New Advances on Bayesian Ying-Yang Learning System With Kullback and Non-Kullback Separation Functionals

Lei Xu
Dept of Computer Science, The Chinese University of Hong Kong, Hong Kong

Proceedings of 1997 IEEE-(INNS) Conference on Neural Networks,
Houston, TX, June. 9-12, Vol. 3, pp1942-1947, 1997.

Abstract

In this paper[1]*, we extend Bayesian-Kullback YING-YANG (BKYY) learning into a much broader Bayesian Ying-Yang (BYY) learning System via using different separation functionals instead of using only Kullback Divergence, and elaborate the power of BYY learning as a general learning theory for parameter learning, scale selection, structure evaluation, regularization and sampling design, with its relations to several existing learning methods and its developments in the past years briefly summarized. Then, we present several new results on BYY learning. First, improved criteria are proposed for selecting number of densities on finite mixture and gaussian mixtures, for selecting number of clusters in MSE clustering and for selecting subspace dimension in PCA related methods. Second, improved criteria are proposed for selecting number of expert nets in mixture of experts and its alternative model and selecting number of basis functions in RBF nets. Third, three categories of Non-Kullback separation functionals namely Convex divergence, L_p divergence and Decorrelation index, are suggested for BYY learning as alternatives for those learning models based on Kullback divergence, with some interesting properties discussed. As examples, the EM algorithms for finite mixture, mixture of experts and its alternative model are derived with Convex divergence.*

1. BYY Learning System and Theory

1.1 BYY Learning System

The learning problems by an information processing system can be summarized into the problem of estimating joint distribution $p(x,y)$ of the observable pattern x in the observable space X and its representation patter y in the representation space Y. We call a passage $M_{y|x}$ for the flow like $x \to y$ a *Yang*/(male) passage since it performs the task of transferring a pattern/(a real body) into a code/(a seed). We call a passage $M_{x|y}$ for the flow $y \to x$ as a *Ying*/(female) passage it performs the task of generating a pattern/(a real body) from a code/(a seed). $M_{y|x}$ and $M_{x|y}$ are complement to each other and together implement an entire circle $x \to y \to x$. Interestingly, under the Bayesian framework, we also have two representations $p(x,y) = p(y|x)p(x)$ and $p(x,y) = p(x|y)p(y)$. We use a Yang/(visible) model M_x representing $p(x)$ (i.e., modeling the space X), and we use a Ying/(invisible) model M_y representing $p(y)$ (i.e., modeling the space Y). Moreover, $M_{y|x}$ is represented by $p_{M_{y|x}}(y|x)$ and $M_{x|y}$ by $p_{M_{x|y}}(x|y)$. Together, we have a *YANG* machine $M_1 = \{M_{y|x}, M_x\}$ to implement $p_{M_1}(x,y) = p_{M_{y|x}}(y|x)p_{M_x}(x)$ and a *YING* machine $M_2 = \{M_{x|y}, M_y\}$ to implement $p_{M_2}(x,y) = p_{M_{x|y}}(x|y)p_{M_y}(y)$. A pair of YING-YANG machines is called a YING-YANG pair or a YING-YANG system. Such a formalization compliments to a famous Chinese ancient philosophy that *every entity in universe involves the interaction between YING and YANG.*

The task of specification of a Ying-Yang system is called *learning* in a broad sense. For this purpose, we need to specify four components $p_{M_x}(x)$, $p_{M_{y|x}}(y|x)$, $p_{M_{x|y}}(x|y)$ and $p_{M_y}(y)$ as well as the type and scale of variables x, y.

First, x is given by a practical problem without other choice, usually it is assumed that $x \in R^d$. But y can be real $y \in R^k$, integer $y \in [1, 2, \cdots, k]$ and binary $y = [y_1, \cdots, y_k]$, $y_i \in [0, 1]$, where y represents the complexity of representation space or equivalently the scale of a YING-YANG pair structure. **Second**, $p_{M_x}(x)$ is specified at some nonparametric estimate from a given training data set, one example is the kernel estimate (Devroye, 1987)

$$p_h(x) = \frac{1}{N}\sum_{i=1}^{N} K_h(x - x_i), \quad K_h(x) = \frac{1}{h^d}K(\frac{x - x_i}{h}). \tag{1}$$

where $\int |K(x)|dx < \infty$, $\int K(x)dx < 1$. **Next**, the rest three can be fixed or from a *parametric* family. Each is specified by both structure, e.g., density function form $p(x|y,\cdot)$ in $p_{M_{x|y}}(x|y) = p(x|y, \theta_{x|y})$, and parameter $\theta_{x|y}$.

For convenience, we denote $M_a = \{S_a, \theta_a, k\}$ for $a \in \{x|y,\ y|x,\ y\}$, i.e., M_a denotes a component with structure or desnity form S_a, parameter θ_a and scale k. Also, we denote $M_S = M_x = \{K, h, N\}$ with $K = K(x)$, smooth parameter h and sample size N.

The task of specifying $S = \{S_{x|y}, S_{y|x}, S_y\}$ is called *structural design.* The task of specifying k is called *scale selection.* The task of specifying $\Theta = \{\theta_{x|y}, \theta_{y|x}, \theta_y\}$ is called *parameter learning or estimation*, also called *learning* simply in a narrow sense. The task of specifying M_x is

[1] Supported by This project was supported by the HK RGC Earmarked Grants CUHK250/94E, CUHK484/95E, and Ho Sin-Hang Education Endowment Fund HSH 95/02.

A part of this work (particularly on those with Non-Kullback separation functionals) has been presented in my invited talks at WCNN96 and ICONIP96 in Sept. 15-18 and Sept. 24-27, 1996 respectively, although it is just formally published here.

called sampling design. All the four tasks together specify a Ying-Yang pair. The whole specification process can be regarded as a Ying-Yang interaction process with four possible types of *marital dynamics*: (a) *marry*, (b) *divorce*, (c) *YING chases & YANG escapes*, and (d) *YANG chases & YING escapes*, described by a combination of minimization (chasing) and maximization (escaping) on a so called *separation functional*:

$$F_s(M_1, M_2) = F_s(p_{M_{y|x}}(y|x)p_{M_x}(x), p_{M_{x|y}}(x|y)p_{M_y}(y)) \geq 0,$$
$$with\ F_s(M_1, M_2) = 0,\ if\ \ and\ only\ if$$
$$p_{M_{y|x}}(y|x)p_{M_x}(x) = p_{M_{x|y}}(x|y)p_{M_y}(y) \qquad (2a)$$

Since this system bases on the interaction between the two complement YING and YANG Bayesian representations, we call it *Bayesian Ying-Yang Learning System*. Particularly, when $F_s(M_1, M_2)$ is the Kullback divergence:

$$KL(M_1, M_2) =$$
$$\int_{x,y} p_{M_{y|x}}(y|x)p_{M_x}(x) \ln \frac{p_{M_{y|x}}(y|x)p_{M_x}(x)}{p_{M_{x|y}}(x|y)p_{M_y}(y)} dx\,dy \qquad (2b)$$

we return to Bayesian-Kullback YING-YANG (BKYY) Learning (Xu, 1995a&96a&c). In sec. 4, three categories of Non-Kullback separation functionals will be discussed.

