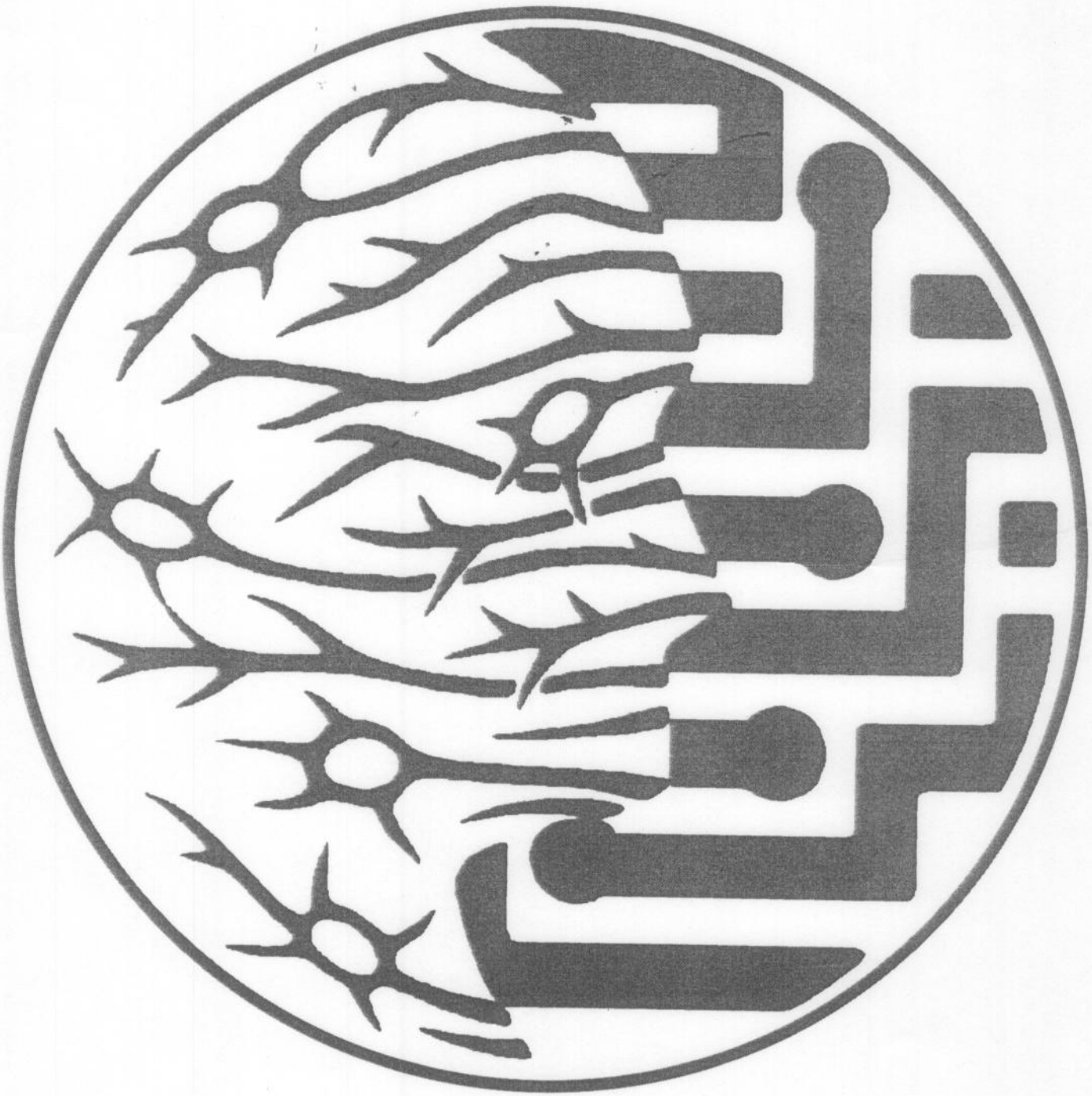
WORLD CONGRESS ON NEURAL NETWORKS SAN DIEGO, CALIFORNIA

