TCM doctors who participate in the studies.

In the past decade, a new approach to TCM syndrome classification has emerged in the literature [7-17]. It is called the latent tree analysis approach. The objective is to make TCM patient sub-classification as evidence-based as possible. It starts with data on symptom occurrence. Unlike in the case of the previous paragraph, judgments about syndrome types are not included. In other words, the data do not have class labels. Instead of supervised learning, unsupervised learning method is used for data analysis. Specifically the data are analyzed using latent tree model to identify symptom co-occurrence patterns, and the patterns are used to group patients into possibly overlapping clusters. Those clusters are then used to define patient sub-classes and to establish classification rules.

The paper presents a systematic review of the latent tree analysis approach. We strive to achieve completeness, clarity and cleanness. By completeness we mean that all the key components of the approach will be covered and the paper will be as self-contained as possible. By clarity we mean that the materials will be explained in details using simple language so that TCM researchers can easily follow the paper and can apply the method after reading this paper. By cleanness we mean that points once considered important in the original literature but actually not essential and potentially misleading will be excluded from the presentation. In addition, several important improvements to the approach are introduced.

Due to the nature of the materials, it is difficult to organize the paper in the usual material-method-result-conclusion format. It is instead organized as follows. In Section 2 we describe a data set on vascular mild cognition impairment (VMCI) that will be used for illustration throughout the paper. In Section 3 we explain the data analysis methods to be used in this paper, namely latent tree analysis (LTA) and latent class analysis (LCA). In Section 4
we present the results of LTA on the VMCI data and highlight the fact that, from an application viewpoint, LTA is a tool for detecting symptom co-occurrence and symptom mutual exclusion patterns. The patterns are represented by latent variables. In Section 5 we discuss the interpretation of latent variables to determine their statistical meanings and TCM connotations, where expert judgments are required. In Section 6 we show how to divide patients into clusters using LCA with selected latent variables as features. Patient subclass definitions are thereby obtained. In Section 7 we investigate the establishment of classification rules for the resulting patient subclasses. Finally, we end the paper with discussions of related work in Section 8 and concluding remarks in Section 9.

2. A Data Set about VMCI

The data set used in this paper came from a cross-sectional survey conducted between February 2008 and February 2012 at seven hospitals and nearby communities from Northern China. Included in the study were subjects aged 50 or above who met a set criteria for VMCI that were set up based on the NINDS-CSN Vascular Cognitive Impairment Harmonization Standards, the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-R), the Mini–Mental State Examination (MMSE), the Montreal Cognitive Assessment, and the Clinical Dementia Rating (CDR) [18]. Excluded from the study were subjects who had difficulties in communication or in participating psychological assessments, and subjects who scored 17 or above on the Hamilton Rating Scale for Depression (HAM-D). A total number of 803 were recruited for the study at the end.

A checklist of 93 symptoms was used in the survey. The list can be found in [18] and all the symptoms also appear in Figure 3. A questionnaire was completed for each subject where questions were asked whether each of the symptoms. The final data consists of 93 columns, each corresponding to a symptom, and 803 rows, each corresponding to a subject in the study. The values in the data set are either 0 or 1, which indicate the absence and presence of a symptom on a subject respectively.

3. Data Analysis Methods

The data analysis methods used in this paper are based on probabilistic models that describe relationships among discrete variables. Some of the variables are observed, while the others are latent, that is, unobserved.
3.1. Latent Tree Models and Latent Class Models

The models that we use are called latent tree models (LTMs). An LTM describes the relationship among a set of variables at two levels. At the qualitative level, it is an undirected tree where the observed variables are located at the leaf nodes, whereas the latent variables are located at the internal nodes. At the quantitative level, it describes the relationship between each pair of neighboring variables using a conditional probability distribution.

Figure 1(a) shows an example LTM taken from [9]. Qualitatively, it asserts that a student’s Math grade (MG) and Science grade (SG) are influenced by his analytical skill (AS); his English grade (EG) and History grade (HG) are influenced by his literacy skill (LS); and the two skills are correlated. Here, the grades are observed variables, while the skills are latent variables.

For simplicity, assume all the variables have two possible values ’low’ and ’high’. The dependence of MG on AS is quantitatively characterized by the conditional distribution \( P(MG|AS) \), which is also shown in Figure 1. It says that a student with high AS tends to get high MG and a student with low AS tends to get low MG. Similarly, the dependences of other grade variables on the skill variables are quantitatively characterized by the distributions \( P(SG|AS) \), \( P(EG|LS) \) and \( P(HG|LS) \) respectively. They are not shown to save space.

To specify the quantitative relationships among the latent variables, it is convenient to root the model at one of the latent variables and regard it as a directed model --- a tree-structure Bayesian network [19]. If we use AS as the root, then we need to provide, as given in Figure 1, the marginal distribution of the root \( P(AS) \) and the distribution \( P(LS|AS) \) of LS conditioned on its parent AS. If LS were chosen as the root instead, we would need to provide \( P(LS) \) and \( P(AS|LS) \). The choice of root does not matter because different choices give rise to equivalent directed models [20].

Latent tree models with a single latent variable are called latent class models (LCMs). Figure 1(b) shows an example LCM, where Intelligence is the sole latent variable. Qualitatively, it asserts that a student’s grades in the four subjects are all influenced by his intelligence level.

Different models make different independence assumptions. In Figure 1(b), the four grade variables are mutually independent conditioned on the latent variable Intelligence. This is known as the local independence assumption. In Figure (a), MG and SG are independent conditioned on the latent variable AS, but EG and HG are not. Similarly, EG and HG are independent conditioned on the latent variable LS, but MG and SG are not.

Historically, LCMs predate LTMs. LTMs are introduced as a generalization of LCMs in
[20], where they are called hierarchical latent class models. A key motivation for the introduction of LTM}s is to analyze TCM symptom data [1, 8, 11].

3.2. Latent Class Analysis

*Latent class analysis (LCA)* refers to the analysis of data using latent class models. As an example, consider a data set about the grades that students from a school obtain on the aforementioned four subjects. To perform LCA on the data, we assume there is a latent variable \( Y \) that is related to the grade variables as shown in Figure 2 (a). The task is to: (1) determine the number of possible values for \( Y \), (2) determine the marginal distribution \( P(Y) \) and the distributions \( P(MG|Y), P(SG|Y), P(EG|Y) \) and \( P(HG|Y) \) of the grade variables conditioned on \( Y \). The first item is known as *model selection*, while the second as *parameter estimation*.