Up to now, only the status *marry* and *divorce*, i.e., $\min_{M_1,M_2} F_s$ and $\max_{M_1,M_2} F_s$ have been studied. As shown in (Xu, 1997b), $\max_{M_1,M_2} F_s$ will result in $p_{M_{x|y}}(x|y) = p_{M_2}(x)$ and it can be further shown (Xu, 1996a&c) becomes maximization information preservation learning (Informax)(Linsker, 1989; Atick & Redlich, 1990) or its variants. Actually, the most useful one is $\min_{M_1,M_2} F_s$, which includes already the most useful special case of Informax, namely, Maximum output entropy. Therefore, we only consider it in this paper.

Generally, this $\min_{M_1,M_2} F_s$ is implemented by the *Alternative Minimization (ALTMIN)* iterative procedure

Step 1: Fix $M_2 = M_2^{old}$, get $M_1^{new} = \min_{M_1} F_s$.

Step 2: Fix $M_1 = M_1^{old}$, get $M_2^{new} = \min_{M_2} F_s$. (3)

which is guaranteed to converge (Xu, 1995a&96a&c).

The above system and theory can be directly applied to those unsupervised and supervised learnings for the information processing types like $x \to y$, $y \to x$, where x can be regarded consisting of two parts $x = (x, z)$ or even more when it is needed.

Moreover, for those tasks of focusing particularly on the relation $x \to z$, we still can use the above system with a slight extension. First, we replace all the x in eqs.(2a&b) by $(x, z)|x = z|x$ and all the y by $y|x$. Second, we notice that $(y|x)|(z|x) = y|(x, z, x) = y|(z, x)$ and $(z|x)|(y|x) = z|(y, x)$, which leads us to

$$F_s(M_1, M_2|x) =$$
$$F_s(p_{M_{y|z,x}}(y|z,x)p_{M_{z|x}}(z|x), p_{M_{z|y,x}}(z|y,x)p_{M_{y|z}}(y|z)) \geq 0,$$
$$with\ F_s(M_1, M_2) = 0,\ if\ \ and\ only\ if$$
$$p_{M_{y|z,x}}(y|z,x)p_{M_{z|x}}(z|x) = p_{M_{z|y,x}}(z|y,x)p_{M_{y|z}}(y|z)$$
$$F_s(M_1, M_2) = \int p_{M_x}(x))F_s(M_1, M_2|x)dx. \qquad (4a)$$
$$KL(M_1, M_2|x) =$$
$$\int_{x,y} p_{M_{y|z,x}}(y|z,x)p_{M_{z|x}}(z|x) \ln \frac{p_{M_{y|z,x}}(y|z,x)p_{M_{z|x}}(z|x)}{p_{M_{z|y,x}}(z|y,x)p_{M_{y|x}}(y|x)} dx\,dy$$
$$KL(M_1, M_2) = \int p_{M_x}(x)KL(M_1, M_2|x)dx. \qquad (4b)$$

where $M_1 = \{M_x, M_{y|z,x}, M_{z|x}\}$ and $M_2 = \{M_{z|y,x}, M_{y|x}\}$, $M_a = \{S_a, \theta_a, k\}$ for $a \in \{y|(z, x),\ z|(y, x),\ y|x\}$. Given a paired data set $\{x_i, z_i\}_{i=1}^N$, we usually let

$$p_{M_{z|x}}(z|x) = p_h(z|x_i) = K_h(z - z_i), \text{ at } x = x_i \qquad (4c)$$

with $K_h(x)$ is the same as in eq.(1). Thus, we still have sampling design $M_S = \{M_x, M_{z|x}\} = \{K, h, N\}$. As a whole, we have the entire structure $S = \{S_{y|z,x}, S_{z|x}, S_{z|y,x}, S_{y|z}\}$ and all the parameters $\Theta = \{\theta_{y|z,x}, \theta_{z|x}, \theta_{z|y,x}, \theta_{y|z}\}$.

1.2 A General Learning Theory

The different choices on specific structures, specific forms of separation functionals and sampling designs makes it possible to specify a large number specifications of a Ying-Yang system and thus a large number of specific learning models and theories. Therefore, we suggest that $\min_{M_1,M_2} F_s(M_1, M_2)$ functions as a unified general statistical learning theory for:

1. Parameter estimation or learning, *which is usually called* learning *in the narrow sense. That is, given S, k and M_S fixed, we determine*

$$\Theta^* = arg \min_{\Theta} F_s(\Theta : S, k, M_S). \qquad (5)$$

2. Scale selection, *or called model size selection. That is, given S, and M_S fixed, we determine*

$$k^* = arg \min_k \{\min_{\Theta} F_s(\Theta, k : S, M_S)\}. \qquad (6)$$

3. Structure evaluation. *That is, given M_S fixed, for two given sets of structures $S^{(1)}$ and $S^{(2)}$, We choose the first one if*

$$J(S^{(1)}) < J(S^{(2)}), \qquad J(S^{(i)}) = \min_k \{\min_{\Theta} F_s(\Theta, k : S^{(1)}, M_S)\} (7)$$

4. Sampling design. *It can be further divided into three. One is called* Sampling smoothing, *i.e., in* parameter learning, *we also adapt h to minimize F_s under given K, N. The second is called* Sampling structure evaluation, *i.e., given N, eq.(6) includes the evaluation on different $K^{(1)}$ and $K^{(2)}$. The third is* Sample complexity. *Given S, and K, h fixed, it can be made by*

$$N^* = E_x[\min_k \{\min_{\Theta} F_s\}]. \qquad (8)$$

where $E_x(J(x)) = \int_x p(x)J(x)dx$.

5. Regularization. *For a limited number N of samples, some regularization can be obtained by using one structure to constrain the others. For example, for a forward net or recognition model, we can design $S_{y|x}$ with more freedom to ensure its representation ability, but design $S_{x|y}$ with less freedom to regularize the learning to get a good generalization. Similarly, for a backward net or generative model, we can design $S_{x|y}$ with more freedom to ensure its representation ability, but design $S_{x|y}$ with less freedom to regularize the learning to get good generalization.*

It should be noted that this theory provides a unified general guideline that applies to all the specifications of Ying-Yang system. As shown previously in (Xu, 1995a&96a&c), as well partly in the latter sections of this paper, we will get the detailed forms of various special cases of this general theory for different specifications of Ying-Yang system. These specifications can be first classified into groups according *structure design*. Each of such groups is usually regarded as a different specific learning model/theory/method. Then, each of these groups can further have different realizations or individuals due to the difference in separation functionals, sampling designs and even the details of implementation algorithms. Therefore, we can first always fix the separation functional given at Kullback divergence and sampling design at an idealistic case of eq.(1) that $p_h(x) = \lim_{N\to\infty, hN\to\infty, h=h_N\to 0} p_h(x)$ (actually in this case, $p_h(x)$ will converge to $p_0(x)$—the original density that $\{x_i\}_{i=1}^N$ comes from), and then under this situation we explore various learning models or theories through different structural designs. Next, we use different separation functionals and sampling designs to get variants of these models or theories with some features.

1.3 The Power of The General Theory

To the current literature, this unified statistical learning theory can provide us at least the following strengths:

First, it is able to unify a quite number of existing major parameter learning models and theories for both supervised and unsupervised learning.

