J.M. Zurada et al. (Eds.): Computational Intelligence: Research Frontier,
LNCS State-of-the Art Survey Volume: WCCI 2008 Plenary/Invited Lectures, LNCS 5050, pp. 48–78, 2008.

Bayesian Ying Yang System, Best Harmony Learning, and Gaussian Manifold based Family

Lei Xu

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, email: lxu@cse.cuhk.edu.hk

Abstract. Two intelligent abilities and three inverse problems are re-elaborated from a probability theory based two pathway perspective, with challenges of statistical learning and efforts towards the challenges overviewed. Then, a detailed introduction is provided on the Bayesian Ying-Yang (BYY) harmony learning. Proposed firstly in (Xu,1995) and systematically developed in the past decade, this approach consists of a two pathway featured BYY system as a general framework for unifying a number of typical learning models, and a best Ying-Yang harmony principle as a general theory for parameter learning and model selection. The BYY harmony learning leads to not only a criterion that outperforms typical model selection criteria in a two-phase implementation, but also model selection made automatically during parameter learning for several typical learning tasks, with computing cost saved significantly. In addition to introducing the fundamentals, several typical learning approaches are also systematically compared and re-elaborated from the BYY harmony learning perspective. Moreover, a further brief is made on the features and applications of a particular family called Gaussian manifold based BYY systems.

1 Introduction

1.1 Two intelligent abilities and three inverse problems

An intelligent system, which could be an individual or a collection of men, animals, robots, agents, and other intelligent bodies, survives in its world with needs of two types of intelligent abilities. As illustrated by Fig.1, implemented by a top-down or outbound pathway, Type-I consists of abilities of discovering the knowledge about its world, including not only understanding ability to explain its world but also motoring ability to track the changes in its world. The knowledge is obtained either from pieces of uncertain evidences (or called samples) about the world or from certain existing authorized sources (e.g., textbooks) that were obtained from samples in past. Therefore, Type-I abilities are actually obtained via processes that we usually call *learning*, during which an intelligent system gradually senses its world from samples and modifies itself to adapt the world. This learning task aims at common features or regularities among an ensemble of uncertain evidences (or called samples) from the world.

J.M. Zurada et al. (Eds.): WCCI 2008 Plenary/Invited Lectures, LNCS 5050, pp. 48–78, 2008.
@ Springer-Verlag Berlin Heidelberg 2008

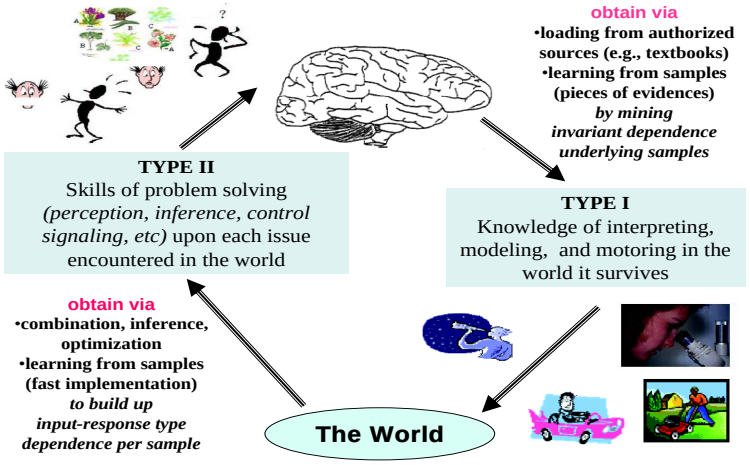

Fig. 1. Two types of intelligent ability and how to get the abilities

On the other hand, implemented by a bottom-up or inbound pathway, Type-II consists of problem solving skills, ranging from perceiving events that are encountered to producing signals that activate the outbound pathway. These skills can be roughly classified into two categories. One is made via evidence combination, inference, optimization, based on a priori knowledge of Type-I. The other is developing a fast implementing device (or called problem solver) for those often encountered events that usually need a rapid response. Specifically, the problem solver is developed via learning from samples either based on the existing Type-I knowledge or in help of a teacher who teaches a desired response as each sample comes (e.g., in supervised pattern recognition, function approximation, control system, ..., etc). This learning task is featured by aiming at the dependence of input-response type per one or several samples encountered.

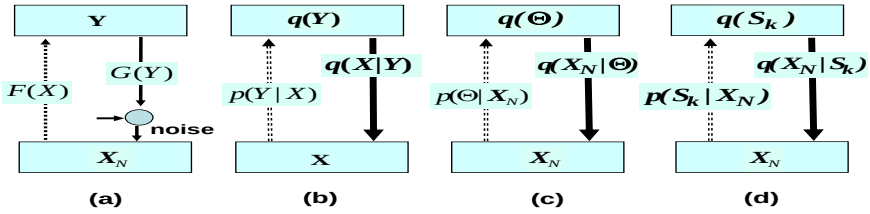

Fig. 2. Three levels of inverse problems

Insights on how two types of of intelligent abilities can be further observed from three levels of inverse problems, illustrated in Fig.2. Provided with an observation x that can be regarded as either generated from an inner representation y or a consequence from a cause y via a given mapping $G : y \to x$, the Type-II ability makes an inverse inference $x \to y$, as shown in Fig.2(a). When $G : y \to x$ is one-to-one, its inverse one-to-one mapping $F : x \to y$ is analytically solvable.

Generally, it is not so simple due to uncertainties. One type of uncertainties is incurred externally by observation noises, which can be described by a distribution $q(x|y,\theta_{x|y})$ for a probabilistic mapping $y \to x$. The other type origins internally from a mapping $G : y \to x$ of many-to-one or infinite many to one, which can be considered by $q(x|y,\theta_{x|y})$ plus a distribution $q(y|\theta_y)$ for every reasonable cause or inner representation y, as shown in Fig.2(b). Actually, $q(x|y,\theta_{x|y})$ and $q(y|\theta_y)$ jointly act as the knowledge of Type I, based on which a Type-II ability is obtained via combination, inference, optimization as illustrated at the left-bottom in Fig.1. Specifically, there are four typical ways to handle it, as listed in the 1st column of Tab. 1.

The first choice is Bayesian inference (BI) that provides a distribution $p(y|x)$ for a probabilistic inverse map $x \to y$ via combining evidences from $q(x|y,\theta_{x|y})$ and $q(y|\theta_y)$ in a normalized way, which involves an integral with a computational complexity that is usually too high to be practical. The difficulty is tackled by seeking a most probable mapping $x \to y$ in a sense of the largest probability $p(y|x)$, called the maximum Bayes (MB) or MAximum Posteriori (MAP). It further degenerates into $y^* = arg\max_y q(x|y,\theta_{x|y})$ when there is no knowledge about $q(y|\theta_y)$. In some cases, making maximization may also be computationally expensive. Instead, the last choice is to Learn a Parametric Distribution (LPD) $p(y|x,\theta_{y|x})$ by which an inverse mapping $x \to y$ can be fast implemented. To get this $p(y|x,\theta_{y|x})$, we need its structure pre-specified and then learn the parameter set $\theta_{y|x}$ from samples either based on $q(x|y,\theta_{x|y})$ and $q(y|\theta_y)$ or in help of a teacher who teaches a desired response to each sample. Actually, this LPD is a special case of the following second type of inverse problems.