In Statistics, the concept of *likelihood* measures how well a model fits data. In LCA, probabilistic parameters are determined using the *maximum likelihood estimate (MLE) principle* [21]. The number of possible values for \( Y \) is often determined using the *Bayes Information Criterion (BIC)* [22]. The BIC score is likelihood plus a penalty term for model complexity. The use of BIC intuitively means that we want a model that fits data well, but do not want it to be overly complex.

Note that there are two equivalent versions of BIC that are negations of each other. In one version, the BIC score takes positive values. Here we want to minimize the BIC score. In the other version it takes negative values. In this case, we want to maximize the BIC score.

In practice LCA is used as a tool for clustering discrete data [23]. Each value of the latent variable \( Y \) represents a probabilistic cluster of individuals and all the values collectively give a *partition* of all individuals. To determine the number of possible values for \( Y \) is to determine the number of clusters, and to determine the probabilistic parameters is to determine the statistical characteristics of the clusters.

LCA is used in WM research to identify subtypes in a group of patients. It has been used to study rheumatoid arthritis [24], chronic fatigue [25], and major depressive disorders [26]. Recently, it has been used to identify the TCM syndrome subclasses among psoriatic patients [27].

3.3. Latent Tree Analysis

*Latent tree analysis* refers to the analysis of data using latent tree models. For a given
data set, there are many possible LTMs. For example, one possible LTM for the student grade
data is shown in Figure 2(b), another in Figure 2(a), and there are also other possible models.
The task is to determine which model is the best for the data. Specifically, we need to
determine: (1) the number of latent variables, (2) the number of possible values for each latent
variable, (3) the connections among the latent and observed variables, and (4) the probability
parameters. The model selection problem here is more difficult than in the case of LCA. It
consists of the first three items.

Several algorithms have been developed for LTA [28]. Extensive empirical studies
have been conducted where the different algorithms are compared in terms of the BIC scores
of the models they obtain and running time [28,29]. The experiments indicate that the EAST
(Extension-Adjustment-Simplification-until-Termination) algorithm [30] finds the best
models on data sets with dozens to around one hundred observed variables, while the BI
(Bridged-Islands) algorithm [29] finds the best models on data sets with hundreds to around
one thousand observed variables. On data sets with dozens to around one hundred variables,
BI is much faster than EAST, while the models it obtains are sometimes inferior to those
found by EAST. EAST is unable to deal with data sets with hundreds or more observed
variables.

The LTM shown in Figure 2(b) can be viewed as two LCMs with their latent variables
connected by an edge. In general, an LTM can be regarded as a collection of LCMs, where
each LCM is based on a distinct subset of observed variables and the latent variables are
connected to form a tree structure. As pointed out earlier, an LCM gives one probabilistic
partition of data. Consequently, an LTM gives multiple partitions of data. Each partition is
based primarily on a distinct subset of observed variables and the different partitions are
correlated. For this reason, LTA is a tool for multidimensional clustering [30]. Multidimensional clustering is what we need when analyzing TCM symptom data because
such data are complex and have different natural facets to them. The patient population can
be meaningfully partitioned in multiple ways along the facets.

3.4. Software Tools

A software tool called Lantern (http://www.cse.ust.hk/faculty/lzhang/ltm/index.htm) has
been developed to facilitate LCA and LTA. The software is designed to run on the desktop of
personal computers. The user can use it to analyze data using various algorithms, inspect the
results, and perform further analyses to be described later in this paper. Separate
implementations of the EAST and BI algorithms are also provided so that users can run data
analysis on servers. On data sets that involve around 100 symptom variables and 1000
samples, EAST typically takes several days, while BI takes a few hours. We recommend that users run EAST on servers rather than on desktops.

4. Latent Tree Analysis of the VMCI Data

We have performed LTA on the VMCI data using the EAST algorithm. The structure of the resulting model is shown in Figure 3. The variables labeled with English phrases are symptom variables, which originate from the data set. Each symptom variable has two possible values 0 and 1, representing the absence and presence of the symptom respectively. The $Y$-variables are latent variables, which are introduced during data analysis. The integer next to a latent variable is the number of its possible values. For example, $Y1$ has 2 possible values, whereas $Y15$ has 3.

The widths of the edges indicate strengths of correlations between neighboring variables. Technically, they represent mutual information computed from model parameters for visualization. We see that $Y1$ is strongly correlated with the symptom variables Sallow_complexion, Asthenia_of_defecation, and Dry_stool_or_constipation, and weakly related to Clear_profuse_urination. Similarly, $Y08$ is strongly correlated with the latent variable $Y12$ and weakly related to $Y04$ and $Y13$.

Each latent variable represents a probabilistic partition of the patients surveyed. For example, $Y01$ has two possible values and hence partitions the patients into two clusters. Information about the partition is given in Table 1. $Y01$ is directly connected to four symptom variables. One of them, namely Clear_profuse_urination, is not included in the table because its relationship with $Y01$ is weak. A formal criterion for such exclusion will be given in Section 6.

We see that the two clusters consist of 83% and 17% of the patients respectively. The three symptoms asthenia of defecation, dry stool or constipation, and sallow complexion occur with high probabilities in the cluster $Y01=s1$ and with low probabilities in the cluster $Y01=s0$. Consequently, the three symptoms tend to co-occur in the cluster $Y01=s1$, whereas they do not in the cluster $Y01=s1$. As such, $Y01$ reveals the probabilistic co-occurrence of the three symptoms. This is the statistical meaning of $Y01$.

Information about four other partitions is given in Tables 2, 3, 4 and 5. It is clear that the $Y29$ reveals the probabilistic co-occurrence of hypodynamia and lassitude; the $Y08$ reveals the probabilistic co-occurrence of sunken pulse and feeble pulse; and the $Y25$ reveals the probabilistic co-occurrence of insomnia and dreamful sleep.

Table 5 shows the partition given by yet another latent variable $Y12$. In the cluster
Y12=s0, slippery pulse occurs with high probability and thin pulse does not occur at all. In the cluster Y12=s1, on the other hand, slippery pulse occurs with low probability and thin pulse occur with high probability. Those indicate that the two symptoms tend to be mutual exclusive. We say that Y12 reveals the probabilistic mutual exclusion of slippery pulse and thin pulse.

The above discussions suggest the following view of LTA from the application perspective: LTA is a tool for detecting probabilistic symptom co-occurrence/mutual-exclusion patterns in data, and it partitions the patient population in multiple ways based on those patterns.