For *unsupervised learning*, as shown in (Xu, 1995a&96a&c), one of its special cases reduces to Maximum likelihood learning on finite mixture model with the EM algorithm and several related results, e.g., a cost function for mixture Gaussian by Hathaway (1986) and Neal & Hinton (1993), to the *Information Geometry* theory and the *em* algorithm by Amari and others(Amari, 1995) and others, to MDL autoencoder with a "bits-back" argument by Hinton & Zemel (1994). The special case can also reduce to multisets modeling learning (Xu, 1995d; Xu, 1994)–a unified learning framework for clustering, PCA-type learnings and self-organizing map. Its second special case reduces to the recent proposed Helmholtz machine (Dayan et al, 1995; Hinton et al, 1995) with new understandings. Its third special case gives a general Independent Component Analysis (ICA) framework (Xu & Amari, 1996) that unifies the information maximization (INFORMAX) approach (Bell and Sejnowski, 1995) and the minimum mutual information (MMI) approach (Amari, Cichocki, and Yang, 1996). Its another special design (Xu, 1996c) leads to LMSER learning and Principal Component Analysis (PCA) (Xu, 1991&93; Oja, 1989). Furthermore, some other special cases will also give us improved new learning models for ICA, linear and nonlinear LMSER learning as well as their localized extensions(Xu, 1997b).

For *supervised learning*, as shown in Xu (1996b&c), one special case includes the popular mixture of expert model (Jacobs, Jordan, Nowlan, & Hinton, 1991; Jordan & Jacobs, 1994) and its alternative model (Xu, Jordan & Hinton, 1994&95) as special cases, from which we can further get new algorithms for improving learnings on RBF networks (Xu, 1997c). Moreover, some special cases will not only lead us to the conventional maximum likelihood learning (or least square learning in particular) for a feedforward network, but also provide two new learning theories and algorithms as the alternatives of the traditional backpropagation algorithm (Xu, 1995a; 1997a).

The above powerful unification provides us not only deep insights on these mentioned popular existing learning models but also further guidance on obtaining their new variants or extensions via cross-fertilization, with some such examples already mentioned above.

Second, some special cases give us several interesting new unsupervised and supervised learning models or theories, which deserve further investigation (Xu, 1997a). Particularly, we can get a general scheme called *co-supervised learning* for handling those training data sets with one part consisting of both input and teaching target and the other part consisting of input only, so that supervised learning and unsupervised learning not only have been unified, but also coexist consistently for best exploiting the information in such a given data set (Xu, 1997a). Also in Sec. 4, we will show that by using the so called *Convex divergence, L_p divergence and Decorrelation index* to replace Kullback divergence, we can also generalize all the above mentioned Kullback divergence related models into a wide spectrum of different alternatives or extensions with some new interesting properties, such as becoming more robust. Furthermore, some effort has been also made on extending this theory to temporal patterns with a number of new models for signal modeling, cognition, prediction and segmentation. Some of them can be regarded as the extensions of Helmholtz machine or maximum information preservation learning to temporal processing. Some of them include and extend the existing Hidden Markov Model (HMM), AMAR and AR models (Xu, 1995b). Particularly, with the state space representation, it has been shown that this theory is equivalent to Kalman filter approach for the linear case, but outperforms Kalman filter and the existing extended Kalman filter considerably in nonlinear cases (Lei & Xu, 1997).

Third, although there are many theories and criteria available for *scale selection* on supervised learning (e.g., the number of hidden units in the feed-forward nets), how to do *scale selection* on unsupervised learning still remains open. This theory can function as a general scale selection theory for unsupervised learning, based on which criteria have been obtained in (Xu, 1995a) and then further refined (Xu, 1996b&c) for the selection of the number of Gaussians in a Gaussian mixture or the number of densities in a finite mixture, particularly, of the number of clusters in the conventional least mean square error (MSE) clustering analysis or vector quantization, e.g., by the k-means or LBG algorithm. Also based on this theory, a criterion has also been first obtained in (Xu, 1995c) and then refined in (Xu, 1996c) for the selection of the subspace dimension in *Principal Component Analysis (PCA)* related approach. In Sec.2, an improved version of this criterion will be given. Moreover, the theory can also solve the model scale selection problems of other Ying-Yang models for unsupervised learning (Xu, 1997b) and supervised learning (Xu, 1997a). Particularly, we can get new criteria for selecting the number of hidden units in feed-forward nets(Xu, 1997a). Furthermore, in (Xu, 1996c&d), criteria have been obtained for the selection of number of experts in the mixture of experts model and its alternative model, as well as of the number of basis functions in RBF nets; while in Sec. 3 of this paper, the improved versions of these criteria will be further proposed.

Fourth, although there are many theories and techniques for regularization on supervised learning, how to do *regularization* on unsupervised learning also remains open. As stated in the previous subsection, our theory also provides a new regularization scheme which applies to both unsupervised learning and supervised learning.

The last but not the least, as stated in Sec.1.2, the theory has also provide us a guide line to implement the more sophisticated evaluation of structure designing and sampling designing.

1.4 Relations to Other Approaches

The case $\min_{M_1,M_2} KL(M_1,M_2)$ relates to the well known information geometry theory (Amari, 1995; Byrne, 1992; Csiszar, 1975) in that both uses the Kullback divergence for measuring the difference between two joint densities. But, there are several key differences. *First*, the BYY learning theory considers the joint densities represented by two models, while the information geometry learning theory considers a missing data joint density and model joint density, which can be regarded as a special case of two models. *Second*, the BYY learning theory does not require joint densities necessarily in the exponential family, as required by the information geometry learning theory. *Third*, most importantly, the information geometry learning theory considers each joint density as an entire body without looking into their internal structures; while the BYY learning theory deliberately considers two complement but equivalent Bayesian structures for two joint densities. *Next*, the information geometry learning theory only considers the problem of parameter learning; while the BYY learning theory is proposed as a general unified theory for parameter learning, scale selection, structure evaluation, regularization and sampling design.

The sprit of considering the simultaneous modeling of the forward and backward passages in one system has been suggested by a number of previous researchers under the different formulations with different motivations, including ART theory and architecture (Carpenter & Grossberg, 1987), Pattern Theory (Grenander, 1976-1981); the Helmholtz machine (Hinton et al, 1995; Dayan et al, 1995), Forward-inverse model (Kawato, 1993). Mumford's pattern-theoretic architectures (Mumford, 1994), Bidirectional information flow cortex model(Ullman, 1994), Least MSE reconstruction principle (Xu, 1991; 1993). Being different from these existing efforts, the BYY learning theory attempts to formalize this sprit at a high level statistical theory that considers globally the whole probability distribution of the input pattern domain and its inner representation domain via two complement asymmetric Bayesian representations for the two passages and serves as a unified statistical learning theory. On one hand, it can include or closely relate some of these mentioned models, such as the Helmholtz machine and the Least MSE reconstruction principle. On the other hand, the relationship between the rest models remains unclear yet and deserves further exploration.