	(a) Inverse Inference on y	(b) Parameter Learning	(c) Model Selection
BI	$p(y\|x)=\frac{q(x\|y,\theta_{x\|y})q(y\|\theta_y)}{q(x\|\theta)}$ $q(x\|\theta)=\int q(x\|y,\theta_{x\|y})q(y\|\theta_y)dy$	$p(\theta\|X_N)=\frac{q(X_N\|\theta)q(\theta)}{q(X_N\|\mathbb{S})}$ $q(X_N\|\mathbb{S})=\int q(X_N\|\theta)q(\theta)d\theta$	$p(\mathbf{k}\|X_N)=\frac{q(X_N\|\mathbb{S}_k)q(\mathbf{k})}{q(X_N\|\aleph)}$ $q(X_N\|\aleph)=\sum_{\mathbf{k}} q(X_N\|\mathbb{S}_k)q(\mathbf{k})$
MB	$\max_y[q(x\|y,\theta_{x\|y})q(y\|\theta_y)]$	$\max_\theta[q(X_N\|\theta)q(\theta)]$	$\max_{\mathbf{k}}[q(X_N\|\mathbb{S}_k)q(\mathbf{k})]$
ML	$\max_y q(x\|y,\theta_{x\|y})$	$\max_\theta q(X_N\|\theta)$	$\max_{\mathbf{k}} q(X_N\|\mathbb{S}_k)$
LPD	$p(y\|x,\theta_{y\|x})$	$p(\theta\|X_N)$	$p(\mathbf{k}\|X_N)$
BI for Bayesian Inference, MB for Maximum Bayes or called Maximum Posteriori (MAP) ML for Maximum Likelihood or Marginal Likelihood, LPD for Learned Parametric Distribution			

Table 1 Typical methods for three levels of inverse problems.

The second level of inverse problems considers the situations that $q(x|y,\theta_{x|y})$ and $q(y|\theta_y)$ are unknown but provided with their parametric structures. As illustrated in Fig.2(c), the scenario becomes that we have a set of samples $\mathcal{X}_N = \{x_t\}_{t=1}^N$ from a map $\Theta \to \mathcal{X}_N$, and the task is getting an inverse mapping $\mathcal{X}_N \to \Theta$, usually referred by the term *estimation* or *parameter learning* for Θ. This Θ consists of $\theta_{x|y}, \theta_y$, as well as $\theta_{y|x}$ (if the above LPD is

considered together). There could be different directions to pursuit an inverse mapping $\mathcal{X}_N \rightarrow \Theta$. One most widely studied one is that similar to Fig.2(b), with uncertainties considered by two distributions $q(\mathcal{X}_N|\Theta)$ and $q(\Theta)$. Usually, $q(\mathcal{X}_N|\Theta)$ is described by $q(\mathcal{X}_N|\Theta) = \int q(\mathcal{X}_N|\mathcal{Y}_N, \theta_{x|y}) q(\mathcal{Y}_N|\theta_y) d\mathcal{Y}_N$. When the samples in $\mathcal{X}_N$ are independently and identically distributed (i.i.d.), we have $q(\mathcal{X}_N|\Theta) = \prod_{t=1}^{N} q(x_t|\Theta)$ with $q(x_t|\Theta)$ given by the choice BI of the 1st column in Tab. 1.

Based on $q(\mathcal{X}_N|\Theta)$ and $q(\Theta)$, again there are four ways for getting an inverse mapping $\mathcal{X}_N \rightarrow \Theta$, as shown in the 2nd column of Table 1. The simplest and most widely studied one is the maximum likelihood (ML) learning $\max_{\Theta} q(\mathcal{X}_N|\Theta)$. With a priori distribution $q(\Theta)$ in consideration, we are lead to the choice MB of the 2nd column in Table 1, i.e., $\max_{\Theta}[q(\mathcal{X}_N|\Theta)q(\Theta)]$, on which extensive studies have been made under different names [25, 32, 42], and are collectively referred in term of Bayesian school. The challenge is how to get an appropriate $q(\Theta)$, which needs a priori knowledge that we may not have. Related efforts also include those made under *Tikhonov regularization* [40, 26] or regularization approaches. Conceptually, we may also consider the BI choice in the 2nd column of Table 1 for a probabilistic inverse mapping by a distribution $p(\Theta|\mathcal{X}_N)$, while it encounters an integral over Θ.

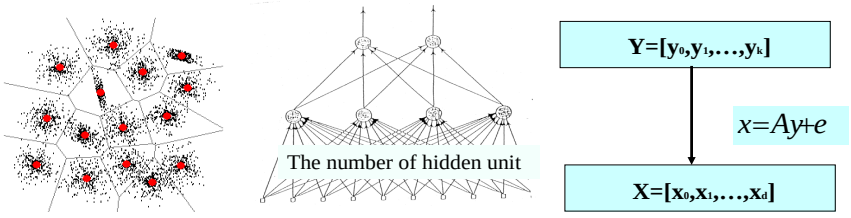

Fig. 3. A combination of a series of individual simple structures

Being too difficult to compute except some special cases, this integral over Θ is encountered not just as above but also in the 3rd column of Table 1. An alternative is using a particularly designed parametric structure in place of $p(\Theta|\mathcal{X}_N)$, i.e., the choice LPD in the 2nd column of Table 1. Moreover, even in implementing the ML learning, we have to handle either a summation or a numerical integral over y for getting $q(x|\Theta)$ (see the choice BI of the 1st column in Table 1), which also involves a huge computing cost except special cases. Instead, the choice LPD in the 1st column of Table 1 is considered via a particularly designed parametric structure $p(y|x, \theta_{y|x})$. Studies on learning either or both of $p(y|x, \theta_{y|x})$ and $p(\Theta|\mathcal{X}_N)$ jointly with the *parameter learning* for Θ have been made in the Helmholtz free energy based learning [15, 11], BYY Kullback learning [64], and BYY harmony learning [64, 47]. Detailed discussions are referred to Sec.3.2.