5. Interpreting Latent Variables

LTA can reveal symptom co-occurrence/mutual-exclusion patterns hidden in data. Once the patterns are detected, we need to determine what they mean from the TCM perspective. This is called model interpretation and it requires domain knowledge and expert judgments [14,16,17,30]. In this section, we illustrate the process with a few examples.

We begin with Y29, which captures the probabilistic co-occurrence of hypodynamia and lassitude. To determine the TCM connotation of the pattern, we ask this question: What TCM syndrome(s) can cause the co-occurrence of the two symptoms? Our answer is qi deficiency because, according to TCM textbook [e.g., 31], lassitude and hypodynamia are the key symptoms of qi deficiency.

The concept of qi deficiency has several aspects to it. Y29 is about only one of the aspects, namely the manifestations of qi deficiency on vitality. There are other latent variables that are related to other aspects of qi deficiency. For example, Y08 is apparently about the manifestations of qi deficiency on pulse.

Y25 captures the probabilistic co-occurrence of insomnia and dreamful sleep. According to TCM textbook, the two symptoms can be caused by any of several syndromes: yin deficiency, fire, blood deficiency, qi deficiency and dampness. As such, Y25 is related to all of them.

Y01 reveals that the three symptoms asthenia of defecation, dry stool or constipation, and sallow complexion tend to co-occur. According to TCM textbook, the first two of the three symptoms can be caused by qi deficiency, while the last two can be caused by blood deficiency. As such, Y01 is related to both of those two syndromes. Qi deficiency is considered the primary syndrome related to Y01 because it explains the leading symptom in
the co-occurrence pattern, which is the most important factor to consider when distinguishing between patients with the pattern from those without. Since blood deficiency is a secondary syndrome relative to Y01, it is not included in the table.

\( Y12 \) reveals the probabilistic mutual exclusion of slippery pulse and thin pulse. According to TCM textbook, slippery pulse can be caused by dampness, while thin pulse can be caused by qi deficiency or blood deficiency. As such, \( Y12 \) is related to all of the three syndromes.

In summary, there are four possible scenarios regarding the interpretation of a latent variable. First, the latent variable captures a symptom co-occurrence pattern and there is a unique syndrome that can explain the pattern (e.g., \( Y29, Y08 \)). Second, the latent variable captures a symptom co-occurrence pattern and the pattern can be explained by one of several syndromes (e.g., \( Y25 \)). Third, the latent variable captures a symptom co-occurrence pattern and the explanation of the pattern requires the combination of two or more syndromes (e.g., \( Y01 \)). In this case, the syndrome that explains the leading symptom is considered the primary interpretation. Fourth, the latent variable captures a symptom mutual exclusion pattern and different symptoms in the pattern are explained by different syndromes (e.g., \( Y12 \)).

Several remarks are in order. First, the interpretation of a pattern is about identifying an appropriate explanation for the pattern. It is not about determining the syndrome type of a patient given that the pattern is present. To interpret the co-occurrence of hypodynamia and lassitude, for instance, the correct question to ask is: What syndrome(s) can cause the occurrence of the two symptoms? The wrong question is: What syndrome class does a patient belong to if he has the two symptoms? To determine the syndrome type of a patient, more information needs to be considered in addition to those two symptoms.

Second, while \( Y29 \) is about qi deficiency, it is not appropriate to, as suggested in previous literature [12,14,16,17,30], interpret the cluster \( Y29=s1 \) as a the class of patient with qi deficiency. The concept of qi deficiency involves many other symptoms in addition to hypodynamia and lassitude. A more appropriate interpretation \( Y29=s1 \) would be 'low vitality due to qi deficiency'.

Third, when interpreting a symptom co-occurrence pattern, one should try to interpret the pattern as a whole, rather than interpreting the individual symptoms. Take the co-occurrence of asthenia of defecation and dry stool or constipation as an example. The symptom dry stool or constipation can be explained by the syndrome factor fire. However, the syndrome factor fire does not explain the other symptom and hence is not an appropriate interpretation for the pattern. A more appropriate explanation for the pattern is qi deficiency, because it explains the occurrence of both symptoms.
Sometimes, a co-occurrence pattern cannot be interpreted as a whole. One example is the co-occurrence of the three symptoms asthenia of defecation, dry stool or constipation, and sallow complexion. So, we interpret the co-occurrence of the first two symptoms using qi deficiency and the co-occurrence of the last two symptoms using blood deficiency. And qi deficiency is considered the primary interpretation.

Finally, model interpretation requires expert judgments. It is possible that different experts interpret the same pattern in different ways. This issue is beyond data analysis, and must be resolved through discussions among TCM researchers. Model interpretation is where TCM researchers need to spend the most efforts when using the latent tree analysis approach.

6. Establishing the Definitions of Patient Subclasses

In this section we discuss how to use the latent variables identified by LTA to establish patient subclass definitions. This issue was first investigated in [16].

In the previous section, we have examined the TCM connotations of the latent variables $Y_{01}, Y_{08}, Y_{12}, Y_{25},$ and $Y_{29}$. It turns out that they are all related to qi deficiency. There are two other latent variables that are also related to qi deficiency, namely $Y_{20}$ and $Y_{26}$. As shown in Table 6, $Y_{20}$ reveals the probabilistic co-occurrence of the three symptoms fat tongue, tongue with ecchymosis, tooth marked tongue. This co-occurrence can be caused by qi deficiency, dampness or blood stasis. $Y_{26}$ reveals the probabilistic co-occurrence of chest oppression and palpitation, which can be caused by qi deficiency as well as yang deficiency or qi stagnation.

The seven latent variables mentioned above are about different aspects of qi deficiency. For example, $Y_{29}$ is about the impact of qi deficiency on vitality; $Y_{08}$ is about the impact of qi deficiency on pulse; $Y_{25}$ is related to the impact of qi deficiency on sleep; $Y_{26}$ is related to the impact of qi deficiency on chest; and so on. In the next step of data analysis, we partition patients into several clusters by considering all the aspects jointly. The step is hence called joint clustering [16].

Joint clustering is carried out using the model shown in Figure 4. The top part of the model is a latent class model where the seven latent variables $Y_{01}, Y_{08}, ..., Y_{29}$ are regarded as features and a new latent variable $Z_1$ is introduced to represent the patient clusters to be identified. This underscores the key idea of joint clustering, which is to partition the patients using latent class analysis (LCA) based on all the latent variables that are related to a particular syndrome. Symptom variables are added at the bottom so that the values of the...
latent variables can be inferred from data. The computational task is to determine the number of possible values for the joint clustering variable $Z_1$ and the probability parameters. This is done using Latern through a procedure similar to standard LCA.