2. New Results on BKYY Unsupervised Learning

2.1 Improved number selection criteria for finite mixture and MSE clustering

Given the following design:

$$p_{M_{x|y}}(x|y) = p(x|S_{x|y}, \theta_y),\ p_{M_{y|x}}(y|x) = \sum_{j=1}^k P(j|x)\delta(y-j),$$
$$p_{M_y}(y) = \sum_{j=1}^k \alpha_j \delta(y-j) \text{ with } P(j|x) > 0, \alpha_j > 0,$$
$$\sum_{j=1}^k P(j|x) = 1 \text{ and } \sum_{j=1}^k \alpha_j = 1, \qquad (9)$$

It can be shown that as $Nh^d \to \infty$ and $h \to 0$, the minimization of $KL(M_1, M_2)$ given by eq.(2b) is equivalent to $\min_{\Theta_k} KL(\Theta_k)$ with $\Theta_k = \{\alpha_j, \theta_j\}_{j=1}^k$ and

$$KL(\Theta_k) = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^k P(j|x_i) \ln P(j|x_i)$$
$$-\frac{1}{N}\sum_{i=1}^N \sum_{j=1}^k P(j|x_i) \ln p(x_i|S_{x|j}, \theta_j) - \sum_{j=1}^k \alpha_j \ln \alpha_j \qquad (10a)$$

As shown in (Xu, 1995a, 1996a&b&c), this is equivalent to $\max_{\Theta_k} L(\Theta_k)$ with

$$L(\Theta_k) = \frac{1}{N}\sum_{i=1}^N p(x_i, \Theta_k),$$
$$p(x_i, \Theta_k) = \sum_{j=1}^k \alpha_j p(x_i|S_{x|j}, \theta_j) \qquad (10b)$$

and the ALTMIN eq.(3) is the same as the EM algorithm:

$$\text{E Step: } P(j|x_i) = \frac{\alpha_j p(x_i|S_{x|j}, \theta_j)}{p(x_i, \Theta_k)}, \quad \alpha_j = \frac{1}{N}\sum_{i=1}^N P(j|x_i).$$
$$\text{M Step: } \theta_j^{new} = \max_{\theta_j} \sum_{i=1}^N P(j|x_i) \ln p(x_i|\theta_j). \qquad (11)$$

That is, we get the maximum likelihood learning on finite mixture eq.(10b) with the EM algorithm eq.(11). Moreover, let $J(k) = \min_{\Theta_k} KL(\Theta_k)$, we can use $k^* = \min_k J(k)$ for selecting a correct scale k (Xu, 1995a, 1996a&b&c).

Moreover, assume that $\{x_i\}_{i=1}^N$ comes from

$$p^o(x, \Theta_{k^o}^o) = \sum_{j=1}^{k^o} \alpha_j^o p(x|S_{x|j}^o, \theta_j^o), \qquad (12)$$

as $Nh^d \to \infty$ and $h_N \to 0$, we can prove that $KL(\Theta_k)$ reaches its minimum, when $k = k^o$, $\Theta_{k^o}^o = \Theta_k$ and $S_{x|j} = S_{x|j}^o$, $j = 1, \cdots, k$. The condition also becomes necessary as long as $S_{x|j}^o$, $j = 1, \cdots, k$ satisfy a very mild regular condition (Xu, 1996c, 1997b). In this case, we have $J(k^o) < J(k)$ for $k \neq k^o$. Moreover, for the special case of the same structure $S_{x|j}^o$ (i.e., $p(x|S_{x|j}^o, \theta_j^o) = p(x|\theta_j)$), when the function's form $p(x|\cdot)$ in eq.(11) is the same one as in eq.(12), this regular condition will simply become that $p(x|\theta_j), j = 1, \cdots k$ are linear independent when θ_j is different from each other, which is automatically satisfied by Gaussian (Cheung & Xu, 1997).

However, when N is finite, actually $J(k)$ gradually reduces as k increases and becomes reducing very slowly after $k \geq k^o$. We can improve this weak point by dropping the first term in eq.(10a), i.e., let $\Theta_k^* = arg\min_{\Theta_k} KL(\Theta_k)$ given by eq.(10a), we define

$$k^* = \min_k J(k), \text{ and } J(k) = J(\Theta_k^*) =$$
$$-\frac{1}{N}\sum_{i=1}^N \sum_{j=1}^k P^*(j|x_i) \ln p(x_i|S_{x|j}, \theta_j^*) - \sum_{j=1}^k \alpha_j^* \ln \alpha_j^*. \qquad (13)$$

where $P^*(j|x_i)$ is given by eq.(11) at Θ_k^*. As $Nh^d \to \infty$ and $h_N \to 0$, under the same condition as above, we still can prove that $J(k^o) < J(k)$ for $k \neq k^o$:

Proof: From $KL(M_1, M_2) \geq 0$, we have $P(\Theta_k) \geq Q(\Theta_k)$:

$$P(\Theta_k) = \sum_{j=1}^k \int_x P(j|x)p^o(x, \Theta_{k^o}^o) \ln P(j|x)p^o(x, \Theta_{k^o}^o)dx$$
$$Q(\Theta_k) = \sum_{j=1}^k \int_x P(j|x)p^o(x, \Theta_{k^o}^o) \ln [\alpha_j p(x|S_{x|j}, \theta_j)]dx$$

where the equality holds if and only if $P(j|x_i) = \alpha_j p(x_i|S_{x|j}, \theta_j)/p^o(x, \Theta_{k^o}^o)$. Naturally, it is the case when $k = k^o$ and $P^o(j|x_i) = \alpha_j^o p(x_i|S_{x|j}^o, \theta_j^o)/p^o(x, \Theta_{k^o}^o)$, in which we have $P(\Theta_{k^o}^o) = Q(\Theta_{k^o}^o)$:

$$P(\Theta_{k^o}^o) = \sum_{j=1}^k \int_x P^o(j|x)p^o(x, \Theta_{k^o}^o) \ln P^o(j|x)p^o(x, \Theta_{k^o}^o)dx$$
$$Q(\Theta_k^o) = \sum_{j=1}^k \int_x P^o(j|x)p^o(x, \Theta_{k^o}^o) \ln [\alpha_j^o p(x_i|S_{x|j}^o, \theta_j^o)]dx$$

Next, we consider $k \neq k^o$. From $\sum_{j=1}^k P(j|x_i) = 1$, we have

$$\sum_{j=1}^k \alpha_j p(x|S_{x|j}, \theta_j) = p^o(x, \Theta_{k^o}^o) = \sum_{j=1}^{k^o} \alpha_j^o p(x|S_{x|j}^o, \theta_j^o). \quad (14)$$

Under the above same mild regular condition (Xu, 1996c, 1997b), or when $p(x|S_{x|j}, \cdot) = p(x|S_{x|j}^o, \cdot) = p(x|\cdot)$ as long as $p(x|\theta_j), j =$

$1, \cdots k$ are linear independent when θ_j is different from each other, we can observe that eq.(14) can not true when $k < k^o$. That is, in this case we have $P(\Theta_k) > Q(\Theta_k)$. While, when $k > k^o$, eq.(14) holds only when $p(x|S_{x|j_r}, \theta_{j_r}) = p(x|S^o_{x|r}, \theta^o_r), r = 1, \cdots, k^o$ and for some r there are at least two $r_1 \neq r_2$ such that $p(x|S_{x|j_{r_1}}, \theta_{j_{r_1}}) = p(x|S_{x|j_{r_2}}, \theta_{j_{r_2}}) = p(x_i|S^o_{x|r}, \theta^o_r)$ with $\alpha_{j_{r_1}} > 0, \alpha_{j_{r_2}} > 0$ and $\alpha_{j_{r_1}} + \alpha_{j_{r_2}} = \alpha^o_r$. Thus, from eq.(14) we have $P(j_{r_1}|x) > 0, P(j_{r_2}|x) > 0,\ P(j_{r_1}|x) + P(j_{r_2}|x) = P^o(r|x)$ and

$$\sum_{j=1}^{k^o} P^o(j|x) \ln P^o(j|x) > \sum_{j=1}^{k} P(j|x) \ln P(j|x)$$

which in turn leads us to $P(\Theta_k) < P(\Theta^o_{k^o})$.

