Proceedings of the 4th WSEAS/IASME Int. Conf. on System Science and Simulation in Engineering, Tenerife, Spain, December 16-18, 2005 (pp139-144)

Automatic Oriental Medical Diagnosis via BYY Learning Based Discrete Independent Factor Analysis

JEONG-YON SHIM
Division of General Studies, Computer Science, Kangnam University,
San 6-2, Kugal-ri, Kihung-up, YongIn Si, KyeongKi Do, KOREA
Email:mariashim@kangnam.ac.kr

LEI XU
Dept. of Computer Science and Engineering Chinese University of Hong Kong
Shatin, NT, Hong Kong, CHINA
email: lxu@cse.cuhk.edu.hk

Abstract: An oriental medical diagnosis is featured by inferring hidden causal factors of diseases from observed symptoms. This paper introduces a computer aided diagnosis in help of a linear independent factor model that interprets the observed symptoms as generated from hidden independent causes in term of discrete variables. The model is computed by algorithms obtained from the BYY harmony learning.

Key-Words: Oriental medical diagnosis, discrete independent factor analysis, BYY harmony learning.

1. Introduction

In western medical area, various medical diagnostic systems have been developed by using ever developing computer and information technologies. Starting from the system MYCIN [11, 6] which stands for If-Then production rule based system, many medical expert systems have been implemented and successfully applied to medical diagnostic area, such as EMYCIN [8], and PUFF [3], etc. These systems are usually built with manual help to transfer human knowledge into rules, in a lack of automatic learning ability. There are also open problems that are difficult to tackle, e.g., how to handle conflict and redundancy in the rule base. In the past two decades, studies on neural networks have also been made for developing medical diagnostic systems, with ability of learning and generalization as well as fault tolerance, but with deficiency on explanation of its results. In oriental medicine, doctor tries to find the hidden causal factors from the observed symptoms of patient and discover the fundamental disharmony among the hidden causal factors. The key healing approach in oriental medicine is to make harmony by compensating insufficient hidden causal factors with herbal medicine.A diagnostic process that involves only a reasoning line of IF-then rules does not suit well to this purpose. The input-output regression by a conventional neural networks is also not fit the purpose well.

This paper further studies a computer aided oriental medical diagnosis in help of a linear independent factor model that interprets observed symptoms as generated from hidden independent causes in terms of discrete variables [10]. The model is computed by algorithms obtained from the BYY harmony learning.

2. Oriental Medical Diagnosis

2.1 Yin-Yan based oriental medicine

The belief that human body is little more than an extremely sophisticated machine has led, in the West, to many extraordinary advanced approaches; for example, the remarkable developments in surgery and drug therapy. Much of the current disaffection with modern medicine amongst patients, however, stems from the limitations of this approach. It fails to recognize that the mind and spirit have an extremely powerful effect upon the body and that the human body is more than the sum of its chemistry and mechanics.

Oriental medicine never considered the mind and body as separated from each other, as Western medicine did for the last two centuries. It has a fundamentally different philosophical basis that permeates to the core of its theory and practice. Oriental medicine has a holistic therapy that treats a patient as a whole, rather than just the symptoms out of context of the person. Sickness is not understood in terms of the pathology of isolated organs, as though they were merely cogs in a medicine, but rather as the dysfunction of a normal harmonious of one whole living entity.

Both the systems of medicine can be practised more or less holistically, depending on the wisdom of the physician. What is remarkable about oriental medi-

cine is that it places diagnosis of the person at the core of its diagnostic process and regards nearly all chronic disease as a manifestation of the individual's particular weakness. When oriental medical treatment is directed at these long-standing weakness or 'imbalances', the patient is often amazed to find that not only is his main complaint improving, but many secondary complaints are also responding. This contrasts significantly with the effect of many of the modern drugs which, because of their side-effects, create secondary complaints rather than improve them.