Until now, we assume that the parametric structures of $q(x|y, \theta_{x|y})$ and $q(y|\theta_y)$, as well as of $p(y|x, \theta_{y|x})$ are provided in advance. In fact, we do not

know how to pre-specify these structures. Usually, we consider a family of infinite many structures $\{S_{\mathbf{k}}(\Theta_{\mathbf{k}})\}$ via combining a set of individual simple structures (or simply called units) via a simple combination scheme, as shown in Fig.3. Every unit can be simply one point, one dimension in a linear space, or one simple computing unit. The types of the basic units and the combination scheme jointly act as a seed or meta structure $\aleph$ that grows into a family $\{S_{\mathbf{k}}(\Theta_{\mathbf{k}})\}$ with each $S_{\mathbf{k}}$ sharing a same configuration but in different scales, each of which is labeled by a scale parameter $\mathbf{k}$ in term of one integer or a set of integers. That is, each specific $\mathbf{k}$ corresponds to one candidate model with a specific complexity. We can enumerate each candidate via enumerating [1] $\mathbf{k}$.

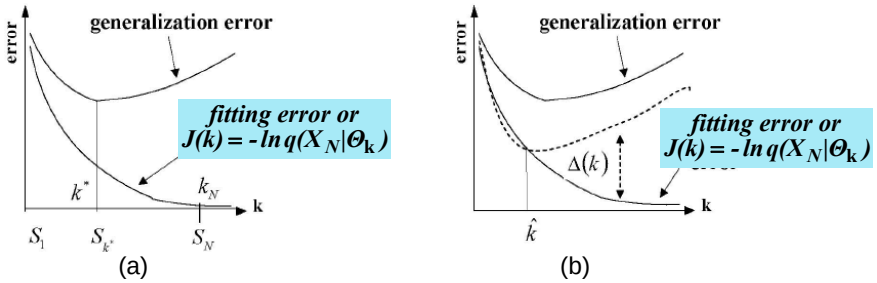

Fig. 4. Model selection : fitting performance vs generalization performance

As shown in Fig.2(d), the third level of inverse problems considers selecting an appropriate $\mathbf{k}^*$ based on $\mathcal{X}_N = \{x_t\}_{t=1}^N$ only, usually referred as *model selection*. We can not simply use the best likelihood value as a measure to guide this selection. As illustrated in Fig.4(a), $J(\mathbf{k}) = -\max_{\Theta} lnq(\mathcal{X}_N|\Theta)$ will keep decreasing as k increases and reaches zero at a value k_N that is usually much larger than the appropriate one, as long as the size N is finite. Though a $S_{\mathbf{k}}(\Theta_{\mathbf{k}})$ with $\mathbf{k}^* \prec \mathbf{k}$ can get a low value $J(\mathbf{k})$ and thus $\mathcal{X}_N$ got well described, it has a poor generalization performance (i.e., performing poorly on new samples with the same regularity underlying $\mathcal{X}_N$). This is also called *over-fitting* problem.

Until now, we assume that the parametric structures of $q(x|y, \theta_{x|y})$ and $q(y|\theta_y)$, as well as of $p(y|x, \theta_{y|x})$ are provided in advance. Actually, we do not know how to pre-specify these structures. Usually, we consider a family of infinite many structures $\{S_{\mathbf{k}}(\Theta_{\mathbf{k}})\}$ via combining a set of individual simple structures (or simply called units) via a simple combination scheme, as shown in Fig.3. Every unit can be simply one point, one dimension in a linear space, or one simple computing unit. The types of the basic units and the combination scheme jointly act as a seed or meta structure $\aleph$ that grows into a family $\{S_{\mathbf{k}}(\Theta_{\mathbf{k}})\}$ with each $S_{\mathbf{k}}$ in a same configuration but in different scales, each of which is labeled by a scale parameter $\mathbf{k}$ in term of one integer or a set of integers. That is, each

[1] We say that $\mathbf{k}_1$ proceeds $\mathbf{k}_2$ or $\mathbf{k}_1 \prec \mathbf{k}_2$ if $S_{\mathbf{k}_1}$ is a part (or called a substructure) of $S_{\mathbf{k}_2}$. When $\mathbf{k}$ consists of only one integer, $\mathbf{k}_1 \prec \mathbf{k}_2$ becomes simply $\mathbf{k}_1 < \mathbf{k}_2$.

specific $\mathbf{k}$ corresponds to one candidate model with a specific complexity. We can enumerate each candidate via enumerating [2] $\mathbf{k}$.

1.2 Efforts towards challenges

In the past 30 or 40 years, several learning principles or theories have been proposed and studied for an appropriate $J(\mathbf{k})$, roughly along three directions.

Those measures summarized in Table 1 are featured by the most probable principle based on probability theory. The efforts of the first direction can be summarized under this principle. As discussed above, the ML choice of the 2nd column in Table 1 can not serve as $J(\mathbf{k})$. Studies on the BI choice of the 2nd column, i.e., $J(\mathbf{k}) = -\max_{\Theta}[q(\mathcal{X}_N|\Theta)q(\Theta)]$, have been made under the name of *minimum message length* (MML)[42]. It can provide an improved performance over $J(\mathbf{k}) = -\max_{\Theta} q(\mathcal{X}_N|\Theta)$ but is sensitive to whether an appropriate $q(\Theta)$ is pre-specified, which is difficult. Studies on the BI choice of the 3rd column in Table 1 have also been conducted widely in the literature. Usually assuming that $q(\mathbf{k})$ is equal for every $\mathbf{k}$, we are lead to the ML (marginal likelihood) choice of the 3rd column, i.e., $J(\mathbf{k}) = -\ln q(\mathcal{X}_N|S_k)$, by which the effect of $q(\Theta)$ has been integrated out. However, the integral over Θ is difficult to compute and thus is approximately tackled by turning it into the following format:

$$J(k) = -\max_{\Theta} \ln q(\mathcal{X}_N|\Theta) + \Delta(\mathbf{k}), \tag{1}$$

where the term $\Delta(\mathbf{k})$ is resulted from a rough approximation such that it is computable. Differences on $q(\Theta)$ and on methods for approximating the integral result in different specific forms. Typical efforts include those under the names of *Bayesian Information Criterion* [34, 23], Bayes Factors [21], the evidence or the marginal likelihood [22], etc. The *Akaike Information Criterion (AIC)* can also be obtained as a special case though it was orginally derived from a different perspective [1, 2].

The second direction follows the well known principle of Ockham Razor, i.e., seeking a most economic model that represents $\mathcal{X}_N$. It is implemented via mimizing a two part coding length. One is for encoding the residuals or errors incurred by the model in representing $\mathcal{X}_N$, which actually corresponds to the first term in eq.(1). The other is for encoding the model itself, which actually corresponds to the second term in eq.(1). Different specific forms maybe obtained due to differences on what measure is used for the length and on how to evaluate the measure, which is usually difficult, especially for the second part coding. Studies have been made under the names of *minimum message length* (MML)[42], *minimum description length* (MDL) [29], best information transfer, etc. After this or that type of approximation, the resulted criteria turn out closely related to or even same as those obtained along the above first direction.