In summary, we have that $P(\Theta^o_{k^o}) \geq Q(\Theta_k)$ with the equality only when $k = k^o, \Theta_k = \Theta^o_{k^o}$. That is, where $-Q(\Theta_k)$ reaches its maximum $P(\Theta^o_{k^o})$. Let $J(k) = \min_{\Theta_k}\{-Q(\Theta_k)\}$ with $p^o(x, \Theta^o_{k^o}) = \lim_{hN \to \infty, h = h_N \to 0} p_h(x)$, we can get eq.(13). In other words, as $Nh^d \to \infty$ and $h_N \to 0$, $J(\Theta_k)$ converges to $-Q(\Theta_k)$. **Q.E.D.**

This $J(k)$ given by eq.(13) improves the original one when N is finite. The reason is that the first term in eq.(10a) is always negative and reduces quickly after $k \geq k^o$. After dropping it, $J(k)$ becomes increasing after $k > k^o$, even for the finite N, as shown by the experimental results given in (Xu, 1997a). In fact, from the term $\alpha_j p(x_i|S_{x|j}, \theta_j)$ in $Q(\Theta_k)$, we can also regard this improved scale selection criterion is a Bayesian model selection criterion for unsupervised learning that considers α_j.

In the special case of Gaussian $p(x_i|S_{x|j}, \theta_j) = G(x, m_j, \Sigma_j)$, $J(k)$ given by eq.(13) simply becomes

$$J_G(k) = J(\Theta^*_k) = \sum_{j=1}^k \alpha^*_j \ln \sqrt{|\Sigma^*_j|}/\alpha^*_j. \qquad (15)$$

from which we can get various special criteria for various special cases of α_j, Σ_j.

Furthermore, if we hardcut $P^*(j|x_i)$ into $I^*(j|x_i)$ with $I^*(j|x_i) = 1$ for $j = arg \max_r P^*(r|x_i)$ and otherwise $I^*(j|x_i) = 0$, we can have that $J(k)$ given by eq.(13) becomes exactly $J_h(k) = J_h(\Theta(k))$ given by Eq.(10) in Xu (1996c), which includes the criterion $J^h_G(k)$ given by Eq.(11b) in Xu (1996c) for the MSE clustering by well known k-means algorithm.

This hardcut is a heuristic treatment, and thus the above conclusion that $J(k^o) < J(k)$ for $k \neq k^o$ is not automatically true even as $Nh^d \to \infty$ and $h_N \to 0$.

To have the conclusion to be true, it must satisfy

$$J(k) - J(k^o) > e(k^o) - e(k), \text{ with } e(k) = J_h(k) - J(k) \qquad (16a)$$

This condition is difficult to test. We further refine it into

$$J(k) - J(k^o) > \max\{|e(k^o)|, |e(k)|\}, \qquad (16b)$$

and the upper bound of $|e(k)|$ can be estimated via a further condition

$$|I^*(j|x) - P^*(j|x)| < \xi < 1 \qquad (16c)$$

from which we can specify some specific condition on $\Theta^o_{k^o}$. This process is usually quite complicated. However, for gaussian mixture, especially the case corresponding to the criterion $J^h_G(k)$ given by Eq.(11b) in Xu (1996c), it is possible to get some sufficient condition on $\Theta^o_{k^o}$ to make eq.(16a) hold.

2.2 An improved criterion for subspace dimension

We consider the design that (a) $p_{M_x}(x) = p_h(x)$ by eq.(1); (b) $p_{M_y}(y) = G(y, 0, \Sigma_y)$ with $\Sigma_y = diag[\lambda_1, \cdots, \lambda_k]$, $\lambda_1 > \cdots > \lambda_k > 0$; (c) $p_{M_{x|y}}(x|y) = G(e_x, W^t y, \sigma^2_{x|y} I)$ with $x = W^t y + e_x$, where e_x is a Gaussian noise with $E(e_x) = 0$, $E(e_x e_x^t) = \sigma^2_{x|y} I_d$, and $E(e_x y^t) = 0$; (d) $p_{M_{y|x}}(y|x) = G(e_y, Wx, \sigma^2_{y|x} I_k)$, with $y = Wx + e_y$, where e_y is a Gaussian noise with $E(e_y) = 0$, $E(e_y e_y^t) = \sigma^2_{y|x} I_k$, and $E(e_y x^t) = 0$.

With this design, as shown in Xu(1995c&96c), we get

$$J(W, k) = 0.5\{\ln \frac{|\Sigma_y|\sigma^{2n}_{x|y}}{\sigma^{2k}_{y|x}} + \frac{1}{\sigma^2_{x|y}}(E_2(W) + k\sigma^2_{y|x})\} + const$$
$$E_2(W) = \int_x p_{M_x}(x)||x - W^t W x||^2 dx \qquad (17)$$

For a fixed k, we see that min $_W J(W, k)$ is equivalent to min $_W E_2(W)$, which is the LMSER self-organization (Xu, 1991& 1993) that performs Principal subspace analysis (Oja, 1989), i.e, its solution W^* satisfies $W^* W^{*t} = I$ and spans the subspace spanned by the k principal components of data on x.

With this $W^* W^{*t} = I$, from $x = W^{*t} y + e_x$ we can get $y = W^* x + e_y = W^* x - W^* e_x$ with $e_y = -W^* e_x$. Thus, e_y is also a Gaussian with $E(e_y) = 0$ and covariance $E(W^* e_x e_x^t W^{*t}) = \sigma^2_{x|y} I_k = \sigma^2_{y|x} I_k$. That is, $\sigma^2 y|x = \sigma^2 x|y = \sigma^2$. We put this together with $W^* W^{*t} = I$ into eq.(17), we can define $J(k) = \min_W J(W, k) + const$. That is

$$J(k) = \ln[|\Sigma_y|\sigma^{2(n-k)}] + k + \frac{1}{\sigma^2} E_2(W^*) \qquad (18)$$

Furthermore, from $x = W^{*t} y + e_x$, $y = W^* x + e_y$, and $e_y = -W^* e_x$, we have $x - W^{*t} W^* x = e_x - W^{*t} W^* e_x$. Moreover, we have $E_2(W^*) = \int_x p_{M_x}(x)\|x - W^{*t} W^* x\|^2 dx = Tr[E(I - W^{*t} W^*) e_x e_x^t (I - W^{*t} W^*)^t] = \sigma^2(n - k)$, and thus

$$J(k) = \ln|\Sigma_y| + (n - k) \ln \frac{E_2}{(n-k)}. \qquad (19)$$

From $E(e_y x^t) = 0$ and $E(e_y e_y^t) = \sigma^2_{y|x} = \sigma^2$, we have

$$\Sigma_y = E(yy^t) = E[(W^* x + e_y)(W^* x + e_y)^t] = W^* S W^{*t} + \sigma^2,$$

where $S = \frac{1}{N}\sum_{i=1}^N x_i x_i^t$ and $W^* S W^{*t}$ is a diagonal matrix consisting of $\lambda^x_1 \geq \lambda^x_2 \geq \cdots \geq \lambda^x_k$ — the first k largest eigenvalues of S. Therefore, eq.(19) further becomes

$$J(k) = \sum_{j=1}^k \ln(\lambda^x_i + \frac{E_2(W^*)}{n-k}) + (n - k) \ln \frac{E_2(W^*)}{(n-k)} \qquad (20)$$

with $E_2(W^*)$ estimated by $\frac{1}{N}\sum_{i=1}^N \|x_i - W^t W x_i\|^2$ and λ^x_i from $S = \frac{1}{N}\sum_{i=1}^N x_i x_i^t$ via eigen-analysis or PCA learning.

Therefore, we can finally use this simplified criterion for selecting the dimension of subspace in the PCA related subspace analysis. That is, we select k^* if $k^* = arg \min_k J(k)$, which is an improvement of the old one given in Xu(1996c).