This improvement in the patient's well-being as a whole is one of the main reasons that patients in the West have been flocking to oriental medical doctor over the last few decades. Oriental medical doctors have always specialized in treating human beings, not illnesses. What an oriental medical doctor is searching for when he sees a patient are the 'hidden factors of disharmony' which have caused the symptoms now afflicting the patient. For example, if a western doctor and an oriental doctor were both to examine a patient with difficulty in breathing, the western doctor might diagnose asthma and the oriental doctor might diagnose a deficiency in the 'Gi' of the lung. It is not that one is correct and the other incorrect; it is just that they both see the symptom through the perception of their own very different medical theories.

The Chinese Yin-Yang philosophy acts as a foundation role in oriental medicine. Yin and Yang depict two sides of one thing, e.g., the dark and sunny sides of hill. Yin and Yang are constantly in transition, just as in nature day and night constantly change into one another. Night is predominantly Yin, day predominantly Yang. Yang has characteristics of Light, Activity, Heaven, Energy, Expansion, Rising, Male, and Fire. On the other hand, Yin has characteristics of Darkness, Rest, Earth, Matter, Contraction, Descending, Female, and Water. One of the principal tasks of oriental medical diagnosis is to 'observe the relationship between Yin and Yang carefully, and to make adjustments to bring about equilibrium'. In order to do this, he must assess various factors in his diagnosis of a person's Qi, according to their Yin-Yang nature shown on the following table.

Table 1. The causal factors according to Yin-Yang

Yang	Yin
Wha(Fire)	Su(Water)
Yeol(Heat)	Naeng(Cold)
Geon(Dry)	Seup(Wet)
whalDongGwaDa	WhalDongGwaSo
(Hyper-active)	(Hypo-active)

Yin and Yang are in a status of constantly changing but in good health a balance is always maintained. Ill health will only occurs when one side starts to 'consume' the other. Yin in excess makes Yang suffer; Yang in excess makes Yin suffer. A preponderance of Yang leads to heat manifestations; a preponderance of Yin brings on cold. The four varieties of imbalance, i.e., Excess of Yin, excess of Yang, Deficiency of Yang and Deficiency of Yin, require radically different treatment.

2.2 Western versus oriental medicine

Though both attempt to find the causes from observed symptoms, western and oriental medicine have radically different views about how we become ill.

In western medicine, especially in its early developing stage, human body is more or less regarded as an extremely sophisticated machine and thus illness comes from disfunction or abnormal of certain parts or organs. A doctor first diagnoses which or where the organs, and then repairs or replaces them via surgery and drug therapy. In this understanding, the diagnosis process consists of a sequence of inferences from observed symptoms to the cause. Starting from the MYCIN system [11, 6], IF-THEN production rule system has been successfully applied to implementing such a western medical diagnosis, with many medical expert systems developed for medical diagnosis, such as EMYCIN [8] and PUFF [3], etc. However, these IF-THEN production rule systems lack automatical learning ability. They have to be built heuristically with manual help to transfer human knowledge into rules. Also, there lack effective ways to handle conflict and redundancy in a rule base.

In oriental medicine, human body is considered systematically as a dynamic system and is featured by a set of hidden factors or state variables that monitors the harmony or balance of the Yin-Yang nature (see Tab.1) in a person's Qi system. It is the 'hidden factors of disharmony' which have caused the symptoms now afflicting the patient. A oriental doctor first diagnoses the hidden factors of disharmony, and then adjusts the human body via herbal, acupuncture, and treatments to let the factors to return in harmony.

The difference between western and oriental diagnosis can be sensed more intuitively via an example that both a western doctor and an oriental doctor were examining a patient with difficulty in breathing, the western doctor might diagnose asthma and the oriental doctor might diagnose a disharmony in the 'Gi' of his lung. A deep insight on the difference between western and oriental diagnosis can be understood from the ways of handling the relation between cause and effect. A western doctor starts from the observed symptoms to find the organ or part in disfunction and the inner or exter-

nal causes that lead to the disfunction and the observed symptoms, while an oriental doctor tries to find the inside 'hidden factors of disharmony' that are regarded as the final reasons responsible for the symptoms now afflicting the patient. That is, the original causes and the effected symptoms are decoupled by the 'hidden factors of disharmony'.