The result of joint clustering with the eight latent variables is a partition with 3 patient clusters, which we denote as $Z_1=s_0$, $Z_1=s_1$ and $Z_1=s_2$ respectively. For reasons that will become clear later, we merge the first two clusters and thereby obtain a two-cluster partition. Information about the two-cluster partition is given in Table 7.

In the table, the symptom variables are arranged in descending order according to their mutual information with the partition, which is shown in the second column. Mutual information measures the amount of information that a symptom variable contains about the partition. It is closely related to the difference in occurrence probabilities of the variable in the two clusters. The larger the mutual information, the larger the difference in occurrence probabilities, and the more important the variable is for distinguishing the two clusters in the partition. We see that the most important symptoms are palpitation, hypodynamia and so on, in that order.

The cumulative information coverage of a variable in the ordered list is a ratio, where the numerator is the amount of information about the partition contained in the variable and all the variables before it, and the denominator is the amount of information about the partition contained in all the variables $[14,30]$. It exceeds 95% at the symptom dreamful sleep. In such a case, we conclude that the symptom and all those before it are sufficient for characterizing the differences between the two clusters. Other symptoms are therefore ignored and not included in the table. Note that this criterion was implicitly used earlier when interpreting $Y_{01}$, $Y_{25}$ and $Y_{29}$, where the variables Clear_profuse_urination, Flushed_face and Loose_stool are ignored respectively.

The last two columns of Table 7 show the occurrence probabilities of the symptoms in the two clusters. We see that the symptoms occur with much higher probabilities in the cluster $Z_1=s_2$ than in the cluster $Z_1=s_0$ or $s_1$. It is hence appropriate to interpret $Z_1=s_2$ as the subclass of patients with qi deficiency and $Z_1=s_0$ or $s_1$ as the subclass of patients without qi deficiency.

If we accept the interpretation and if the patients surveyed are a representative sample of the general VMCI population, then we have established a quantitative definition of qi deficiency subclass for VMCI and it is given in the last two columns of Table 7. The definition tells us which symptoms occur with high frequency in the case of qi deficiency and the exact probability values. It also gives occurrence probabilities of the symptoms on patients without qi deficiency. This is important because the symptoms most important for
determining qi deficiency are those that not only occur with high probability in the case of qi deficiency, but also occur with low probability on patients without qi deficiency.

In addition to establishing a quantitate definition of qi deficiency subclass for VMCI, we have also determined its prevalence among VMCI patients. Specifically, 24% of VMCI patients have qi deficiency.

We end this section with two remarks. First, the two clusters Z1=0 and Z=1 are merged because the six symptoms we end up using to define qi deficiency occur with low probabilities in both of the clusters. Keeping them separate would require the inclusion in Table 7 of information for distinguishing the two clusters, which is not related to qi deficiency and would be distracting to the reader [17].

Second, two identifiability issues could potentially arise with joint clustering [16]. In Figure 4, there are eight latent variables at the middle level. If there were only two, the model would not be identifiable, meaning that the parameters cannot be uniquely determined [20]. In such a case, we eliminate the latent variables at the middle level and connect the symptom variables directly to the joint clustering variable at the top. In Figure 4, each latent variable at the middle level is directly connected to more than one symptom variables. If a latent variable at the middle level were connected directly to only one symptom variable, the model would also be unidentifiable. In such a case, we remove the latent variable and connect the symptom variable directly to the latent variable at the top. These solutions are automatically enforced in the Lantern software.

7. Establishing Classification Rules

Suppose the definition for a patient subclass has been established using the joint clustering method. In this section we discuss how to determine whether a particular patient belongs to the subclass. This issue was first investigated in [17].

7.1 Model-Based Classification

In our approach, the definition of a patient subclass is established using a joint clustering model such as the one shown in Figure 4. For generality, denote the symptoms variables as \( X_1, \ldots, X_n \) and the joint clustering latent variable as \( Z \). Let \( s \) be a state of \( Z \) that corresponds to the patient subclass of interest and use \( \neg s \) the other state(s). The problem is to determine whether \( Z=s \) or \( Z=\neg s \) based on the values of the symptoms variables.
The problem is simple in theory. All we need to do is to compute the posterior distribution \( P(Z|X_1, \ldots, X_n) \) of \( Z \) given the values of the symptom variables, and conclude that \( Z=s \) if and only if the following inequality is true:

\[
P(Z = s|X_1, \ldots, X_n) \geq P(Z = \sim s|X_1, \ldots, X_n). \tag{1}
\]

This method is called \textit{model-based classification}. Although conceptually simple, it lacks operability. It is unrealistic to expect doctors to do probability calculations in the clinic setting. It is of course possible to write a software tool for the doctors to use as a black box. However, it might be difficult for patients and doctors to trust black boxes.

### 7.2 Scored-Based Classification Rules

For the sake of operability, we derive a score-based classification rule to approximate model-based classification. Scores are associated with the symptoms. The total score for a patient is calculated based on the presence and absence of the symptoms. When the total score exceeds a threshold, the patient is classified into the class \( Z=s \).

We start by rewriting inequality (1) into the following equivalent form using Bayes rule:

\[
P(Z = s)P(X_1, \ldots, X_n|Z = s) \geq P(Z = \sim s)P(X_1, \ldots, X_n|Z = \sim s).
\]

To obtain a score-based classification rule, we assume that the symptom variables are mutually independent given \( Z=s \) or \( Z=\sim s \). Strictly speaking, the assumption is not true. Hence approximations are introduced here. The assumption allows us to rewrite the inequality further as follows:

\[
P(Z = s)P(X_1|Z = s) \ldots P(X_n|Z = s) \geq P(Z = \sim s)P(X_1|Z = \sim s) \ldots P(X_n|Z = \sim s).
\]

By taking logarithm of both sides and re-arranging the terms, we get:

\[
\log \frac{P(X_1|Z = s)}{P(X_1|Z = \sim s)} + \ldots + \log \frac{P(X_n|Z = s)}{P(X_n|Z = \sim s)} \geq \log \frac{P(Z = \sim s)}{P(Z = s)} \tag{2}
\]

There are two technical issues here. First, the choice of base for logarithm does not matter. For interpretability of the terms, we assume that the base is 2. Second, the probability values involved in (2) might be 0 sometimes. We deal with this issue by smoothing. The formula for calculating the conditional distribution \( P(X_i|Z) \) is \( P(X_i|Z) = \frac{P(X_i; Z)}{P(Z)} \). Smoothing means to change the formula to \( P(X_i|Z) = \frac{P(X_i; Z) + c}{P(Z) + |X|c} \), where \(|X|\) is the number of possible values of \( X \), \( c \) is the \textit{smoothing parameter}. The smoothing parameter is to be determined by
the user. In Lantern, it is set at 0.000001 by default.