1996 INTERNATIONAL
NEURAL NETWORK SOCIETY
ANNUAL MEETING

TOWN & COUNTRY HOTEL
SAN DIEGO, CALIFORNIA
SEPTEMBER 15-18, 1996



MATHEMATICAL FOUNDATIONS *Oral Presentations*

A Maximum Balanced Mapping Certainty Principle for Pattern Recognition and Associative Mapping[1]

Lei Xu
1. Dept. of Computer Science, The Chinese University of Hong Kong
Shatin, Hong Kong (the correspondence address)
2. Information Science Center, Peking University, Beijing, China

Abstract *A genernal principle, called Maximum Balanced Mapping Certainty (Max-BMC), is proposed for pattern recognition and associative mapping. Three unsupervised special cases have been investigated. One is equivalent to maximum mutual information or informax, with a close relation to a special case of the YING-YANG machine. One provides an extension of a recently proposed minimum uncertainty-unbalance PCA-type learning for a single neuron. The another gives a new model particularly suitable for pattern recognition purpose. Furthermore, the counterparts of the three cases for supervised learning have also been studied. In addition, the exclusive, factorial, and separable representations have been discussed with a new type of asymptotic separability for pattern recognition.*

1. Introduction

We may come to an age of searching a unified scheme for many different unsupervised and supervised learning models that have been developed for pattern recognition and associative mapping in the literature. Recently, a Bayesian-Kullback scheme, called the YING-YANG Machine, has been proposed as such an effort(Xu, 1995a&96). Its one special case reduces to the EM algorithm (Dempster et al, 1977; Hathaway, 1986; Neal & Hinton, 1993) and the closely related *Information Geometry* theory and the *em* algorithm (Amari, 1995a&b), to MDL autoencoder with a "bits-back" argument by Hinton & Zemel (1994) and its alternative equivalent form that minimizes the bits of uncoded residual errors and the unused bits in the transmission channel's capacity (Xu, 1995d), as well as to *Multisets modeling* learning (Xu, 1995e)–a unified learning framework for clustering, PCA-type learnings and self-organizing map. Its other special case reduces to maximum information preservation (Linsker, 1989; Atick & Redlich, 1990; Bell & Sejnowski, 1995). More interestingly its another special case reduces to Helmholtz machine (Dayan et al,1995; Hinton, 1995) with new understandings. The YING-YANG machine includes also maximum likelihood or least square learning. Furthermore, the YING-YANG Machine has also been extended to temporal patterns with a number of new models for signal modeling, with the Hidden Markov Model (HMM), AMAR and AR models (Xu, 1995b) and other existing models included and extended. In addition, it has also been shown in Xu(1995a&c, 1996b) that one special case of the YING-YANG machine can provide us three variants for clustering or VQ, particularly with criteria and an automatic procedure developed for solving how to select the number of clusters in clustering analysis or Gaussian mixtures — a classical problem that remains open for decades.

A number of possible new models have also been suggested by this unified YING-YANG Machine scheme (Xu, 1996a). However, there do exist some learning models for pattern recognition and associative mapping that can not be unified under this unified scheme. In this paper, we propose another general learning principle, called Maximum Balanced Mapping Certainty (Max-BMC), as a complement scheme. In Section 2, we present this principle for unsupervised learnings. We show that one special case is equivalent to maximum mutual information or informax (Linsker, 1989; Atick & Redlich, 1990; Bell & Sejnowski, 1995), with a close relation to a special case of the YING-YANG machine (Xu, 1996a). One other special case has extended a recently proposed minimum uncertainty-unbalance unsupervised learning for a single neuron, with a close relation to Nonlinear Maximum Variance (NMV) for PCA-type tasks (Xu, 1995f). Particularly, another special case gives us a new model which is more suitable for pattern recognition purpose. Section 3 further introduces the counterparts of the three cases for supervised learning. In Section 4, based on this general learning principle, we discuss the exclusive, factorial, and separable representations, with a new type of asymptotic separability suggested.

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

2. The Max-BMC Principle for Unsupervised Pattern Recognition

All the tasks of pattern recognition and associative mapping can be summaried by setting up a mapping from an input pattern space X to a class or representation space Y. Probabilistically, such a mapping can be modeled by a conditional density $P_M(y|x)$ such that this model M recognizes/classifies/maps an input x into a class or a representation y with probability $P_M(y|x)$.