There are pros and cons to both western and oriental medicine. Seeking the original causes, a western doctor is able to find measures to directly remove the causes and thus the patient may recover quickly. However, not only the disorder or illness of a patient may be caused also by some unknown hidden cases, but also a treatment on a particular cause may also incur other effects in addition to remove the effects of this cause. This is why many of the modern drugs create secondary complaints due to their side-effects. In a significant contrast, in oriental medicine the original causes are summarized by hidden factors that are regarded as final causes to the observed symptoms. This factors are classified in the external ones such as Wind, Heat, Cold, Dampness, and Dryness and the internal ones such as Anger, Joy, Worry, Sadness, and Fear. An oriental doctor diagnose them whether in a disharmony and treats to bring back to harmony. The advantage is that it works even in the case of cause of disease unknown to the ancients, such as radiation sickness. Also, treatments directed at these long-standing weakness or 'imbalances', the patient is often amazed to find that not only is his main complaint improving, but many secondary complaints are also responding. Of course, the decoupling from the original causes leads to a relative slow recovering of patient and an increased risk of misdiagnosis if the doctor is not well experienced. This is why in Asia countries, both western and oriental medicine are used in a complementary of each other.

3. Automatic Oriental Medical Diagnosis

3.1 Regression vs independent factor model

As discussed in the previous section, an IF-THEN production rule system does not suit well to oriental diagnosis since it no longer consists of a sequence of inferences from observed symptoms to an original cause. An oriental diagnosis makes the observed symptoms decoupled from a great number of original possible causes but linked to a finite number of conclusive hidden factors, which provides possibilities of building automatic medical diagnosis systems via statistical models and neural networks in help of learning methods developed in past two decades.

The observed symptoms, such as jaundice, headache, dizziness, pink eye, rapid pulse, and yellow urine, are usually measured in a format of $\mathbf{x} = [x^1, \cdots, x^d]^T$ with each element x^i standing for one symptom and the value of x^i describing a degree of the corresponding symptom. We can collect a set of symptoms from N patients to get a set $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^N$ of samples for learning a automatic medical diagnosis system. The hidden factors such as 'fire, heat, wet, and hyper-active', are described in a vector $\mathbf{y} = [y^{(1)}, ..., y^{(k)}]^T$ with each element $y^{(j)}$ being binary and taking either '1' denoting that this factor affects observed symptoms in $\mathbf{x}$ or '0' for denoting that this factor is irrelevant to the symptoms.

If we have the diagnosis records of experienced oriental medical doctors on the observed symptoms $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^N$ of all the N patients, i.e., a set of paired samples $\{\mathbf{x}_t, \mathbf{y}_t^e\}_{t=1}^N$ with $\mathbf{y}_t^e$ being the oriental medical doctors' diagnosis on $\mathbf{x}_t$, we can use a regression model or a conventional three layer neural networks to get an approximate diagnostic function $\mathbf{y}_t = f(\mathbf{x}_t, \theta)$ by supervised learning on its parameters in θ such that the discrepancy between $\mathbf{y}_t^e$ and $\mathbf{y}_t$ is minimized under a give error measure. After learning, the generalization ability of $\mathbf{y} = f(\mathbf{x}, \theta)$ makes it applicable to the observed symptoms of a new patient. For $\mathbf{y} = f(\mathbf{x}, \theta)$ implemented by a conventional three layer neural networks, the discrepancy between $\mathbf{y}_t^e$ and $\mathbf{y}_t$ can be minimized as small as possible for any set of $\{\mathbf{x}_t, \mathbf{y}_t^e\}_{t=1}^N$ as long as the number of hidden units is large enough. However, the generalization ability decreases as N increases. A learning algorithm is obtained from BYY harmony learning [15] such that an appropriate number of hidden units can be determined automatically during learning parameters in θ. The details of applying it for building an oriental medical diagnostic function $\mathbf{y} = f(\mathbf{x}, \theta)$ will be further discussed elsewhere.