Note that on the left hand side of inequality (2), there is one term for each symptom variable. We regard it as the score for that symptom. To be more specific, the score for symptom variable $X_i$ is:

$$\log \frac{P(X_i|Z = s)}{P(X_i|Z = \sim s)}$$

There are actually two scores, one for $X_i = 0$ (absence of the symptom), and another for $X_i = 1$ (presence of the symptom). The term on the right hand is regarded as the threshold. Under this interpretation, the inequality becomes a classification rule and it is an approximation to the rule given by inequality (1).

As an example, consider the patient subclass $Z1 = s2$ (qi deficiency) that was defined in the previous section. The corresponding score-based classification rule is given in the left half of Table 8. Two scores are associated with each symptom, corresponding to the presence and absence of the symptom respectively. For example, a patient gets 2.3 points if he has the symptom palpitation and gets -1.3 points if the symptom is absent. The threshold is 1.6, which is given at the top of the table.

To use the classification rule on a patient, all the symptoms in the table must be examined. The patient gets one score for each symptom, either the score for the presence of the symptom or the score for its absence. If the total score exceeds the threshold, the patient is classified into the subclass $Z1 = s2$. Otherwise, the patient is classified into the subclass $Z1 = \sim s2$.

The classification rule is an approximation to model-based classification. How accurate is the approximation? To answer this question, we applied both methods on the patients in the VMCI data set. It turns out that the rule classifies 93.1% of the patients the same way as model-based classification. Therefore, the accuracy of the rule is 93.1%, as indicated at the bottom of the “accuracy” column in Table 8.

### 7.3 Understanding the Scores

It is important to realize that the scores in Table 8 are calculated from the probability values in Table 7. In fact, the scores are logarithms of probability ratios. Take palpitation as an example. It occurs with higher probability in the subclass $Z1 = s2$ (0.65) than the subclass $Z1 = \sim s2$ (0.13). In other words, it occurs more often in the case of qi deficiency than the case without qi deficiency. Hence the presence of palpitation is positive evidence for qi deficiency.
The strength of the evidence is determined by first calculating the ratio of the two probability values and then taking logarithm, i.e., $\log_2(0.65/0.13)=2.321$. For simplicity, we round up the number to one decimal place and get the score 2.3, which is the score for the presence of palpitation shown in Table 8.

Now consider the absence of palpitation. It happens less often in the subclass $ZI=s2$ (with probability $1-0.65=0.35$) than in the subclass $ZI=\sim s2$ (with probability $1-0.13=0.87$). Hence it is negative evidence for qi deficiency. The strength of the evidence is calculated using the formula $\log_2(0.35/0.87)=-1.313$. Rounding up the number to one decimal place, we get -1.3, which is the score for the absence of palpitation shown in Table 8.

The threshold is also the logarithm of probability ratio. In Table 7, we see that there are more patients in the subclass $ZI=\sim s2$ (74%) than in the subclass $ZI=s2$ (24%). In other words, there are more patients without qi deficiency than with qi deficiency. This is positive prior evidence for no qi deficiency (and hence negative evidence for qi deficiency). The strength of the evidence is calculated using the formula $\log_2(0.76/0.24)\approx1.6$, which is the threshold given in Table 8.

So, the left hand side of inequality (2) is total evidence for the subclass $Z=s$ from the symptoms, whereas the right hand side is prior evidence for the subclass $Z=\sim s$ from the sizes of the two subclasses. A patient is classified into the subclass $Z=s$ if and only if there is more evidence for $Z=s$ than $Z=\sim s$.

Two remarks are in order. First, we see in Table 8 that, in most cases, the presence of a symptom is positive evidence for $ZI=s2$, whereas the absence of a symptom is negative evidence. However, there are exceptions. For example, the presence of flushed face is negative evidence for qi deficiency with score -0.8, while its absence is weak positive evidence with score 0.1.

Second, more symptoms are included in Table 8 than Table 7. This indicates that, although the symptoms at the bottom of Table 8 are not important for characterizing the difference between the two subclasses $Z=s2$ and $Z=\sim s2$, some of them can be still important for patient classification. Take the symptom asthenia of defecation as an example. It occurs infrequently and hence is not an important factor to consider when we try to understand the differences between the two subclasses. However, the score for the presence of the symptom is the second highest (1.5), which means that it is strong evidence for qi deficiency.

### 7.4 Simplification of Classification Rules
The classification rule shown in the left half of Table 8 involves 19 symptoms variables. Some of them have scores with low absolute values. For example, the scores for thin pulse are -0.1 and 0.2, which are both small in absolute values. Such low-score symptoms are not important and we might consider eliminating them from the rule for simplicity.

The elimination of symptoms from a classification rule might affect its accuracy. The impact needs to be assessed before the elimination actually takes place. The Lantern software has a function to facilitate this operation. To illustrate the function, we consider eliminating the k symptoms at the bottom. The accuracies of the simplified rules are shown in the “accuracy” column. The number 0.931 at the bottom is the accuracy for the rule with no symptom removed; The fourth last number 0.933 is the accuracy for the rule with the last three symptoms removed; and so on. We see that, if we keep all the 10 symptoms up to loose stool and remove the last 9 symptoms from the rule, the classification accuracy is still 0.931. We recommend the rule with the first 10 symptoms as the final rule.

7.5 Single-Score Classification Rules

We call the rule given by inequality (2) the double-score classification rule. To apply the rule, a user needs to remember two scores for each symptom, one for the presence of the symptom and one for the absence of the symptom. There is a way to transform the rule so that the user needs to remember only one score for each symptom.

By subtracting a constant from both sides of inequality (2), we get the following single-score classification rule:

\[
\sum_{i=1}^{n} \log \left( \frac{P(X_i | Z = s)}{P(X_i | Z = \sim s)} / \frac{P(X_i = 0 | Z = s)}{P(X_i = 0 | Z = \sim s)} \right) \geq \log \left( \frac{P(Z = \sim s)}{P(Z = s)} \right) - \sum_{i=1}^{n} \log \left( \frac{P(X_i = 0 | Z = s)}{P(X_i = 0 | Z = \sim s)} \right)
\] (3)

There is one term for each symptom variable on the left hand side. We regard it as the score for that symptom. To be more specific, the score for symptom variable \( X_i \) is:

\[
\log \left( \frac{P(X_i | Z = s)}{P(X_i | Z = \sim s)} / \frac{P(X_i = 0 | Z = s)}{P(X_i = 0 | Z = \sim s)} \right)
\]

Note the score is 0 when \( X_i = 0 \), i.e., when the symptom is absent. So, the user needs to remember only the score for \( X_i = 1 \), i.e., the score for the presence of the symptom. It is the difference between the scores for \( X_i = 1 \) and \( X_i = 0 \) in the double-score classification rule. The term on the right hand side is the threshold of the rule. Note that it changes with the number of variables \( n \) included in the rule.