3. New Results for Supervised Learning

1. Improved number selection criteria over that given in (Xu, 1996c&d) are obtained for the model of mixture of experts (Jacobs, Jordan, Nowlan & Hinton, 1991; Jordan & Jacobs, 1994).

We design that: (i) $p_h(x)$ is by eq.(1) and $p_{M_{z|x}}(z|x)$ by eq.(4c); (ii) $p_{M_{y|z,x}}(y|z_i, x_i) = \sum_{j=1}^k P(j|x_i)\delta(y - j)$ with $\sum_{j=1}^k P(j|x) = 1$; (iii) $P_{M_{y|x}}(y|x) = \sum_{j=1}^k \delta(y - j)P(j|x, \psi)$, $p_{M_{z|y,x}}(z|y, x) = p_{M_{z|y,x}}(z|y, x, \theta_y)$. By putting the design into eq.(4b), it can be shown that as $Nh^d \to \infty$ and $h \to 0$, the minimization of $KL(M_1, M_2)$ by eq.(4b) is equivalent to $\min_{\Theta_k} J(\Theta_k)$ with $\Theta_k = \{\psi, \theta_j\}_{j=1}^k$ and

$$J(\Theta_k) = \frac{1}{N}\sum_{i=1}^N \sum_{j=1}^k P(j|x_i) \ln \frac{P(j|x_i)}{P(z_i|x_i, j, \theta_j)P(j|x_i, \psi)} \qquad (21)$$

As shown in (Xu, 996c&d), this is equivalent to $\max_{\Theta_k} L(\Theta_k)$:

$$L(\Theta_k) = \frac{1}{N}\sum_{i=1}^{N} \ln p(z_i|x_i, \Theta_k),$$
$$p(z|x, \Theta_k) = \sum_{y=1}^{k} P(y|x, \psi) p(z|x, y, \theta_y) \qquad (21)$$

and the ALTMIN eq.(3) is the same as the EM algorithm:

$$\text{E Step: } P(y|x_i) = \frac{P(y|x_i, \psi)p(z_i|x_i, y, \theta_y)}{p(z_i|x_i, \Theta(k))}.$$
$$\text{M Step: } \quad \theta_y^{new} = \max_{\theta_y} \sum_{i=1}^{N} P(y|x_i) \ln p(z_i|x_i, y, \theta_y).$$
$$\psi^{new} = \max_{\psi} \sum_{i=1}^{N} \sum_{y=1}^{k} P(y|x_i) \ln P(y|x_i, \psi). \qquad (22)$$

That is, we get maximum likelihood learning for the original model of mixture of experts with the EM algorithm eq.(22). Similar to what we did in Sec.2.1, we get $\Theta_k^* = arg \min_{\Theta_k} J(\Theta_k)$ by eq.(21), and define $J(k) = J(\Theta_k^*)$:

$$J(k) = -\frac{1}{N}\sum_{i=1}^{N} \sum_{y=1}^{k} P^*(y|x_i) \ln [p(z_i|x_i, y, \theta_y)P(y|x_i, \psi)], \quad (23)$$

as an improved criterion for selecting $k^* = \min_k J(k)$ as the number of experts required in a mixture expert model. Also similar to what we did in Sec.2.1, we can prove that $J(k^o) \leq J(k)$ for $k \neq k^o$.

When the regression error of each expert is gaussian, i.e., $p(z|x, y, \theta_y) = G(z, f(x, W_y), \Sigma_y)$, eq.(23) can be simplified into

$$J_G(k) = -\frac{1}{N}\sum_{i=1}^{N} \sum_{j=1}^{k} P^*(j|x_i) \ln P(j|x_i, \psi^*),$$
$$+\frac{1}{2}\sum_{j=1}^{k} \alpha_j^* \ln |\Sigma_j^*|, \quad \alpha_j^* = \frac{1}{N}\sum_{i=1}^{N} P^*(j|x_i),$$
$$\Sigma_j^* = \frac{1}{\alpha_j^* N}\sum_{i=1}^{N} P^*(y|x_i)[z_i - f(x_i, W_j^*)][z_i - f(x_i, W_j^*)]^t, \quad (24)$$

2. Improved number selection criteria over that given in (Xu, 1996c&d) are obtained for the alternative model of mixture of experts (Xu, Jordan, & Hinton, 1994, 1995), which replace $P(y|x_i, \psi)$ by

$$P(y|x_i, \psi) = p(x|y, \psi_y)\alpha_y / \sum_{y=1}^{k} p(x|y, \psi_y)\alpha_y \qquad (25)$$

and the M step in the EM algorithm eq.(22) is replaced by

$$\text{M Step: } \alpha_y = \frac{1}{N}\sum_{i=1}^{N} P(y|x_i).$$
$$\theta_y^{new} = \max_{\theta_y} \sum_{i=1}^{N} P(y|x_i, \psi) \ln p(z_i|x_i, y, \theta_y).$$
$$\psi_y^{new} = \max_{\psi} \sum_{i=1}^{N} p(x_i|y, \psi_y) \ln p(x_i|y, \psi_y). \qquad (26)$$

which actually maximizes $L(\Theta_k) = \frac{1}{N}\sum_{i=1}^{N} p(z, x, \Theta_k)$ with $p(z, x, \Theta_k) = p(x)p(z|x, \Theta_k)$. Accordingly, we can get criteria eq.(23) and eq.(24) become respectively:

$$J(k) = -\sum_{y=1}^{k} \alpha_y^* \ln \alpha_y^*$$
$$-\frac{1}{N}\sum_{i=1}^{N} \sum_{y=1}^{k} P^*(y|x_i) \ln \{p(z_i|x_i, y, \theta_y^*)p(x|y, \psi_y^*)\}, \qquad (27a)$$
$$J_G(k) = -\sum_{y=1}^{k} \alpha_y^* \ln \alpha_y^*$$
$$+\frac{1}{2}\sum_{y=1}^{k} \alpha_y^* \ln |\Sigma_y^*| - \frac{1}{N}\sum_{i,y} P^*(y|x_i) \ln P(x|y, \psi_y^*). \qquad (27b)$$

3. An improved number selection criteria over that given in (Xu, 1996c&d) is obtained for the normalized Radial Basis Function (RBF) nets given by Eqs.(20a&b) in (Xu, 1996c), which are the special case of the above alternative model of mixture experts respectively at

$$P(z|x, y, \theta_y) = G(z, c_y, \Sigma_{zy}, \; P(z|x, y, \theta_y) = G(z, \theta_y^t x + c_y, \Sigma_{zy})$$
$$P(y|x_i, \psi) = \frac{exp[-0.5(x - m_y)\Sigma_x^{-1}(x - m_y)]}{\sum_{y=1}^{k} exp[-0.5(x - m_y)\Sigma_x^{-1}(x - m_y)]}, \qquad (28)$$

with $\alpha_y / |\Sigma_x| = const$. We can get an improved criterion:

$$J_G(k) = \frac{1}{2}\ln |\Sigma_x^*| + \frac{1}{2k}\sum_{y=1}^{k} \ln |\Sigma_{zy}^*| + \ln k \qquad (29)$$

for the two cases, where $\Sigma_{zy}^*, \Sigma_x^*$ are obtained directly from the learning by EM algorithm.

In Xu(1997c), a number of algorithms, including EM, hard-cut EM, adaptive learning, have been proposed for efficiently training the alternative model of mixture experts and RBF networks.