Another direction is towards estimating the generalization performance directly. One typical approach is called *cross-validation* (CV). $\mathcal{X}_N$ is randomly

[2] We say that $\mathbf{k}_1$ proceeds $\mathbf{k}_2$ or $\mathbf{k}_1 \prec \mathbf{k}_2$ if $S_{\mathbf{k}_1}$ is a part (or called a substructure) of $S_{\mathbf{k}_2}$. When $\mathbf{k}$ consists of only one integer, $\mathbf{k}_1 \prec \mathbf{k}_2$ becomes simply $\mathbf{k}_1 < \mathbf{k}_2$.

and evenly divided into $D_i, i = 1, \cdots, m$ parts, each D_i is used to measure the performance of $S_{\mathbf{k}}$ with its $\Theta_{\mathbf{k}}$ determined from the rest samples in $\mathcal{X}_N$ after taking D_i away. Then we use the average performance measures of m times as an estimation of $J(\mathbf{k})$ [39, 30]. One other approach is using the VC dimension based learning theory [41] to estimate a bound of generalization performance via theoretical analysis. A rough bound can be obtained for some special cases, e.g., a Gaussian mixture [44]. Generally, such a bound is difficult to get because it is very difficult to estimate the VC dimension of a learning model.

Even with a $J(\mathbf{k})$ available, evaluating its optimal values involves a discrete optimization nested with a series of implementations of parameter learning for a best $\Theta^*_{\mathbf{k}}$ at each $\mathbf{k}$. The task usually incurs a huge computing cost, while many practical applications demand that learning is made adaptively upon each sample comes. Moreover, the parameter learning performance deteriorates rapidly as $\mathbf{k}$ increases, which makes the value of $J(\mathbf{k})$ evaluated unreliably. Efforts have been made on tackling this challenge along two directions. One is featured by incremental algorithms that attempts to incorporate as much as possible what learned as $\mathbf{k}$ increases step by step, focusing on learning newly added parameters. Such an incremental implementation can save computing costs in certain extent. However, parameter learning has to be made by enumerating the values of $\mathbf{k}$, and computing costs are still very high. Also, it usually leads to suboptimal performance because not only those newly added parameters but also the old parameter set $\Theta_{\mathbf{k}}$ have to be re-learned. Another type of efforts has been made on a widely encountered category of structures that consists of individual substructures, e.g., a Gaussian mixture that consists of several Gaussian components. A local error criterion is used to check whether a new sample x belongs to each substructure. If x is regarded as not belonging to anyone of substructures, an additional substructure is added to accommodate this new x. This incremental implementation is much faster. However, the local evaluating nature makes it very easy to be trapped into a poor performance, except for some special cases that $\mathcal{X}_N = \{x_t\}_{t=1}^N$ come from substructures that are well separated.

The other direction consists of learning algorithms that start with $\mathbf{k}$ at a large value and decrease $\mathbf{k}$ step by step, with extra parameters discarded and the remaining parameter updated. These algorithms are further classified into two types. One is featured by decreasing $\mathbf{k}$ step by step, based on evaluating the value of $J(\mathbf{k})$ at each $\mathbf{k}$. The other is called automatic model selection, with extra structural parts removed automatically during parameter learning. One early effort is Rival Penalized Competitive Learning (RPCL) [65] for a structure that consists of k individual substructures. With k initially given a value larger enough, a coming sample x is allocated to one of the k substructures via competition, and the winner adapts this sample by a little bit, while the rival (i.e., the second winner) is de-learned a little bit to reduce a duplicated allocation. This rival penalized mechanism will discard those extra substructures, making model selection automatically during learning. Various extensions have been made in the past one decade and half. Readers are referred to a recent encyclopedia paper [48].

1.3 Two-pathway approaches and the scope of this paper

RPCL learning was heuristically proposed in lack of theoretical guide. Proposed firstly in [64] and systematically developed in the past decade [47, 49], the Bayesian Ying-Yang (BYY) harmony learning acts as a general statistical theory that guides various learning tasks with model selection achieved automatically during parameter learning, which is featured by using a Bayesian Ying-Yang (BYY) system to model an intelligent system and three level of inverse problems shown in Fig.1 and Fig.2.

The two-pathway idea has been adopted in the literature of modelling a perception system for decades. One early example is the adaptive resonance theory developed in the 1970s [14], featured by a resonance between bottom-up input and top-down expectation in help of a mechanism motivated from a cognitive science view. Efforts have been further made on multi-layer net featured two-pathway approaches, e.g., under the least mean square error based auto-association [6], the LMSER self-organization [66]. However, these early studies were neither motivated nor targeted at a probability theory based perspective as shown in Fig.2.

In addition to those approaches discussed in Table 1, studies on a probabilistic two-path way perspective include the BYY learning, the Helmholtz free energy based learning or *Helmholtz machine* [15, 11], variational approximation methods [20, 19]. Motivated differently, these approaches share certain common features and also have different properties. Firstly proposed in 1995 [64, 56, 50, 51, 47] and developed in the past decade, BYY learning not only acts as a general framework for a unified perspective on these approaches as well as the approaches in Table 1, but also provides a new theory for model selection on a finite size of samples, both on deriving a criterion that outperforms typical model selection criteria in a two-phase implementation, and on developing learning algorithms for several typical learning tasks with an appropriate model scale obtained automatically during parameter learning while with computing cost saved significantly.

In the rest of this paper, Section 2 introduces the fundamentals of Bayesian Ying Yang system and best harmony learning theory, the implementable structures for Yang machine and the distributed log-quadratic inner structures for Ying machine. In Section 3, relations and differences of a number of existing typical learning approaches are rather systematically compared and re-elaborated from the perspective of BYY learning under the principles of best harmony versus best matching. Finally, a further introduction is made on a particular family of BYY systems featured with Gaussian manifolds as components.

2 Bayesian Ying-Yang Learning

2.1 Bayesian Ying-Yang System and Best Harmony Learning

As shown in Fig.5, a unified scenario of Fig.2 is considered by regarding that the observation set $\mathbf{X} = \{x\}$ are generated via a top-down path from its inner representation $\mathbf{R} = \{\mathbf{Y}, \Theta\}$. Given a system architecture, the parameter set Θ

collectively represents the underlying structure of $\mathbf{X}$, while one element $y \in \mathbf{Y}$ is the corresponding inner representation of one element $x \in \mathbf{X}$. A mapping $\mathbf{R} \to \mathbf{X}$ and an inverse mapping $\mathbf{X} \to \mathbf{R}$ are jointly considered via the joint distribution of $\mathbf{X}$ and $\mathbf{R}$ in two types of Bayesian decomposition shown at the right-bottom of Fig.5. In a compliment to the famous ancient Ying-Yang philosophy, the decomposition of $p(\mathbf{X}, \mathbf{R})$ coincides the Yang concept with a visible domain $p(\mathbf{X})$ for a Yang space and a forward pathway by $p(\mathbf{R}|\mathbf{X})$ as a Yang pathway. Thus, $p(\mathbf{X}, \mathbf{R})$ is called Yang machine. Similarly, $q(\mathbf{X}, \mathbf{R})$ is called Ying machine with an invisible domain $q(\mathbf{R})$ for a Ying space and a backward pathway by $q(\mathbf{X}|\mathbf{R})$ as a Ying pathway. Such a Ying-Yang pair is called *Bayesian Ying-Yang (BYY) system*.