To obtain this model M, we first need to specify a family $\mathcal{M}$ from which our model comes from. This family is usually of two types. For the first type, $\mathcal{M}$ is given with $P_M(y|x)$ being a regression model, e.g., $P_M(y|x) = N(f(x), \Sigma)$–a normal density with mean $f(x)$ and variance Σ and $f(x)$ implemented by a feedforward network or radial basis network. Also, $P_M(y|x)$ can be given by normalizing the outputs $[f_1(x), f_2(x), \cdots, f_k(x)]$ of a feedforward network into probabilities, e.g., by softmax $e^{f_j(x)}/\sum_{j=1}^{k} e^{f_j(x)}$.

The second type is called generative model with

$$P_M(y|x) = \frac{P_M(y)P_M(x|y)}{\int_y P_M(y)P_M(x|y)dy}, \quad or \quad P_M(y|x) = \frac{P_M(y)P_M(x|y)}{\sum_{y=1}^{k} P_M(y)P_M(x|y)}. \tag{1}$$

with $P_M(x|y)$ given by a backward network, e.g., $P_M(x|y) = N(g(y), \Sigma_y)$–a normal density with mean $g(y)$ and variance Σ_y and $g(y)$ implemented by a backward network. In the simplest case, we even can let $g(y) = m_y$ being a point, i.e., a usual normal density $P_M(x|y) = N(m_y, \Sigma_y)$.

Next, we need a principle to choose an appropriate one $M \in \mathcal{M}$, based on a given data set.

In unsupervised cases, we have only an input data set $\{x_i\}_{i=1}^{N}$ to base on.

Given a model M, $P_M(y|x)$ describe the certainty that we make a decision of classifying or mapping x into y. In the ideal case, we hope that this decision should be correct and fully confident. Since we have no supervisor here, it is difficult to check whether it is correct. However, we can maximize $P_M(y|x)$ or $f(P_M(y|x))$ to let our decision more confident, where $f(.)$ is a strictly monotonic increasing function on the interval $[0, 1]$. Considering the whole distribution of x and y, we maximize

$$E[f(P_M(y|x))] = \int_{x,y} P(x,y) f(P_M(y|x)) dxdy, \quad f(.) \text{ is strictly monotonic increasing on } [0,1]. \tag{2}$$

We further approximate $P(x,y) = P_M(y|x)P_0(x)$ with the empirical estimate $P_0(x) = \lim_{h\to 0} P_h(x) = \frac{1}{N}\sum_{i=1}^{N} \delta(x - x_i)$ for the input density. Putting this into eq.(2), we get

$$J_c = \frac{1}{N}\sum_{i=1}^{N} \int_y P_M(y|x_i) f(P_M(y|x_i)) dy, \quad or \quad J_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M(y|x_i) f(P_M(y|x_i)) \tag{3}$$

The larger is the J_c, the more confident we are on our mapping, that is, the mapping certainty is maximized. However, this maximization alone may result in a trivil solution that all the $\{x_i\}_{i=1}^{N}$ are mapped into a single y_0 with $P_M(y_0|x_i) = \delta(y - y_0)$ or $P_M(y_0|x_i) = 1$. In contrary, we hope that the mapped results can well distribute on the full range of Y space. To achieve this target, we minimize $E[f(P_M(y))]$, i.e.,

$$J_b = \int_y P_M(y) f(P_M(y)) dy, \text{ or } J_b = \sum_{y=1}^{k} P_M(y) f(P_M(y)), \quad P_M(y) = \int_x P_M(y|x)P_0(x)dx = \frac{1}{N}\sum_{i=1}^{N} P_M(y|x_i) \tag{4}$$

The minimization of J_b will let input data more balancely mapped onto the full range of space Y.

As a whole, the trade-off between $\max J_c$ and $\min J_b$ gives us a principle as follows:

$$\max_M (J_c - J_b), \qquad \text{with } J_c,\ J_b \text{ given by eq.(3) and eq.(4)}, \tag{5}$$

We call it by Maximum Balanced Mapping Certainty (Max-BMC) since it maximizes the mapping (decision) certainty and keep the largest balance on the distribution in the space Y.