The above approach has a serious weakness that not only getting a set of $\{\mathbf{x}_t, \mathbf{y}_t^e\}_{t=1}^N$ is expensive but also the performance of the obtained $\mathbf{y} = f(\mathbf{x}, \theta)$ depends how good are those oriental medical doctors who provided the set $\{\mathbf{x}_t, \mathbf{y}_t^e\}_{t=1}^N$. Moreover, it is not able to be extended for diagnosing diseases that have not been considered in $\{\mathbf{x}_t, \mathbf{y}_t^e\}_{t=1}^N$ by those oriental medical experts. Alternatively, a new diagnostic function needs to be learned for diagnosing new diseases. Furthermore, $\mathbf{y} = f(\mathbf{x}, \theta)$ approximately provides a diagnosis on $\mathbf{x}$ but unable to provide a reasonable explanation on its decisions and thus has been regarded as not reliable enough for practical purpose.

3.2 Discrete independent factor analysis

Following the oriental medical theory, we explain that observed symptoms $\mathbf{x}$ are caused by the hidden factors in $\mathbf{y}$. For simplicity, we model it by a linear model plus taking observation noise in consideration. That is

$$\mathbf{x} = A\mathbf{y} + \mu + e, \; A = [a_1, \cdots, a_m], \qquad (1)$$

where the noise e is usually independent from $\mathbf{y}$ and from Gaussian with zero mean and covariance matrix Σ, and thus we also have $\mu = E\mathbf{x} - AE\mathbf{y}$ with Eu denoting the mean vector of u. Equivalently, eq.(1) can be described by the following distribution

$$q(\mathbf{x}|\mathbf{y},\phi) = G(\mathbf{x}|A\mathbf{y}+\mu,\Sigma),\ \phi = \{A,\Sigma\}. \quad (2)$$

where $G(u|m,\ \Sigma)$ means a Gaussian distribution of variable u with mean m and covariance matrix Σ.

The earliest effort on eq.(1) at the case that $\mathbf{y}$ comes from also Gaussian can be traced back to various studies in the literature of statistics [2, 7] under the name of factor analysis. studies have been also made in other literatures under different names. One is called *linear generative model* since it describes how $\mathbf{x}$ is generated via a linear model. The other is called *latent or hidden model* since y is not directly visible from observation. In general, the hidden factors in $\mathbf{y} = [y^{(1)},...,y^{(k)}]^T$ are usually regarded as mutually independent components from each other, i.e.,

$$q(\mathbf{y}|\psi) = \prod_{j=1}^{k} q(y^{(j)}|\psi_j), \quad (3)$$

which is justified because it is quite nature to believe that hidden factors should be of the least redundancy among each other, otherwise a compound factor can be further decomposed into simpler factors. In this case, eq.(1) is called independent factor analysis.

A typical case is that each $y^{(j)}$ takes only 0 or 1 subject to a Bernoulli distribution:

$$q(y^{(j)}|\psi_j) = q_j^{y^{(j)}}(1-q_j)^{1-y^{(j)}}, \quad (4)$$

In this case, eq.(1) is called binary independent factor analysis, multiple cause model, latent trait models, and item response theory [4, 16].

For an oriental diagnosis in such a setting, observed symptoms $\mathbf{x}$ are regarded as being caused by the hidden factors in $\mathbf{y}$ under disturbance of a Gaussian noise. $y^{(j)} = 1$ indicates the existence of a casual factor that contributes to cause illness symptoms, while $y^{(j)} = 0$ represents that the symptoms is irrelevant to this casual factor. Each q_j describes probability that the casual factor $y^{(j)}$ may like to cause illness symptoms in a priori independent of person, e.g., it may reflect the heath condition of certain environment and an oriental doctor uses it in his diagnosis implicitly and get known about it according to his past experiences in this environment.

Basing on eq.(3) and eq.(2), we can further get the following posteriori probability

$$p(\mathbf{y}|\mathbf{x}) = \frac{G(\mathbf{x}|A\mathbf{y}+\mu,\Sigma)\prod_{j=1}^{k} q(y^{(j)}|\psi_j)}{q(\mathbf{x}|\phi,\psi)} \quad (5)$$

$$q(\mathbf{x}|\phi,\psi) = \sum_{\mathbf{y}} G(\mathbf{x}|A\mathbf{y}+\mu,\Sigma)\prod_{j=1}^{k} q(y^{(j)}|\psi_j),$$

which describes the degree or the probability that the particular causes in $\mathbf{y}$ are diagnosed as being responsible for the observed symptoms in $\mathbf{x}$.