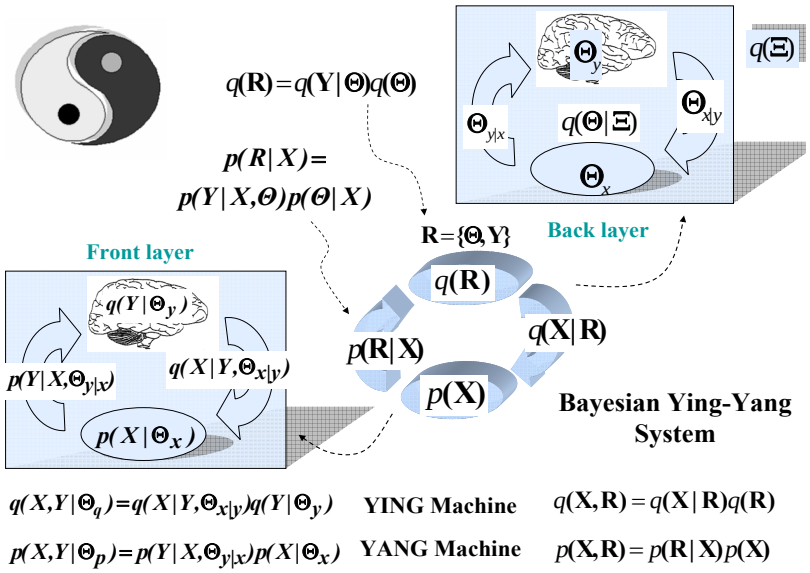

Fig. 5. Bayesian Ying-Yang System

As shown in Fig.5, the system is further divided into two layers. The front layer is actually the one shown in Fig.2(b), with a parametric Ying-Yang pair at the left-bottom of Fig.5, which consists of four components with each associated with a subset of parameters $\Theta = \{\Theta_p, \Theta_q\}$, where $\Theta_p = \{\theta_{y|x}, \theta_x\}$ and $\Theta_q = \{\theta_y, \theta_{x|y}\}$. This Θ is accommodated on the back layer with a priori structure $q(\Theta|\Xi_q)$ to back up the front layer, the back layer may be modulated by a meta knowledge from a meta layer $q(\Xi)$. Correspondingly, an inference on Θ is given by $p(\Theta|\mathbf{X}, \Xi_p)$ that integrates information from both the front layer and the meta layer. Putting together, we have

$$\begin{aligned} q(\mathbf{X}, \mathbf{R}) &= q(\mathbf{X}|\mathbf{Y}, \theta_{x|y}) q(\mathbf{Y}|\theta_y) q(\Theta|\Xi_q), \\ p(\mathbf{X}, \mathbf{R}) &= p(\Theta|\mathbf{X}, \Xi_p) p(\mathbf{Y}|\mathbf{X}, \theta_{y|x}) p(\mathbf{X}|\theta_x). \end{aligned} \tag{2}$$

The external input is only a set of samples $\mathcal{X}_N = \{x_t\}_{t=1}^N$ of $\mathbf{X} = \{x\}$, based on which we form an estimate of $p(\mathbf{X}|\theta_x)$ either directly or with a unknown sclar parameter $\theta_x = h$, as shown in Tab.2. Based on this very limited knowledge, the goal of building up the entire system is too ambitious to pursuit. We need to further specify certain structures of $p(\mathbf{X}, \mathbf{R})$ and $q(\mathbf{X}, \mathbf{R})$. Summarized in Tab 2 are typical scenarios of both $p(\mathbf{X}, \mathbf{R})$ and $q(\mathbf{X}, \mathbf{R})$, and further details will be introduced in the subsequent two subsections.

Similar to the discussions made at the end of Sec.1.1, the Ying Yang system is also featured by a given meta structure $\aleph$ that grows into a family $\{S_{\mathbf{k}}(\Theta_{\mathbf{k}})\}$ with each $S_{\mathbf{k}}$ sharing a same configuration but in different scales of $\mathbf{k}$. The meta structure $\aleph$ consists $\aleph_q, \aleph_p$ for the Ying machine and the Yang machine respectively, from which we get the structures of $q(\mathbf{X}, \mathbf{Y}|\Theta_q)$ and $p(\mathbf{X}, \mathbf{Y}|\Theta_p)$ in different scales. Though it is difficult to precisely define, the scale $\mathbf{k}$ of an entire system is featured by the scale or complexity for representing R, which is roughly regarded as consisting of the scale $\mathbf{k}_Y$ for representing Y and the number n_f of free parameters in Θ.

As shown in Tab.2, different structures of the Ying machine $q(\mathbf{X}, \mathbf{Y}|\Theta_q)$ are considered to accommodate the world knowledge and different types of dependences encountered in various learning tasks. First, an expression format is needed for each inner representation $\mathbf{Y}$. It has four typical choices as shown in Tab.2. The general case is the last one, i.e., $\mathbf{Y} = \{\mathbf{Y}_v, \mathbf{L}\}$ with $\mathbf{Y}_v = \{y_v\}$, $\mathbf{L} = \{\ell\}$. Each ℓ takes a finite number of integers to denote one of several labels for tasks of pattern classification, choice decision, and clustering analyses, etc, while each y_v is a vector that acts as an inner coding or cause for observations. Moreover, $q(\mathbf{Y}_v|\theta_y)$ describes the structure dependence among a set of values that $\mathbf{Y}_v$ may take. Second, $q(\mathbf{X}|\mathbf{Y}_v, \theta_{x|y})$ describes the knowledge about the dependence relation from inner representation to observation. Third, in addition to these structures, the knowledge is also represented by Θ jointly, which is confined by a background knowledge via a priori structure $q(\Theta|\Xi)$ with a unknown parameter set Ξ_q. Some choices are shown in Tab.2.