4. Three Categories of Separation Functionals and Their Related Learning Models

(This part of work has been presented in my invited talks at WCNN96 and ICONIP96 in Sept. 15-18 and Sept. 24-27, 1996 respectively, although have not been included in (Xu, 1996c&d). Instead, a more general non-Kullack functional f(.) was discussed in (Xu, 1996e) for a learning principle related to BYY learning with divorcing dynamic, where convex function f(.) is discussed in its Sec.4. Actually, it was that work motivated the work given here.)

1. Convex Divergence We consider a class of functionals on two densities $p_1(x), p_2(x)$ given by

$$F_s(p_1, p_2) = f(1) - \int_x p_1(x) f(\frac{p_2(x)}{p_1(x)})dx \geq 0,$$
$$f(u) \; is \; a \; strict \; convex \; on \; (0, +\infty) \qquad (30)$$

Obviously $F_s(p_1, p_2) = 0$ when $p_1 = p_2$. We have also $F_s(p_1, p_2) > 0$ when $p_1 \neq p_2$ since $\int_x p_1(x) f(\frac{p_2(x)}{p_1(x)})dx \leq f(\int_x p_1(x)\frac{p_2(x)}{p_1(x)}dx) = f(1)$. Substituting p_1 by $p_{M_{y|x}}(y|x)p_{M_x}(x)$ and p_2 by $p_{M_{x|y}}(x|y)p_{M_y}(y)$ in eq.(2a), we have $F_s(M_1, M_2)$.

There are some typical examples of this convex divergence:

(a) $f(u) = \ln u$ which leads us to Kullback Divergence.

(b) $f(u) = -u^\beta, \beta > 1$, called as Minus Convex divergence.

(c) When $f(u) = u^\beta, 0 < \beta < 1$, we called as Positive Convex (PC) divergence. One of its particular interesting case is that $\beta = 0.5$, which leads to a symmetric Root-Inner-Product (RIP) divergence:

$$F_s(p_1, p_2) = 1 - \int_x \sqrt{p_1(x)p_2(x)}dx \qquad (31)$$

Remarks: (1) We will have a similar situation for $f(u)$ being strict concave on $(0, +\infty)$ since $-f(u)$ is a strict concave when $f(u)$ is a strict convex. (2) RIP has a nice symmetric feature that the Kullback divergence does not have. (3) Non-Kullback cases of convex divergence may lose a favorable feature of Kullback divergence of holding partially triangle inequality (Amari, 1995), which can define an orthogonal projection.

2. L_p Divergence L_p distance may also be extended as the separation functional: $F_s(M_1, M_2) =$

$$\int_{x,y} p(x)|g(p_{M_{y|x}}(y|x)p_{M_x}(x)) - g(p_{M_{x|y}}(x|y)p_{M_y}(y))|^p dx dy \quad (32)$$

where $g(u)$ is any function such that $g(p_1(x)) = g(p_2(x))$ if and only if $p_1(x) = p_2(x)$.

3. De-correlation Index The correlation coefficient is extended into: $F_s(M_1, M_2) =$

$$1 - \frac{\int_{x,y} p(x) g_{y|x}(x, y) g_{x|y}(x, y) dx dy}{\sqrt{\int_{x,y} p(x) g^2 g_{y|x}(x, y) dx dy}\sqrt{\int_{x,y} p(x) g^2 g_{x|y}(x, y) dx dy}} \geq 0$$
$$g_{y|x}(x, y) = g(p_{M_{y|x}}(y|x)p_{M_x}(x)),$$
$$g_{x|y}(x, y) = g(p_{M_{x|y}}(x|y)p_{M_y}(y)) \qquad (33)$$

where $g(u)$ is any function such that $g(u_1)g(u_2) > 0$, if $u_1 > 0 \; u_2 > 0$ and $g(p_1(x)) = g(p_2(x))$ iff $p_1(x) = p_2(x)$. Particularly when $g(u) = u$, we have the ordinary De-correlation Index.

4. Examples of Related Learning Models. As mentioned at the end of Sec.2.1, using the three categories of separation functionals to replace Kullack divergence, we can, at least theoretically, obtain their counterparts of those models using Kullack divergence. Due to space limit, here we only consider two examples.

Robust Finite Mixture. By putting the design eq.(9) into eq.(2a) instead of eq.(2b), although we can not get the expanded form like eq.(10a), we can still get that the minimization of $F_s(M_1, M_2)$ with the fixed M_2 will results in the E step in the EM algorithm eq.(11). With this obtained $P(j|x_i)$ put into eq.(2a), we consider the case of Convex Divergence and get the form like eq.(10b):

$$L_f(\Theta_k) = \frac{1}{N}\sum_{i=1}^{N} f(p(x_i, \Theta_k)) = \frac{1}{N}\sum_{i=1}^{N} f(e^{\ln p(x_i, \Theta_k)}) \qquad (34)$$

That is, the minimization of $F_s(M_1, M_2)$ is equivalent to the maximization of $L_f(\Theta_k)$. So we get a generalized ML learning for finite mixture.

When $f(u)$ is monotonically increasing for positive u, e.g., $f(u) = u^{\beta}$, $0 < \beta < 1$, $f(e^u) = e^{\beta \ln u}$ is also a monotonically increasing for $\xi \in (-\infty, 0]$. Since this $e^{\beta\xi}$ puts more attention on the value of ξ near 0, and the maximization of $L_f(\Theta_k)$ gives more weights to those samples with $p(x_i, \Theta_k))$ near 1. In other words, the learning is more relayed on those samples around each density center, while those boundary samples are discounted. Thus, the learning will give more robust estimation by discounting outliers. Since the more close the β around 1, the more rapid the $e^{\beta\xi}$ changes around 1, the more robust the learning will be.

On the other hand, when $f(u)$ is monotonically decreasing for positive u, e.g., $f(u) = -u^{\beta}, \beta > 1$, the effect is the minimization of $e^{-\beta\xi}, \xi = \ln p(x_i, \Theta_k)$, $e^{-\beta\xi}$ is also monotonically decreasing for $\xi \in (-\infty, 0]$. That is, more attention is put on those boundary values. The learning seems to emphasizes more on the discrimination between the samples from different models. However, whether it has this type feature is still not quite clear yet and needs to be further explored.

From the E step of eq.(11) and the fact that

$$\frac{df(e^{\ln p(x_i, \Theta_k)})}{d\theta_j} = f'(p(x_i, \Theta_k))\alpha_j p(x_i|S_{x|j}, \theta_j)\frac{d\ln p(x_i|S_{x|j}, \theta_j)}{d\theta_j}$$

$$= f'(p(x_i, \Theta_k))p(x_i, \Theta_k)P(j|x_i)\frac{d\ln p(x_i|S_{x|j}, \theta_j)}{d\theta_j},$$

we get that the M Step in eq.(11) should be replaced by

M Step: θ_j^{new} is given by solving

$$\sum_{i=1}^{N} f'(p(x_i, \Theta_k))p(x_i, \Theta_k)P(j|x_i)\frac{d\ln p(x_i|S_{x|j}, \theta_j)}{d\theta_j} = 0 \qquad (35)$$

Particularly, for Gaussian $p(x_i|S_{x|j}, \theta_j) = G(x, m_j, \Sigma_j)$, it is explicitly given as:

$$\begin{aligned} MStep: \quad & w(y, x_i) = f'(p(x_i, \Theta_k))p(x_i, \Theta_k)P(j|x_i) \\ & m_y^{new} = \frac{1}{N}\sum_{i=1}^{N} w(y, x_i)x_i, \qquad (36) \\ & \Sigma_y^{new} = \frac{1}{N}\sum_{i=1}^{N} w(y, x_i)(x_i - m_y^{new})(x_i - m_y^{new})^t \end{aligned}$$

we actually get a weighted variant of the EM algorithm. When $f(u)$ is monotonically increasing for positive u, we call eq.(35) as *Robust EM (REM)* algorithm for finite mixture, and eq.(36) as *Robust EM (REM)* algorithm for Gaussian mixture.