The task of learning $q(\mathbf{x}|\mathbf{y},\phi)$ and $q(\mathbf{y}|\psi)$ is made on a set of observed symptoms $\mathbf{X} = \{\mathbf{x}_t\}_{t=1}^{N}$, without requiring the diagnosis records of experienced oriental medical doctors on these observed symptoms. What we need is to verify the performance by the automatic diagnosis system in consultation with experienced oriental medical doctors. Also, the system can be adaptively updated as the symptoms of new patients come to cover new diseases. Moreover, it sets up a bridge for investigating oriental medicine from a modern science perspective.

The scope that eq.(4) is applicable is limited because $y^{(j)}$ only takes a binary value. Though increasing k apparently increases the representation scope of a binary vector $\mathbf{y} = [y^{(1)},...,y^{(k)}]^T$, the independence by eq.(3) makes an actual increasing of the representation scope become limited since every factor may not be really independent from all the others. E.g., each Yin-Yang pair in Table 1 can not be regarded as independent from each other. Without violating the independence by eq.(3), one solution of this problem is considering that $y^{(j)}$ takes several discrete labels. E.g., we can use $y^{(j)} = 1$ to indicate the existence of a Yang factor and $y^{(j)} = -1$ to indicate the existence of a Yin factor, while $y^{(j)} = 0$ represents that this pair of Yin factor and Yang factor are in balance. In general, we can extend eq.(4) into

$$q(y^{(j)}|\psi_j) = \sum_{i=1}^{\kappa_j} \alpha_{ji}\delta(y^{(j)} - \ell_i),\ \sum_{i=1}^{\kappa_j}\alpha_{ji} = 1, \quad (6)$$

where $\alpha_{ji} > 0,\ \ell_i, i = 1,\cdots,\kappa_j$ are pre-specified lablel. Each of them can be either an integer (e.g., -1, 0, 1) or even a real number.

Also, it can be observed that eq.(6) degenerates back to being equivalent to eq.(4).

4. BYY Harmony Learning

4.1 ML learning vs BYY harmony learning

The task of learning $q(\mathbf{x}|\mathbf{y},\phi)$ and $q(\mathbf{y}|\psi)$ consists of specifying the parameter set $\theta = \{\phi,\psi\}$, which is called parameter learning, and specifying the integers $\{k,\{\kappa_j\}\}$, which is called model selection in the sense that different values of the integers correspond different models with a same configuration but in different scales.

The parameter learning can be made via maximum likelihood (ML) learning on $q(\mathbf{x}|\phi,\psi)$ in eq.(5), implemented by an expectation-maximization algorithm. However, it is very expensive in computing because the extensive summation $\sum_{\mathbf{y}}$ of $\prod_{j=1}^{k}\kappa_j$ terms has to be encountered in every iteration step.

Moreover, the maximum likelihood learning is poor in model selection. Conventionally, model selection has to be implemented in two phases. In the first phase, we obtain a set of candidate models by the maximum likelihood learning for a set of candidate models by enumerating k. In the second phase, we select one appropriate model based on some model selection criterion. Popular examples of model selection criteria include Akaike's information criterion [1], the consistent Akaike's information criterion [5], cross validation and the minimum description length criterion [9] which formally coincides with the Bayesian inference criterion. This process costs extensively.

Firstly proposed in 1995 [17] and then systematically developed in subsequent years, the Bayesian Ying Yang (BYY) harmony learning acts as a general statistical learning framework, with not only a number of existing major learning problems and learning methods are revisited as special cases, but also a new learning mechanism that makes model selection implemented either *automatically* during parameter learning or *subsequently after* parameter learning via a new class of model selection criteria obtained from this mechanism, including their special cases for determining the numbers in $\mathbf{k} = \{k, \{\kappa_j\}\}$. Also, this BYY harmony learning has motivated three types of regularization. Readers are referred to [12, 13] for a systematical introduction.