A single-score classification rule for the patient subclass \( Z1=s2 \) (qi deficiency) is given
on the right half of Table 8. We see that the score for palpitation is 3.6, which equals the difference between the two scores 2.3 and -1.3 for palpitation in the double-score classification rule. The threshold is 9.7, which equals the threshold of the single-score rule 1.6 minus the sum of the scores for the absence of all the 10 symptoms used in the rule. It is much higher than the threshold for the double-score rule. The two rules are equivalent and hence have the exactly the same accuracy.

To apply the classification rule on a patient, a user can examine the 10 symptoms one by one. The patient gets a score for each symptom that occurs. There are no scores for the symptoms that are not present on the patient. If the total score exceeds the threshold, the patient can be classified into the subclass $Z_1 = s_2$. Because all the scores are positive, there is no need to continue examining the other symptoms is the total score of some of the symptoms exceed the threshold. If the total score is still below the threshold after all the 10 symptoms are examined, the patient is classified into the subclass $Z_1 = \neg s_2$.

Two remarks are in order. First, the transformed score for $X_i = 1$ is has a nice interpretation. To see this, note that:

$$\log\left(\frac{P(X_i = 1|Z = s)}{P(X_i = 0|Z = s)} / \frac{P(X_i = 1|Z = \neg s)}{P(X_i = 0|Z = \neg s)}\right) = \log\left(\frac{P(X_i = 1|Z = s)}{P(X_i = 0|Z = s)} / \frac{P(X_i = 1|Z = \neg s)}{P(X_i = 0|Z = \neg s)}\right)$$

The first term inside the parenthesis is the odds for $X_i = 1$ in the cluster $Z = s$, while the second term is the odds for $X_i = 1$ in the cluster $Z = \neg s$. So, the score is simply a log odds ratio, a standard term in Statistics to quantify how strongly the presence or absence of one property (the symptom $X_i$) is associated with the presence or absence of another property (whether the patient is in the cluster $Z = s$). The score for $X_i = 1$ is large if it occurs with high probability in $Z = s$ and low probability in $Z = \neg s$.

Second, the shaded part of the right half of the Table 8 shows what happens if we include more symptoms in the single-classification rule. For example, if we include the symptom flushed face, the score for the symptom is -0.9, the threshold and the accuracy remain unchanged; If we further slippery slippery pulse, the score for the symptom is -0.3, the threshold is decreased from 9.7 to 9.6, and the accuracy becomes 0.935; and so on.

8. Discussions

This paper is concerned with the problem of how to classify the patients of a WM disease from the TCM perspective. We have presented the latent tree analysis approach for solving the problem. Next we discuss several issues concerning the approach and some related work.
8.1 Double-Score Classification Rules versus Single-Score Classification Rules

For simplicity, we call the two classification rules given in Table 8 as Rule 1 and Rule 2 respectively. In Rule 2, each symptom is associated with one score. For example, the score for palpitation is 3.6. In Rule 1, on the other hand, each symptom is associated with two scores, one for the presence of the symptom and another for its absence. For example, the score for the presence of palpitation is 2.3, indicating that the presence of the symptom is positive evidence for qi deficiency. The score for the absence of palpitation is -1.3, indicating that the absence of the symptom is negative evidence for qi deficiency.

We argue that Rule 1 matches human reasoning better than Rule 2. Imagine that a TCM practitioner has considered some other evidence and has formed some belief about whether a patient has qi deficiency. He next considers the symptom palpitation. According to Rule 1, his belief would increase if he finds out that the patient has palpitation, and his belief would decrease if he finds out that the patient does not have palpitation. This matches our intuition. According to Rule 2, on the other hand, his belief would stay the same if he finds out that the patient does not have palpitation. This is somewhat counter-intuitive. The impact of the absence of palpitation can be appreciated only if the threshold is taken into consideration: Because the total score does not increase after he examines palpitation, there is now less hope to reach the threshold.

On the other hand, Rule 2 is obviously simpler and easier to use. Moreover, it allows the practitioner to reach a conclusion after examining one of some of the symptoms. In the case where all symptoms in a rule have positive scores, if the total score exceeds the threshold after examining only some of the symptoms, the patient can be classified into \( Z^2=s \) without considering the remaining symptoms.

8.2 Integer-Valued Classification Rules

In Section 7.4 and 7.5, we have presented two ways to simplify classification rules. Another idea is to convert real-valued classification rules into integer-valued rules [17], so as to be consistent with the literature where classification rules are always given using integers [e.g., 3].

A real-valued classification rule can be converted into an integer-valued rule by applying rounding to the symptom scores and the threshold. Obviously, rounding affects the accuracy of the rule. To minimize the impact, we can multiply all the scores and the threshold by a
scaling factor and then applying rounding. This operation is supported by the Lantern software.

There are two drawbacks with integer-valued classification rules. First, the symptom scores and the threshold are no longer logarithms of probability ratios or log of odds ratio. They do not have clear semantics. Second, different researchers might use different scaling factors even if they work on the same problem. This renders their results not comparable. For the field to move forward, it is important that results from different research groups are comparable. For the same reason, it is not advisable to enforce that all classification rules have the same threshold.

8.3 Comparisons with Previous Classification Rules

The final outputs of the latent tree analysis approach are quantitative patient subclass definitions and classification rules. In this subsection, we compare them with similar rules from the literature. As a specific example, we reproduce below part of the classification rule for qi deficiency given in [3]:

**Symptom scores:** Weak limbs and body (4), hypodynamia (5), palpitation (3), frequent nocturnal urination (5), frequent daytime urination (5), lassitude (4), fear of wind (4), easy to catch cold (6), asthenia of urination (4), sallow complexion (4), pale tongue (3), feeble pulse (4), thin pulse (3), …

**Threshold:** 14

For simplicity, we refer to this rule as Rule 3.