In sequel, we examine three special cases for the function $f(.)$:

(1) $f(x) = x$, in this case, eq.(5) becomes

$$J_c = \frac{1}{N}\sum_{i=1}^{N} \int_y P_M^2(y|x_i) dy, \quad or \quad J_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M^2(y|x_i) d \qquad J_b = \int_y P_M^2(y) dy, \quad or \quad J_b = \sum_{y=1}^{k} P_M^2(y) \tag{6}$$

This is actually a further development of a unsupervised learning rule for a single neuron in Xu(1995f), called Minimum uncertainty-unbalance, which minimizes:

$$J_{cb} = \frac{1}{N}\sum_{i=1}^{N} p_w(y=1|x_i) p_w(y=0|x_i) - \{\frac{1}{N}\sum_{i=1}^{N} p_w(y=1|x_i)\}\{\frac{1}{N}\sum_{i=1}^{N} p_w(y=0|x_i)\}$$

where $p_w(y=1|x_i) = \frac{1+s(w^T x_i + c)}{2}$ and $s(.)$ being sigmoid function with $s(0) = 0, s(-\infty) = -1, s(\infty) = 1$. Noticing $p_w(y=0|x_i) = 1 - p_w(y=1|x_i)$, the above equation can be rewriten into $-J_{cb} = \sum_{i=1}^{N} p_w^2(y = 1|x_i) - p^2(y=1)$ with $p(y=1) = \frac{1}{N}\sum_{i=1}^{N} p_w(y=1|x_i)$. This $-J_{cb}$ is just the special case of eq.(5) and eq.(6) for only one neuron. As shown in Xu(1995f), it is closely related to the Nonlinear Maximum Variance (NMV) learning rule

$$\max E[s(w^T x_i + c) - Es(w^T x_i + c)]^2 = \max\{Es^2(w^T x_i + c) - [Es(w^T x_i + c)]^2\}$$

which is an extension of linear PCA learning to the nonlinear case.

(2) $f(x) = \ln(x)$, in this case, eq.(5) becomes

$$J_c = \frac{1}{N}\sum_{i=1}^{N}\int_y P_M(y|x_i)\ln P_M(y|x_i)dy, \quad or\ J_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M(y|x_i)\ln P_M(y|x_i),$$
$$J_b = \int_y P_M(y)\ln P_M(y)dy, \quad or\ J_b = \sum_{y=1}^{k} P_M(y)\ln(P_M(y) \tag{7}$$

That is, this Max-BMC is equivalent to maximum mutual information or informax (Linsker, 1989; Atick & Redlich, 1990; Bell & Sejnowski, 1995), and also a special case of the YING-YANG machine (Xu, 1996a).

(3) $f(x) = s(x)$ and $s(x)$ is a sigmoid function with $s(0.5) = 0.5$ and is convex on $[0.5, 1]$ but concave on $[0, 0.5]$. e.g., $s(x) = 0.5tanh(x - 0.5) + 0.5$. In this case, eq.(5) becomes

$$J_c = \frac{1}{N}\sum_{i=1}^{N}\int_y P_M(y|x_i)s(P_M(y|x_i))dy, \quad or\ J_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M(y|x_i)s(P_M(y|x_i))dy,$$
$$J_b = \int_y P_M(y)s(P_M(y))dy, \quad or\ J_b = \sum_{y=1}^{k} P_M(y)s(P_M(y)) \tag{8}$$

Comparing $\ln(x)$ with x, we find that $J_c - J_b$ in eq.(7) attempts to discount those values of probabilities near 1 (which is helpful to increase the decision's certainty) and to over-count those near 0 (which is actually not desirable). In contrast, $s(x)$ can also discount those values of probabilities near 0. As a whole, it pays more attention on deciding those data which are difficult to make a decision (which is surely desirable to increase decision certainty). Thus, eq.(8) should be more suitable for those tasks with binary outputs, such clustering or classification. This is a new model that deserves further theoretical and experimental explorations.