As to the Yang machine $p(\mathbf{X}, \mathbf{Y}|\Theta_p)$, we already have the above discussed input $p(\mathbf{X}|\theta_x)$. Similar to the case of Fig.2(b), the structures of $p(\mathbf{Y}|\mathbf{X}, \theta_{y|x}) = p(\mathbf{Y}_v|\mathbf{X}, L, \theta_{y|x})p(L|\mathbf{X}, \theta_{y|x})$ not only make a fast implementation of a desired problem solving but also act as an inverse role of the Ying machine $q(\mathbf{X}, \mathbf{Y}|\Theta_q)$. If we are also provided with the structure of $p(\Theta|\mathbf{X}, \Xi_p)$, what still remains unknown consists of $\mathbf{k}$ and $\Xi = \{\Xi_q, \Xi_p\}$. An analogy of this Ying Yang system to the ancient Ying-Yang philosophy motivates to determine the unknowns under a best harmony principle, which is mathematically implemented by maximizing the following harmony measure

$$
\begin{aligned}
&\max_{\{\mathbf{k},\Xi\}} H(p\|q, \mathbf{k}, \Xi),\ H(p\|q, \mathbf{k}, \Xi) = \int p(\mathbf{R}|\mathbf{X})p(\mathbf{X}) \ln [q(\mathbf{X}|\mathbf{R})q(\mathbf{R})] d\mathbf{X} d\mathbf{R} \\
&\qquad = \int p(\Theta|\mathbf{X}, \Xi) H_f(\mathbf{X}, \Theta, \mathbf{k}, \Xi) d\Theta, \\
&H_f(\mathbf{X}, \Theta, \mathbf{k}, \Xi) = \sum_L p(L|\mathbf{X}, \theta_{y|x}) H_f(\mathbf{X}, L, \Theta, \mathbf{k}, \Xi), \qquad (3) \\
&H_f(\mathbf{X}, L, \Theta, \mathbf{k}, \Xi) = \int p(\mathbf{Y}_v|\mathbf{X}, L, \theta_{y|x}) p(\mathbf{X}|\theta_x) \times
\end{aligned}
$$

$q(\mathbf{X},\mathbf{R})$	$q(\mathbf{Y}\|\boldsymbol{\theta}_{\mathbf{y}})$	$q(\mathbf{X}\|\mathbf{Y},\boldsymbol{\theta}_{\mathbf{x}\|\mathbf{y}})$	$q(\theta\|\Xi)=\prod_{\xi\in\{\mathbf{x}\|\mathbf{y},\mathbf{x},\mathbf{y}\|\mathbf{x},\mathbf{y}\}}q(\boldsymbol{\theta}_{\xi}\|\Xi_{\xi})$
Choice 1	$q(\mathbf{L}\|\boldsymbol{\theta}_{\mathbf{L}}),\mathbf{L}=\{\ell\}$ $\ell=\mathbf{1},\cdots,\mathbf{k}$	$q(\mathbf{X}\|\mathbf{L},\boldsymbol{\theta}_{\mathbf{L}})$	**Ignored**
Choice 2	$q(\mathbf{Y}\|\boldsymbol{\theta}_{\mathbf{y}}),\mathbf{Y}=\{y\}$ real $y=[y_1,\cdots,y_m]^{\mathrm{T}}$	$G(X\|g(Y,\boldsymbol{\theta}_{\mathbf{X}\|\mathbf{Y}}),\Sigma_{\mathbf{X}\|\mathbf{Y}})$ for real X	**Non-informative**
Choice 3	$q(\mathbf{Y}\|\boldsymbol{\theta}_{\mathbf{y}}),\mathbf{Y}=\{y\}$ binary $y=[y_1,\cdots,y_m]^{\mathrm{T}}$	$q(X\|Y,\boldsymbol{\theta}_{\mathbf{X}\|\mathbf{Y}})$ both X,Y are binary	$q(\theta)\propto\frac{1}{\sum_{t=1}^{N}q(u_t\|\theta)}$
Choice 4	**Hybrid** $q(\mathbf{Y},\mathbf{L}\|\boldsymbol{\theta}_{\mathbf{y}})=q(\mathbf{Y}\|\mathbf{L},\boldsymbol{\theta}_{\mathbf{y}})q(\mathbf{L}\|\boldsymbol{\theta}_{\mathbf{L}})$	**Hybrid** $q(\mathbf{X}\|\mathbf{Y},\mathbf{L},\boldsymbol{\theta}_{\mathbf{X}\|\mathbf{Y},\mathbf{L}})$	**Parametric** $q(\theta\|\Xi)$

	$p(X\|\theta_{\mathbf{x}})$	$p(L\|X,\boldsymbol{\theta}_{\mathbf{y}\|\mathbf{x}})$
Choice A	$\delta(X-X_N)$	**Free of structure**
Choice B	**Sample-based** $p_h(\mathbf{X})=\prod_{t=1}^{N}G(x\|x_t,h^2I)$	**Bayesian structure** $p(L\|X,\Theta_L)=\frac{q(L\|\theta_y)q(X\|L,\theta_{x\|y},\theta_y)}{\sum_L q(L\|\theta_y)q(X\|L,\theta_{x\|y},\theta_y)}$ $q(X\|L,\Theta_L)=\int q(X\|Y,\theta_{x\|y})q(Y\|L,\theta_y)dY$
Choice C	**Batch-based** $p_h(\mathbf{X})=\prod_{t=1}^{N}p_h(x_t)$ $p_h(x_t)=\frac{1}{N}\sum_{t=1}^{N}G(x\|x_t,h^2I)$	$\pi_L(X,\Theta_p)=c_L+\mathrm{B}_L\mathrm{vec}(\mathrm{X})+\beta_L\mathrm{vec}(\mathrm{X})^{\mathrm{T}}Q_L\mathrm{vec}(\mathrm{X})$ $p(L\|X,\Theta_L)=\frac{e^{\pi_L(X,\Theta_L)}}{\sum_L e^{\pi_L(X,\Theta_L)}}$, $\Theta_L=\{\mathrm{B}_L,Q_L,c_L\}$ could be (1) consisting of free unknown parameters, (2) either Q_L is the Hessian of $\ln[q(L\|\theta_y)q(X\|L,\theta_{x\|y},\theta_y)]$ with respect to vec(X) or a part of Q_L is the counterpart part of the Hessian, while the rest is unknown parameters.