The first difference between the Rule 3 and Rules 1 and 2 is that the scores and threshold in Rules 1 and 2 have clear semantics, while those in Rule 3 do not. For example, the score for the presence of palpitation in Rule 1 is 2.3. This means that palpitation occurs $2^{2.3} \approx 4.9$ times more often among patients with qi deficiency than among those without qi deficiency. The score for the absence of palpitation is -1.3. This means that the absence of palpitation is $2^{-1.3} \approx 2.5$ times more common among patients without qi deficiency than among those with qi deficiency. The threshold is 1.6. It means that the patients without qi deficiency are $2^{1.6} \approx 3$ times as many as those with qi deficiency. In Rule 3, the score for palpitation is 4. It is not interpretable in terms of symptom occur frequencies. The threshold 14 is not interpretable either.

The second difference between Rule 3 and Rules 1 and 2 is that they are derived from different types of data. Rule 3 is derived from labeled data that contains not only information about symptoms but also judgments by TCM practitioners about the syndrome types of
patients. In general, the purpose in analyzing such data is to extract the experiences of TCM practitioners and present them in an easy-to-use manner. In contrast, Rules 1 and 2 are derived from unlabeled data that contains only information about symptoms. There are no judgments about syndrome types. When analyzing such data, we first try to identify symptom occurrence patterns hidden in data and then use those patterns to define patient subclasses that correspond to syndromes. Finally, classification rules are derived to distinguish between different patient subclasses. Whereas Rule 3 characterizes a syndrome concept that is the minds of TCM practitioners, Rules 1 and 2 characterizes a patient subclass that is defined based on data patterns and that can be interpreted to a syndrome.

8.4 Previous Work on Defining Syndrome Subclasses

A key idea of this paper is to perform cluster analysis on unlabeled symptom data in hope to find patient subclasses that match TCM syndrome concepts. In cluster analysis, patients are partitioned into groups based on patterns of symptom occurrence. If the statistical characteristics of some of the groups match some TCM syndromes, then they can be used as quantitative definitions for the corresponding syndromes.

The idea has been pursued by other researchers [27,32] and alternative methods have been proposed. Those methods partition the patients in one way and the partition is based on all symptom variables in data. In contrast, our method partitions the patients in multiple ways and different partitions are based on different and possibly overlapping subsets of symptom variables. For the VMCI data, one partition is given in Table 8. It is obtained based on 19 symptom variable shown in Figure 4. Other partitions will be reported in another paper [18]. Our method is better suited for the TCM syndrome classification problem because each syndrome conceptually partition patients into two classes, those with the syndrome and those without, and different syndromes give different partitions.

It should be noted that cluster analysis has previously been used more often to group symptom variables than to group patients, and the symptom variable clusters are interpreted as syndromes. This is problematic because symptoms variables being similar to each other (and hence grouped together) does not necessarily imply the symptoms co-occur [33].

Factor analysis has also been used in an attempt to define syndromes in terms of symptoms [34]. Factor analysis assumes observed variables (symptoms) to be linear combinations of latent variables (syndromes). The coefficients are called factor loadings. Researchers typically report factor loadings for latent variables and interpret the latent variables based on observed variables with high factor loadings. However, since the latent
variables in factor analysis are continuous, it does not give partitions of patients as LTA does. In addition, the factors are usually assumed to be independent of each other, while TCM syndromes are usually considered correlated.

9. Conclusions

The latent tree analysis approach to TCM syndrome classification is systematically presented in this paper. Following the approach, TCM researchers can find the answers to the following questions about a WM disease starting from unlabeled symptom data: What TCM syndrome subclasses are there among patients with the disease? What are the sizes of the subclasses? What are the statistical characteristics of each subclass? How can a doctor determine whether a patient belongs to a particular subclass?

The approach consists of 6 steps: (1) Collect symptom data about the patient population under study; (2) Analyze the data using latent tree models to obtain latent variables that capture probabilistic symptom co-occurrence/mutual-exclusion patterns hidden in data; (3) Interpret the latent variables to determine their TCM connotations; (4) Group the latent variables based on their TCM meanings; (5) Conduct joint clustering based on each group of latent variables to determine patient classes; (6) Establish classification rules for determining which class or classes a particular patient belongs to. Among the six steps, steps 2, 5 and 6 are carried out using computers, while steps 1, 3 and 4 are done by TCM researchers.

Components of the approach are previously scattered in the literature. This is the first time all the pieces are put together in one paper. Non-essential and potentially misleading materials are excluded. Significant improvements are introduced regarding the interpretation of latent variables and the establishment of classification rules. In particular, the materials presented in Sections 7.3, 7.4, 7.5, 8.1 and 8.3 are novel. A software package is developed to facilitate the use of the approach.

References


(Chinese)


Figure 1. The subfigure (a) and the tables illustrate the concept of latent tree models using an example that involves two latent variables (the skill variables) and four observed variables (the grade variables). The subfigure (b) illustrates the concept of latent class models where Intelligence is the only latent variable.

(a) A latent tree model

(b) A latent class model

AS = Analytical Skill, MG = Math Grade
LS = Literacy Skill

Figure 2. The subfigure (a) shows the model for latent class analysis. There is only one latent variable Y. The task is to determine the number of values for Y and the probability parameters. The subfigure (b) shows a model that might be obtained from latent tree analysis, where it is necessary to determine the number of latent variables and connections among them.

(a) The model for latent class analysis

(b) A possible model from latent tree analysis
Figure 3. Structure of the latent tree model obtained on the VMCI data: The variables labeled with English phrases are symptom variables, while the Y-variables are latent variables. The integer next to a latent variable is the number of its possible values.
Table 1. Partition given by Y01

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Y01=s0</th>
<th>Y01=s1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asthenia_of_defecation</td>
<td>0.02</td>
<td>0.61</td>
</tr>
<tr>
<td>Dry_stool_or_constipation</td>
<td>0.21</td>
<td>0.82</td>
</tr>
<tr>
<td>Sallow_complexion</td>
<td>0.06</td>
<td>0.40</td>
</tr>
</tbody>
</table>

Table 2. Partition given by Y29

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Y29=s0</th>
<th>Y29=s1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hypodynamia</td>
<td>0.16</td>
<td>1.00</td>
</tr>
<tr>
<td>Lassitude</td>
<td>0.12</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 3. Partition given by Y08

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Y08=s0</th>
<th>Y08=s1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunken_pulse</td>
<td>0.02</td>
<td>0.64</td>
</tr>
<tr>
<td>Feeble_pulse</td>
<td>0.00</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 4. Partition given by Y25

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Y25=s0</th>
<th>Y25=s1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Insomnia</td>
<td>0.16</td>
<td>0.78</td>
</tr>
<tr>
<td>Dreamful_sleep</td>
<td>0.23</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Table 5. Partition given by Y12

<table>
<thead>
<tr>
<th>Syndrome</th>
<th>Y12=s0</th>
<th>Y12=s1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slippery_pulse</td>
<td>0.85</td>
<td>0.16</td>
</tr>
<tr>
<td>Thin_pulse</td>
<td>0.00</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 6. Information about several latent variables: Table shows the probabilistic patterns revealed by the latent variables, the relevant syndromes, and the aspects of the syndromes that the patterns are about. All the 7 latent variables are related to the syndrome qi deficiency.