The key idea of Bayesian Ying Yang system is to consider the joint distribution of x, y via two types of Bayesian decomposition of the joint density

$$p(\mathbf{x}, \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}),\ q(\mathbf{x}, \mathbf{y}) = q(\mathbf{x}|\mathbf{y})q(\mathbf{y}).$$

Without any constraints, the two decompositions should be theoretically identical. However, it usually not the case since the four components in the two decompositions are usually subject to certain structural constraints.

The *fundamental learning principle* is to make p, q be best harmony in a twofold sense:
• The difference between p, q should be minimized.
• p, q should be of the least complexity.

Mathematically, a functional $H(p\|q)$ is used to measure the degree of harmony between p and q, which is called harmony measure. That is, we have

$$\begin{aligned} &\max_{\theta,\mathbf{k}} H(\theta, \mathbf{k}),\ H(\theta, \mathbf{k}) = H(p\|q), \qquad (7)\\ &= \int p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \ln [q(\mathbf{x}|\mathbf{y})q(\mathbf{y})] d\mathbf{x}d\mathbf{y} - \ln z_q, \end{aligned}$$

where $\mathbf{k} = \{k, \{\kappa_j\}\}$ and θ consists of all the unknown parameters in $p(\mathbf{y}|\mathbf{x})$, $q(\mathbf{x}|\mathbf{y})$, and $q(\mathbf{y})$ as well as $p(\mathbf{x})$ (if any). The task of determining θ is called *parameter learning*, and the task of selecting $\mathbf{k}$ is called *model selection* since a collection of specific BYY systems with different values of $\mathbf{k}$ corresponds to a family of specific models that share a same system configuration but in different scales. Furthermore, the term $Z_q = -\ln z_q$ imposes regularization on learning, via three types of representations [13]. The simplest case is $z_q = 1$, which means that the term Z_q is neglected.

In our case, we simply consider $z_q = 1$ and

$$p_0(\mathbf{x}) = \tfrac{1}{N} \textstyle\sum_{t=1}^{N} \delta(\mathbf{x} - \mathbf{x}_t). \qquad (8)$$

Correspondingly, the learning is called empirical learning. Also, we have $q(\mathbf{y}) = q(\mathbf{y}|\psi)$ by eq.(3) and $q(\mathbf{x}|\mathbf{y}) = q(\mathbf{x}|\mathbf{y}, \phi)$ by eq.(2). Moreover, there are three typical system architectures due to combination of the structures for $q(\mathbf{x}|\mathbf{y}, q(\mathbf{y})$, and $p(\mathbf{y}|\mathbf{x})$. In this paper we consider a backward architecture featured by that $p(\mathbf{y}|\mathbf{x})$ is free to be determined via learning. With these above specific settings, it follows from eq.(7) that a free $p(\mathbf{y}|\mathbf{x})$ is determined as follows:

$$\begin{aligned} &p(\mathbf{y}|\mathbf{x}) = \delta(\mathbf{y} - \hat{\mathbf{y}}),\ \hat{\mathbf{y}} = arg \max_{\mathbf{y}} d(\mathbf{x}, \mathbf{y}), \qquad (9)\\ &d(\mathbf{x}, \mathbf{y}) = \ln G(\mathbf{x}|A\mathbf{y} + \mu, \Sigma) + \sum_{j=1}^{k} \ln q(y^{(j)}|\psi_j). \end{aligned}$$

Correspondingly, eq.(7) is simplified into

$$\max_{\theta,\mathbf{k}} H(\theta, \mathbf{k}),\ H(\theta, \mathbf{k}) = \tfrac{1}{N} \textstyle\sum_{t=1}^{N} d(\mathbf{x}_t, \hat{\mathbf{y}}_t), \qquad (10)$$

which can be implemented either in a parallel way such that model selection is made automatically during parameter learning (referred to [15, 14] for details) or in a two-phase implementation such that model selection is made after parameter learning, as will be discussed in the next subsection.