Robust Mixtures of Experts. For the mixture of expert model by eq.(21), we can similarly get $L_f(\Theta_k) =$

$$\frac{1}{N}\sum_{i=1}^{N} f(p(z_i|x_i, \Theta_k)) = \frac{1}{N}\sum_{i=1}^{N} f(e^{\ln (p(z_i|x_i, \Theta_k)}). \qquad (37)$$

Moreover, the M Step in eq.(22) should be replaced by

M Step: θ_j^{new} is given by solving

$$\sum_{i=1}^{N} f'(p(z_i|x_i, \Theta_k))p(z_i|x_i, \Theta_k)P(j|x_i)\frac{d\ln p(z_i|x_i, j, \theta_j)}{d\theta_j} = 0 \quad (38)$$

and ψ^{new} is given by solving

$$\sum_{i=1}^{N}\sum_{j=1}^{k} f'(p(z_i|x_i, \Theta_k))p(z_i|x_i, \Theta_k)P(j|x_i)\frac{d\ln P(j|x_i, \psi)}{d\psi} = 0 \ (39)$$

with the E step being still the same as that in eq.(22) to get $P(j|x_i)$. Similarly, the M Step in eq.(26) should be replaced by

M Step: get α_y by eq.(26), get θ_j^{new} by solving eq.(38) and get ψ_j^{new} by solving

$$\sum_{i=1}^{N} f'(p(z_i|x_i, \Theta_k))p(z_i|x_i, \Theta_k)P(j|x_i)\frac{d\ln p(x_i|j, \psi_j)}{d\psi_j} = 0 \qquad (40)$$

References

(Due to limited space, a number of references are omitted here, and can be find in the reference lists of Xu, 1995a; 1996a&b&c&d)

Amari, S(1995), *Neural Networks 8*, No.9, 1379-1408.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977), *J. Royal Statist. Society, B39*, 1-38.

Carpenter, G.A. and Grossberg, S.(1987), *Computer Vision, Graphics, and Image Processing, V37*, pp.54-115.

Cheung, Y.M., and Xu, L (1997), in this *Proc. of IEEE ICNN97.*

Devroye, L.(1987), *A Course in Density Estimation*, Birhhauser.

Grenander, U (1976-1981), *Lectures in Pattern Theory I, II and III: Pattern Analysis, Pattern Synthesis and Regular Structures*, Berlin: Springer-Verlag., Berlin, 1976-1981.

Dayan, P., Hinton, G. E., & Neal, R. N. (1995), *Neural Computation* Vol.7, No.5, 889-904.

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977), *J. Royal Statist. Society, B39*, 1-38.

Hinton, G. E., et al, (1995), *Science 268*, pp1158-1160.

Jacobs, R.A., Jordan, M.I., Nowlan, S.J., and Hinton, G.E., (1991), *Neural Computation, 3*, pp 79-87

Jordan, M. & Jacobs, R. (1994), *Neural Computation 6*, 181-214.

Kawato, M. et al (1993), *Networks 4*, pp415-422.

Lei, Y.Q and Xu, L. (1997), "Linear And Nonlinear Filtering Based on the Principle of Ying-Yang Machine ", submitted to a journal.

Mumford, D (1994), In C.Koch and J.Davis eds, *Large-Scale Theories of the Cortex*, Cambridge, MA: MIT Press, pp12-152.

Oja.E.(1989), *Int. J. Neural Systems 1*, 1989, 61-68.

Ullman, S. (1994), In the same eds as above, pp257-270.

Xu, L. (1997a), "Unsupervised, Supervised and Cosupervised Bayesian Ying-Yang Learning: New Developments", to appear on *Lecture Notes in Computer Science*, *Proc. Intl Workshop on Theoretical Aspects of Neural Computation*, May 26-28, Hong Kong, 1997, Springer-verlag.

Xu, L. (1997b), "Bayesian Ying-Yang Learning, LMSER Self-Organization and Independent Component Analysis", to appear on *Proc. ICONIP97*, 24-28, Nov, 1997, Dunedin, New Zealand.

Xu, L. (1997c), "Batch and Adaptive Matched Competitive Learning Algorithms for Mixture of Experts and RBF Networks", submitted to a Journal.

Xu, L. (1996a), "A Unified Learning Scheme: Bayesian-Kullback YING-YANG Machine", *Advances in NIPS 8*, David S. Touretzky, Michael C. Mozer and Michael E. Hasselmo, eds, MIT Press: Cambridge, MA, pp444-450.

Xu, L. (1996b), "How Many Clusters ? : A YING-YANG Machine Based Theory For A Classical Open Problem In Pattern Recognition", Proc. IEEE ICNN96, Vol.3, pp1546-1551.

Xu, L. (1996c),"Bayesian-Kullback YING-YANG Learning Scheme: Reviews and New Results", *Progress in Neural Information Processing, Proc. ICONIP96*, Sept. 24-27, pp59-67. Springer-verlag.

Xu, L. (1996d), "Bayesian-Kullback YING-YANG Machines for SupervisedLearning", Invited Talk, Proc. WCNN96 (Sept. 15-18), pp193-200.

Xu, L. (1996e), "A Maximum Balanced Mapping Certainty Principle for Pattern Recognition and Associative Mapping", Proc. WCNN96 (Sept. 15-18), pp946-949.

Xu, L, and Amari, S (1996), " A general independent component analysis framework based on Bayesian-Kullback Ying-Yang Learning", *Progress in Neural Information Processing, Proc. ICONIP96*, pp59-67. Springer-verlag.

Xu, L. (1995a), "YING-YANG Machine: a Bayesian-Kullback scheme for unified learnings and new results on vector quantization", Keynote talk, Proc. Intl Conf. on Neural Information Processing (ICONIP95), Oct 30 - Nov. 3, 1995, pp977-988.

Xu, L.(1995b), "YING-YANG Machine for Temporal Signals", Keynote talk, Proc of international Conference on Neural Networks and Signal Processing 1995, Vol.I, pp644-651, Nanjing, 10-13, 1995.

Xu, L. (1995c), "New Advances on The YING-YANG Machine", Invited paper, Proc. of 1995 Intl. Symposium on Artificial Neural Networks, ppIS07-12, Dec. 18-20, Hsinchu, Taiwan.

Xu, L. (1995d), " A Unified Learning Framework: Multisets Modeling Learning", Proc. WCNN95(July 17-21), Vol.I, pp35-42.

Xu, L., Jordan, M.I., & Hinton, G. E. (1995), " An Alternative Model for Mixtures of Experts", *Advances in NIPS 7*, eds., Cowan, J.D., Tesauro, G., and Alspector, J., MIT Press, 1995, pp633-640.

Xu, L., Jordan, M.I. and Hinton, G.E. (1994), "A Modified gating network for the mixtures of experts architecture", *Proc. of WCNN'94*, San Diego, Vol.2, 405-410.

Xu, L, (1993), "Least MSE Reconstruction: A Principle for Self-Organizing Nets", *Neural Networks*, Vol.6, pp627-648, 1993.