3. The Max-BMC Principle for Supervised Pattern Recognition

In supervised cases, we have a data set $\{x_i, y_i\}_{i=1}^N$ to base. Now the distribution of y is controlled by the data $\{y_i\}_{i=1}^N$, we no longer need the term J_b for this purpose. Thus, we can maximize only the following simplified J_c to choose $M \in \mathcal{M}$:

$$J_c = \frac{1}{N}\sum_{i=1}^{N} P_M(y_i|x_i)f(P_M(y_i|x_i))$$
$$J_c = \frac{1}{N}\sum_{i=1}^{N} P_M^2(y_i|x_i),\ for\ f(x) = x, \quad J_c = \frac{1}{N}\sum_{i=1}^{N} P_M(y_i|x_i)\ln P_M(y_i|x_i),\ for\ f(x) = \ln(x)$$
$$J_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M(y|x_i)s(P_M(y_i|x_i)), \quad for\ f(x) = s(x). \tag{9}$$

The case for $f(x) = x$ is the same as a new supervised learning model obtained from a special case of the YING-YANG machine (Xu, 1996a). All the others in eq.(9) are also new models that deserve further theoretical and experimental explorations.

4. Exclusive, Factorial, and Separable Representations

For J_c in eq.(3), in the case that $f(.)$ is convex, we have

$$J_c \le f(\frac{1}{N}\sum_{i=1}^{N}\int_y P_M^2(y|x_i)dy) \le f(1), \quad or\ J_c \le f(\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M^2(y|x_i)) \le f(1) \tag{10}$$

That is, it is up-bounded by $f(1)$. For $f(x) = x$, $f(1) = 1$ is achieved when $\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M^2(y|x_i)) = 1$, which holds only when $P_M(y|x_i) = 1$ for a single y_0 and $P_M(y|x_i) = 0$ for all the other y's. In other words, each x_i is classified into one class with probability 1, or each x_i is exclusively represented by a single y. For $f(x) = \ln x$, similarly $f(1) = 0$ is also is achieved when each x_i is exclusively represented by a single y. So, we see that $\max J_c$ is to ensure an *exclusive* representation for each input x_i.

On the other hand, $\max(-J_b)$ is actually the entropy of $P_M(y)$. When $y = [y_1, \cdots, y_k]$ with each binary $y_j = 1$ or $y_j = 0$, $\max(-J_b)$ is achieved when $P_M(y) = \prod_{j=1}^{k} P_M(y_j)$, i.e., whether y_i taking 1 or 0 is independent of what values $y_j, j \neq i$ takes. This is called *factorial* representation, which is different from the above *exclusive* representation by which there is only one $y_j = 1$ with all $y_i = 0, i \neq j$.

So we see that the Max-BMC Principle is a combination of *exclusive* representation on $P(y|x)$ and *factorial* representation on $P(y)$.

In sequel, we further study the condition that $\frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M^2(y|x_i) = 1$. When N is finite, it holds only when $P_M(y|x_i)$ is exclusive for each x_i. When $N \to \infty$, this requirement can be relaxed into

$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M^2(y|x_i) = 1, \quad for\ f(x) = x \tag{11}$$
$$\lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N}\sum_{y=1}^{k} P_M(y|x_i)\ln P_M(y|x_i) = 0, \quad for\ f(x) = \ln(x) \tag{12}$$

which may not imply that $P_M(y|x_i)$ is exclusive for each x_i. In this case, we call those $P_M(y|x_i)$ *asymptotic exclusive*. Clearly, *exclusive* implies *asymptotic exclusive*, but its inverse is not true.