4.2 Learning algorithm and selection of $\mathbf{k}$

If we know all $\{\hat{\mathbf{y}}_t\}_{t=1}^N$, it follows from $\nabla H(\theta, k) = 0$ that we can get A, μ, Σ and all q_j solved analytically. However, getting $\hat{\mathbf{y}}_t$ per sample $\mathbf{x}_t$ is an integer programming task with θ known already. Thus, a better choice is getting parameters updated per sample coming, as suggested firstly in [16] and also further discussed in [15]. The details are introduced as follows.

As $\mathbf{x}_t$ comes, we first get $\hat{\mathbf{y}}_t$ by eq.(9), and then update A, μ, Σ to increase $\ln G(\mathbf{x}_t|A\hat{\mathbf{y}}_t + \mu, \Sigma)$ by the following adaptive updating rules

$$\begin{aligned} &e_t = \mathbf{x}_t - A^{old}\hat{\mathbf{y}}_t - \mu^{old},\ \mu^{new} = \mu^{old} + \eta e_t,\\ &A^{new} = A^{old} + \eta e_t \hat{\mathbf{y}}_t^T,\\ &\Sigma^{new} = (1 - \eta)\Sigma^{old} + \eta e_t e_t^T, \qquad (11) \end{aligned}$$

and update every ψ_j to increase $\sum_{j=1}^{k} \ln q(y^{(j)}|\psi_j)$ by

$$\forall i,\ \ \alpha_{ji}^{new} = \begin{cases} \frac{\alpha_{ji}^{old} + \eta}{1+\eta}, & if\ \hat{y}_t^{(j)} = \ell_i, \\ \frac{\alpha_{ji}^{old}}{1+\eta}, & otherwise. \end{cases} \qquad (12)$$

In the equations, $\eta >$ is a learning step size that can be selected differently for updating different parameter. In many cases, we can assume that the noise is white, i.e., $\Sigma = \sigma^2 I$, and its corresponding updating is simplified into

$$\sigma^{2\,new} = (1-\eta)\sigma^{2\,old} + \eta\|e_t\|^2. \quad (13)$$

In a summary, as each $\mathbf{x}_t$ comes we get $\hat{\mathbf{y}}_t$ by eq.(9) and implement eq.(11), eq.(12), and eq.(13).

Moreover, model selection, i.e., deciding $\mathbf{k}$, can be made in a two phase style. First, we enumerate $\mathbf{k}$ incrementally and at each specific value we get the best parameter value $\theta^*_{\mathbf{k}}$. Then, we select a best $\mathbf{k}^*$ by

$$\min_{\mathbf{k}} J(\mathbf{k}), J(\mathbf{k}) = -H(\theta_{\mathbf{k}^*}, \mathbf{k}), \quad (14)$$

If there are more than one values of $\mathbf{k}$ such that $J(\mathbf{k})$ gets the same minimum,we take the smallest. In our problem, it takes the following simplified form:

$$\min_{\mathbf{k}} J(\mathbf{k}), J(\mathbf{k}) = 0.5d\ln\sigma^2 - \sum_{j=1}^{k}\sum_{i=1}^{\kappa_j}\alpha_{ji}\ln\alpha_{ji}. \quad (15)$$

After learning, we make diagnose $\hat{\mathbf{y}}_t$ by eq.(9) on the symptoms in $\mathbf{x}_t$ per patient. Moreover, we can compute $p(\hat{\mathbf{y}}_t|\mathbf{x}_t)$ by eq.(5) to describe the degree of confidence of this diagnosing. However, the computation of $q(\mathbf{x}|\phi,\psi)$ is expensive. Alternatively, we can compare the best diagnosis $\hat{\mathbf{y}}_t$ by eq.(9) with its competing diagnosis ${}^c\mathbf{y}_t = arg\min_{\mathbf{y}\neq\hat{\mathbf{y}}_t} d(\mathbf{x}_t,\mathbf{y})$ in help of the following Bayesian factor ratio

$$B_f = \frac{p(\hat{\mathbf{y}}_t|\mathbf{x}_t)}{p({}^c\mathbf{y}_t|\mathbf{x}_t)} = e^{d(\mathbf{x}_t,\hat{\mathbf{y}}_t)-d(\mathbf{x}_t,{}^c\mathbf{y}_t)}, \quad (16)$$

which describes a degree of discriminative power of making diagnosing $\hat{\mathbf{y}}_t$. The larger the B_f is, the more confident to make the diagnosis $\hat{\mathbf{y}}_t$.