	$p(\boldsymbol{\theta}\|X_N,\Xi)$	$p(\mathbf{Y}\|\mathbf{X},\mathbf{L},\boldsymbol{\theta}_{\mathbf{y}\|\mathbf{x}})$
Choice A	*Free of structure*	*Free of structure*
Choice B	*Bayesian structure* $p(\boldsymbol{\theta}\|X_N)=\frac{q(X_N\|\boldsymbol{\theta})q(\boldsymbol{\theta}\|\Xi)}{\int q(X_N\|\boldsymbol{\theta})q(\boldsymbol{\theta}\|\Xi)d\boldsymbol{\theta}}$	*Bayesian structure* $p(y\|x,L,\theta_{y\|x})=\frac{q(x\|y,L,\theta_{x\|y})q(y,L\|\theta_y)}{\int q(x\|y,L,\theta_{x\|y})q(y\|L,\theta_y)dy}$
Choice C	*Ying machine induced* $G(\mathrm{vec}(\theta)\|\mu(X_N,\Theta),\Sigma(X_N,\Theta))$ • $\mu(X_N,\Theta)$ is either $\theta^*(X_N,\Xi)=\mathrm{argmax}_\theta H_f(X_N,\theta,k,\Xi)$ or a prespecified linear or nonlinear function of X_N with unknown parameters Ξ. • $\Sigma^{-1}(X_N,\Theta)$ is either the Hessian of $H_f(X_N,\theta,k,\Xi)$ with respect to *vec(θ)* or a simplfied approximation of this Hessian.	*Ying machine induced* *For Y consisting of real variables* $G(vec(Y)\|\mu(X,\phi_{\mu,\mathrm{L}}),\Sigma(Y,\phi_{\Sigma,\mathrm{L}}))$ • $\mu(X,\phi_{\mu,\mathrm{L}})$ is a prespecifi ed function of X with unknown parameters $\phi_{\mu,\mathrm{L}}$. • $\Sigma^{-1}(Y,\phi_{\Sigma,\mathrm{L}})$ is either the Hessian of $ln[q(X\|Y,L,\theta_{x\|y})q(Y\|L,\theta_y)]$ with respect to *vec(Y)* or a simpilfied approximat ion of this Hessian. *For Y consisting of binary variables* $p(Y\|X,L,\theta_{y\|x})$ is in a given structure based on $\eta_L(X,\Theta_p)=c_{b,L}+\mathrm{B}_{b,L}\mathrm{vec}(\mathrm{X})+\mathrm{vec}(\mathrm{X})^{\mathrm{T}}Q_{b,L}\mathrm{vec}(\mathrm{X})$ similar to the Choice(C) of $p(L\|X,\theta_{y\|x})$.

Table 2 Typical scenarios of $q(\mathbf{X},\mathbf{R})=q(\mathbf{X}|\mathbf{Y},\theta_{x|y})q(\mathbf{Y}|\theta_y)q(\Theta|\Xi_q)$ and $p(\mathbf{X},\mathbf{R})=p(\Theta|\mathbf{X},\Xi_p)p(\mathbf{Y}|\mathbf{X},\theta_{y|x})p(\mathbf{X}|\theta_x)$

$$\times \ln [q(\mathbf{X}|\mathbf{Y}_v, L, \theta_{x|y})q(\mathbf{Y}_v|L, \theta_y)q(L|\theta_L)q(\Theta|\Xi_q)]d\mathbf{Y}_v.$$

On one hand, maximizing $H(p\|q)$ forces $q(\mathbf{X}|\mathbf{R})q(\mathbf{R})$ to match $p(\mathbf{R}|\mathbf{X})p(\mathbf{X})$. Due to the constraints on the given Ying and Yang structures, a perfect matching $p(\mathbf{R}|\mathbf{X})p(\mathbf{X}) = q(\mathbf{X}|\mathbf{R})q(\mathbf{R})$ may not be really reached but still be approached as possible as it can. At this equality, $H(p\|q)$ becomes the negative entropy that describes the complexity of system. Further maximizing $H(p\|q)$ with $\mathbf{k}, \Xi$ is actually minimizing the complexity of system, which provides a model selection ability on $\mathbf{k}$. Such an ability can also be observed from other perspectives, with details referred to [52, 49].

The first difficulty we encounter is where to get the structure of $p(\Theta|\mathbf{X}, \Xi_p)$ that specifies a probabilistic inverse mapping $\mathcal{X}_N \to \Theta$ shown in Fig.2(c). One possibility is the BI choice in the second column of Tab.1 or written as the choice B for $p(\Theta|\mathbf{X}, \Xi_p)$ in Tab.2. As previously discussed, it usually involves a difficult computation for not only an integral over Θ but also an integral over Y. To avoid this difficulty, we usually consider the choice A and choice C in Tab.2.

First, we consider Choice A, i.e., a $p(\Theta|\mathbf{X}, \Xi_p)$ free of structure. Maximizing $H(p\|q)$ with respect to such a $p(\Theta|\mathbf{X}, \Xi_p)$ leads to

$$p(\Theta|\mathbf{X}, \Xi_p) = \delta(\Theta - \Theta^*),\ \Theta^* = \max_{\Theta} H_f(\mathbf{X}, \Theta, \mathbf{k}, \Xi). \tag{4}$$

That is, the problem becomes seeking a best harmony between the front layer Ying-Yang pair. But it also incurs a problem. With $p(\mathbf{X}|\theta_x)$ given empirically from $\mathcal{X}_N$, the mapping to Θ^* from $\mathcal{X}_N$ of random samples is probabilistic. However, $\delta(\Theta - \Theta^*)$ can not take this uncertainty in consideration. Actually, $\Theta^*(\mathcal{X}_N)$ by eq.(4) only takes over the information of the first order statistics from the Ying machine. In other words, maximizing $H(p\|q)$ with respect to a free $p(\Theta|\mathbf{X}, \Xi_p)$ can only make a best Ying Yang harmony in term of the first order statistics.

This uncertainty is considered by $p(\Theta|\mathbf{X}, \Xi_p)$ in the Choice B or Choice C such that a best Ying Yang harmony in term of not only the first order statistics but also the statistics of the second order or higher. To be detailed in the next subsection, the approximation will make $H(p\|q, \mathbf{k}, \Xi)$ in eq.(3) approximately turned into the following format:

$$H(p\|q, \mathbf{k}, \Xi) = H_f(\mathcal{X}_N, \Theta^*, \mathbf{k}, \Xi) + \Delta(\Theta^*, \mathbf{k}, \Xi), \tag{5}$$

where Θ^* and $H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi)$ are given in eq.(4), and $\Delta(\Theta^*, \mathbf{k}, \Xi)$ either involves no integral over Θ or an integral over a subset of Θ that is analytically solvable. If the meta parmeters Ξ is given, we can directly maximize the above $H(p\|q, \mathbf{k}, \Xi)$ to select $\mathbf{k}$. If the meta parameters Ξ is unknown, we need to make $\max_{\Xi} H(p\|q, \mathbf{k}, \Xi)$ too. Actually, getting Θ^* by eq.(4) depends on Ξ. In other words, the process of seeking an approriate Ξ^* is coupled with finding Θ^*. In general, we can estimate Ξ^* and Θ^* jointly by iterating the following two steps:

$$\begin{aligned} &\Theta\ step: \Theta^{(t+1)} = \Theta^{(t)} + \eta \nabla_{\Theta} H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi^{(t)})_{\Theta=\Theta^{(t)}}, \\ &\quad or \quad \Theta^{(t+1)} = arg \max_{\Theta} H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi^{(t)}) \text{ if it is analytically solvable,} \end{aligned} \tag{6}$$

$$\Xi \ step : \Xi^{(t+1)} = \Xi^{(t)} + \eta \nabla_{\Xi \ at \ \Xi^{(t)}} [H_f(\mathcal{X}_N, \Theta^{(t+1)}, \mathbf{k}, \Xi) + \Delta(\Theta^{(t+1)}, \mathbf{k}, \Xi)],$$

which starts with an initialization $\Theta^{(0)}$ and $\Xi^{(0)}$ and reaches a convergence.