Moreover, from $\sum_{y=1}^{k} P_M(y|x_i) = 1$, we have $1 = \sum_{y=1}^{k} P_M^2(y|x_i) + \sum_{p=1}^{k} \sum_{p=1,p\neq q}^{k} P_M(p|x_i)P_M(q|x_i)$. Thus, the condition eq.(11) for $f(x) = x$ is equivalent to

$$\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \sum_{p=1}^{k} \sum_{p=1,p\neq q}^{k} P_M(p|x_i)P_M(q|x_i) = 0 \tag{13}$$

which holds when $\lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} P_M(p|x_i)P_M(q|x_i) = 0$ for any pair $p \neq q$. In other words, asymptotically there is no overlap between classes p, q. That is, the data of classes p, q are asymptotically separable. Here, the meaning of "separable" is more general and relaxed than the conventional deterministic "linear separable" "nonlinear separable" in pattern recognition literature, which does not permit $P_M(p|x_i)P_M(q|x_i) = 0$ on any pair $p \neq q$ even for a single point x_i and thus is too idealistic in practice. In the special case that $\mathcal{M}$ is given by the generative model eq.(1) with Gaussian density $P_M(x|y) = N(m_y, \Sigma_y)$, the condition eq.(13) is exactly the Cond-EMG— a necessary and sufficient condition for the correct convergence by the EM algorithm on Gaussian mixture (Ma & Xu, 1996).

With the above regards, we can take the conditions eq.(11) or eq.(12) as the definition of *asymptotic separable data or distributions*, which are more general and practical for pattern recognition purpose. Moreover, from eq.(10) we know that making eq.(12) satisfied only results in the maximized upbound $\ln 1 = 0$ for J_c with $f(x) = \ln x$, which does not imply the maximization of this J_c itself.

5. Conclusions

The Maximum Balanced Mapping Certainty (Max-BMC) can serve as a general principle for both unsupervised and supervised pattern recognition. It unifies some existing models and also provides several new models. It is a combination of *exclusive* representation on $P(y|x)$ and *factorial* representation on $P(x)$. Based on it, we can also obtain more general and practical definitions on separability for pattern recognition.

References

Amari, S(1995a), " Information geometry of the EM and em algorithms for neural networks", *Neural Networks 8.*

Amari, S(1995b), *Neural Computation 7*, pp13-18.

Atick, J.J. & Redlich, A.N. (1990), *Neural Computation* Vol.2, No.3, pp308-320.

Bell A. J. & Sejnowski, T. J.(1995), *Neural Computation* Vol.7, No.6, 1129-1159.

Byrne, W. (1992), *IEEE Trans. Neural Networks 3*, pp612-620.

Csiszar, I., (1975), *Annals of Probability 3*, pp146-158.

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.

Hathaway, R.J.(1986), *Statistics & Probability Letters 4*, pp53-56.

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

Hinton, G. E. & Zemel, R.S. (1994), *Advances in NIPS 6*, pp3-10.

Linsker, R. (1989), *Advances in NIPS 1*, pp186-194.

Neal, R. N.& Hinton, G. E(1993), *A new view of the EM algorithm that justifies incremental and other variants*, preprint.

Ma, J.W., & Xu, L., (1996), " A necessary and sufficient condition for the correct convergence by the EM algorithm on Gaussian mixture ", to be submitted.

Xu, L. (1996a), "A Unified Learning Scheme: Bayesian-Kullback YING-YANG Machine", to appear on *Advances in Neural Information Processing Systems 8*, David S. Touretzky, Michael C. Mozer and Michael E. Hasselmo, eds, MIT Press: Cambridge, MA.

Xu, L. (1996b), "How Many Clusters ? : A YING-YANG Machine Based Theory For A Classical Open Problem In Pattern Recognition", to appear on *Proc. IEEE ICNN96.*

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), 1995, pp977-988.

Xu, L.(1995b), "YING-YANG Machine for Temporal Signals", Keynote talk, Proc IEEE intl Conf. Neural Networks & Signal Processing, 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, Taiwan.

Xu, L. (1995d), "Cluster Number Selection, Adaptive EM Algorithms and Competitive Learnings", Invited paper, Proc. Intl Conf. on Neural Information Processing (ICONIP95), Oct 30 - Nov. 3, 1995, Vol.II, pp1499-1502.

Xu, L. (1995e), Invited paper, Proc. WCNN95, Vol.I, pp35-42. Invited paper, Proc. IEEE ICNN 1994, ppI315-320.

Xu, L. (1995f), " Advances on Three Streams of PCA Studies", Invited paper, Proc. Intl Conf. on Neural Information Processing (ICONIP95), Oct 30 - Nov. 3, 1995, Vol.I, pp480-483.