5. References

[1] Akaike, H. (1974), "A new look at the statistical model identification", *IEEE Tr. Automatic Control*, **19**, 714-723.

[2] Andrews, R., & Rubin, H. (1956), "Statistical inference in factor analysis", *Proc. Berkeley Symp. Math. Statist. Prob. 3rd 5*, UC Berkeley, 111-150.

[3] Aikins, J. S., J. C. Kunz, E. H. Shortliffe and R. J. Fallat, (1983), "PUFF: an expert system for interpretation of pulmonary function data", *Computers and Biomedical Research*, Vol.16, pp199-208, 1983.

[4] Bartholomew, D.J. and Knott, M. (1999), "Latent variable models and factor analysis", *Kendall's Library of Satistics, Vol. 7*, Oxford University Press, New York, 1999.

[5] Bozdogan, H. (1987) "Model Selection and Akaike's Information Criterion: The general theory and its analytical extension", *Psychometrika*, **52**, 345-370.

[6] Buchanan, B. G., and E. H. Shortliffe, (1984), "Rule-Based Expert Systems: The MYCIN Experiments", *the Stanford Heuristic Programming Project*, Reading, MA: Addison-Wesley, 1984.

[7] McDonald, R. (1985), *Factor Analysis and Related Techniques*, Lawrence Erlbaum.

[8] van Melle, W., (1982), " System Aids in Constructing Conultation Programs: EMYCIN", Ann Arbor, MI: UMI Research Press, 1982.

[9] Rissanen, J. (1999), "Hypothesis selection and testing by the MDL principle", *Computer Journal*, **42** (4), 260-269.

[10] Shim, J.Y. and Xu, L., (2002), "Oriental Medical Data Mining and diagnosis Based On Binary Independent Factor Model", *Advances in Neural Networks World* , A. Grmela and N.E. Mastorakis, eds, pp117-122.

[11] Shortliffe, Edward H., (1981), "Consultation Systems For Physicians: The Role of Artificial Intelligence Techniques". In Webber, Bonnie L.and Nilsson, Nils J. (Eds). *Readings in Artificial Intelligence*, Tioga Publishing Company. Palo Alto, CA, pp323-333, 1981.

[12] Xu, L (2004a), "Advances on BYY harmony learning: information theoretic perspective, generalized projection geometry, and independent factor autodetermination", *IEEE Trans on Neural Networks, Vol. 15, No. 4*, pp885-902.

[13] Xu, L. (2004b), "Bayesian Ying Yang Learning (I) & (II)",in *Intelligent Technologies for Information Analysis*, N. Zhong & J. Liu (eds), Springer, pp607-697, 2004.

[14] Xu, L. (2004c), "BI-directional BYY Learning for Mining Projected Polyhedra and Topological Map Structures", in press, *Proc. of IEEE ICDM2004 Workshop on Foundations of Data Mining*, Brighton, UK, Nov. 01 - 04, 2004.

[15] Xu, L. (2003), "BYY Learning, Regularized Implementation, and Model Selection on Modular Networks with One Hidden Layer of Binary Units", *Neurocomputing*, Vol.51, p227-301.

[16] Xu, L. (1998), "Bayesian Kullback Ying-Yang Dependence Reduction Theory", *Neurocomputing 22 (1-3)*, 81-112, 1998.

[17] Xu, L., (1995), "Bayesian-Kullback Coupled YING-YANG Machines: Unified Learnings and New Results on Vector Quantization", *Proc. Intl. Conf. on Neural Information Processing (ICONIP96)*, Oct 30-Nov.3, 1995, Beijing, pp.977-988.