<table>
<thead>
<tr>
<th>Latent variable</th>
<th>Pattern</th>
<th>Related Syndromes</th>
<th>Aspect</th>
</tr>
</thead>
<tbody>
<tr>
<td>Y01</td>
<td>Co-occurrence: asthenia of defecation, dry stool or constipation, and sallow complexion</td>
<td>qi deficiency, blood deficiency</td>
<td>Defecation</td>
</tr>
<tr>
<td>Y08</td>
<td>Co-occurrence: sunken pulse, feeble pulse</td>
<td>qi deficiency</td>
<td>Pulse</td>
</tr>
<tr>
<td>Y12</td>
<td>Mutual-exclusion: slippery pulse, thin pulse</td>
<td>dampness, qi deficiency, blood deficiency</td>
<td>Pulse</td>
</tr>
<tr>
<td>Y20</td>
<td>Co-occurrence: fat tongue, tongue with ecchymosis, tooth marked tongue</td>
<td>qi deficiency, dampness, blood stasis</td>
<td>Tongue</td>
</tr>
<tr>
<td>Y25</td>
<td>Co-occurrence: insomnia, dreamful sleep</td>
<td>yin deficiency, fire, blood deficiency, qi deficiency, yang deficiency, dampness</td>
<td>Sleep</td>
</tr>
<tr>
<td>Y26</td>
<td>Co-occurrence: chest oppression, palpitation</td>
<td>qi deficiency, yang deficiency, qi stagnation</td>
<td>Chest</td>
</tr>
<tr>
<td>Y29</td>
<td>Co-occurrence: hypodynamia, lassitude</td>
<td>qi deficiency</td>
<td>Vitality</td>
</tr>
</tbody>
</table>
Figure 4. Model for joint clustering.

Table 7 Two-cluster partition obtained by joint clustering based on the latent variables related to qi deficiency: MI -- Mutual information, CIC -- Cumulative Information Coverage

<table>
<thead>
<tr>
<th></th>
<th>MI</th>
<th>CIC</th>
<th>Z1=s0 or s1</th>
<th>Z1=s2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Palpitation</td>
<td>0.12</td>
<td>0.44</td>
<td>0.13</td>
<td>0.65</td>
</tr>
<tr>
<td>Hypodynamia</td>
<td>0.10</td>
<td>0.69</td>
<td>0.36</td>
<td>0.86</td>
</tr>
<tr>
<td>Chest_oppression</td>
<td>0.09</td>
<td>0.79</td>
<td>0.21</td>
<td>0.67</td>
</tr>
<tr>
<td>Lassitude</td>
<td>0.05</td>
<td>0.83</td>
<td>0.21</td>
<td>0.57</td>
</tr>
<tr>
<td>Insomnia</td>
<td>0.05</td>
<td>0.94</td>
<td>0.29</td>
<td>0.65</td>
</tr>
<tr>
<td>Dreamful_sleep</td>
<td>0.03</td>
<td>0.96</td>
<td>0.38</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Table 8. Two equivalent classification rules for $ZI=s_2$ vs $ZI=-s_2$: The double-score classification is given on the left hand side of the table. There are two scores for each symptom, corresponding to the absence and presence of the symptom respectively. A patient is classified into class $ZI=s_2$ if and only if the total score exceeds the threshold. The accuracy of the rule with respect to model-based classification is shown at the bottom in the “accuracy” column. Other values in the column show the accuracies of the rule if only some of symptoms (those at the bottom) are eliminated. The rule with the first 10 symptoms has accuracy 0.931. It is recommended as the final rule. The single-score classification rule is given on the right hand side of the table. There is only one score for each symptom. It is for the presence of the symptom. The score of the absence of any symptom is zero. The threshold is much higher than the natural classification rule, and it changes with the number of the symptoms included in the rule.

<table>
<thead>
<tr>
<th>Double-Score Classification</th>
<th>Single-Score Classification Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule</td>
<td>Threshold: 1.6</td>
</tr>
<tr>
<td></td>
<td>score</td>
</tr>
<tr>
<td></td>
<td>Threshold</td>
</tr>
<tr>
<td>Absent</td>
<td>Present</td>
</tr>
<tr>
<td>Palpitation</td>
<td>-1.3</td>
</tr>
<tr>
<td>Hypodynamia</td>
<td>-2.1</td>
</tr>
<tr>
<td>Chest_oppression</td>
<td>-1.3</td>
</tr>
<tr>
<td>Lassitude</td>
<td>-0.9</td>
</tr>
<tr>
<td>Insomnia</td>
<td>-1.0</td>
</tr>
<tr>
<td>Dreamful_sleep</td>
<td>-0.8</td>
</tr>
<tr>
<td>Asthenia_of_defecation</td>
<td>-0.2</td>
</tr>
<tr>
<td>Dry_stool_or_constipation</td>
<td>-0.3</td>
</tr>
<tr>
<td>Sallow_complexion</td>
<td>-0.1</td>
</tr>
<tr>
<td>Loose_stool</td>
<td>-0.1</td>
</tr>
<tr>
<td>Flushed_face</td>
<td>0.1</td>
</tr>
<tr>
<td>Slippery_pulse</td>
<td>0.1</td>
</tr>
<tr>
<td>Fat_tongue</td>
<td>-0.0</td>
</tr>
<tr>
<td>Sunken_pulse</td>
<td>-0.1</td>
</tr>
<tr>
<td>Thin_pulse</td>
<td>-0.1</td>
</tr>
<tr>
<td>Clear_profuse_urination</td>
<td>-0.0</td>
</tr>
<tr>
<td>Tongue_with_ecchymosis</td>
<td>-0.0</td>
</tr>
<tr>
<td>Tooth_marked_tongue</td>
<td>-0.0</td>
</tr>
<tr>
<td>Feeble_pulse</td>
<td>-0.0</td>
</tr>
</tbody>
</table>