In a summary, the best Ying Yang harmony by maximizing $H(p\|q, \mathbf{k}, \Xi)$ is made via the following two stage implementation :

$$\begin{aligned} &Stage\ I: \ \text{get } \Xi^*, \Theta^* \text{ by eq.(6) for } \forall \mathbf{k} \in \mathcal{K}, \ \mathcal{K} \text{ is a set of values of } \mathbf{k}; \quad (7) \\ &Stage\ II : \mathbf{k}^* = arg \min_{\mathbf{k} \in \mathcal{K}} J(\mathbf{k}), J(\mathbf{k}) = -H_f(\mathcal{X}_N, \Theta^*, \mathbf{k}, \Xi^*) + \Delta(\Theta^*, \mathbf{k}, \Xi^*). \end{aligned}$$

As mentioned previously, the scale $\mathbf{k}$ of a BYY system is contributed from two parts. One is featured by $\mathbf{k}_Y$ for representing Y and the rest is featured by the number n_f of free parameters in Θ. The degrees of difficulty for estimating the two parts are quite different. When $q(\mathbf{Y}|\theta_y)$ is in a so called scale reducible structure, an appropriate $\mathbf{k}_Y$ will be determined automatically during parameter learning on Ξ^* and Θ^* by eq.(6), with $\mathbf{k}$ initialized at one big enough value. The details are referred to Sec.2.3. Interestingly, the model selection problem in many typical learning tasks [49, 52] can be reformulated into a BYY system for selecting merely this $\mathbf{k}_Y$ part. This favorable feature makes both parameter learning for Ξ^*, Θ^* and model selection for $\mathbf{k}_Y$ implemented simultaneously by only implementing eq.(6), which can significantly reduce the computational cost that is needed in a two stage implementation by eq.(7). However, the performance of this automatic model selection will deteriorate as the sample size N reduces. In such a case, we can implement both the stages in eq.(7) with a computational cost similar to those conventional two stage implementations of typical model selection criteria. Still, eq.(7) will provide an improvement over those typical criteria since the contribution by $\mathbf{k}_Y$ has been addressed more accurately, though the contribution featured by the number n_f of free parameters in Θ is roughly estimated in a way similar to those typical criteria.

2.2 Yang machine: Implementable scenarios

With $p(\mathbf{X}|\theta_x)$ given empirically from $\mathcal{X}_N$, i.e., Choice A in Tab. 2, it follows from eq.(3) that we further have

$$\begin{aligned} &H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi) = \sum_L p(L|\mathcal{X}_N, \theta_{y|x}) H_f(\mathcal{X}_N, L, \Theta, \mathbf{k}, \Xi), \\ &H_f(\mathcal{X}_N, L, \Theta, \mathbf{k}, \Xi) = \int p(\mathbf{Y}_v|\mathcal{X}_N, L, \theta_{y|x}) \mathcal{L}(\mathcal{X}_N, L, \mathbf{Y}, \Theta_q) d\mathbf{Y}_v - Z(\Theta|\Xi_q), \\ &\mathcal{L}(\mathcal{X}_N, L, \mathbf{Y}_v, \Theta_q) = \ln [q(\mathcal{X}_N|\mathbf{Y}_v, L, \theta_{x|y}) q(\mathbf{Y}_v|L, \theta_y) q(L|\theta_L)], \\ &Z(\Theta|\Xi_q) = -\ln q(\Theta|\Xi_q), \ \Theta_q = \{\theta_{x|y}, \theta_y, \theta_L\}. \end{aligned} \quad (8)$$

There still remains an integral over $\mathbf{Y}_v$, which is handled differently according to the choices of $p(\mathbf{Y}_v|\mathbf{X}, L, \theta_{y|x})$ in Tab.2, whenever there is no confusion, y_v is denoted by y for simplicity. For the Choice A, maximizing $H(p\|q)$ with respect to a $p(\mathbf{Y}_v|\mathbf{X}, L, \theta_{y|x})$ free of structure leads to

$$p(\mathbf{Y}_v|\mathbf{X}, L, \theta_{y|x}) = \delta(\mathbf{Y}_v - \mathbf{Y}^*_{vL}(\Theta_q)), \ \mathbf{Y}^*_{vL}(\Theta_q) = \max_{\mathbf{Y}_v} \mathcal{L}(\mathcal{X}_N, L, \mathbf{Y}_v, \Theta_q),$$

$$H_f(\mathcal{X}_N, L, \Theta, \mathbf{k}, \Xi) = \mathcal{L}(\mathcal{X}_N, L, \mathbf{Y}^*_{vL}(\Theta_q), \Theta_q) - Z(\Theta|\Xi_q). \quad (9)$$

The computational difficulty incurred by the integral over $\mathbf{Y}_v$ has been avoided. But it also incurs two problems. First, the above $\mathbf{Y}^*_{vL}(\Theta_q)$ may not have a differentiable expression with respect to Θ_q, or even no analytical expression. Thus, a gradient based algorithm for $\max_\Theta H(p\|q, \Theta)$ can not take the relation $\mathbf{Y}^*_{vL}(\Theta_q)$ in consideration, which makes learning fragile to local optimal performance. Second, the mapping from a set $\mathcal{X}_N$ of random samples to the inner representations is probabilistic while $\delta(\mathbf{Y}_v - \mathbf{Y}^*_{vL}(\Theta_q))$ can not take this uncertainty in consideration, since it only takes over the information of the first order statistics from the Ying machine. Similar to the discussion after eq.(4), considered in eq.(9) is a best Ying Yang harmony only in term of the first order statistics. It is improved by $p(\mathbf{Y}_v|\mathbf{X}, L, \theta_{y|x})$ in the Choice C of Tab.2 such that a best Ying Yang harmony in the front layer via approximately considering the second order statistics.

Considering a Taylor expansion of $Q(\xi)$ around $\xi^* = \max_\xi Q(\xi)$ up to the second order and noticing $\nabla_\xi Q(\xi) = 0$ at $\xi = \xi^*$, we approximately have

$$\int p(\xi)Q(\xi)d\xi \approx Q(\xi^*) + \frac{1}{2}Tr[\Sigma H_Q(\xi^*)],\ \Sigma = \int p(\xi)(\xi - \xi^*)(\xi - \xi^*)^T d\xi, \quad (10)$$

where the Hessian matrix $H_Q(\xi) = \partial^2 Q(\xi)/\partial\xi\partial\xi^T$ is negative definite in a neighborhood of ξ^*. Moreover, $q(\xi) = e^{-Q(\xi)}/\int e^{-Q(\xi)}d\xi$ defines a distribution. If we use $G(\xi|\mu, \Sigma)$ to approximate $q(\xi)$, the solution is $\mu = \xi^*, \Sigma = H_Q^{-1}(\xi^*)$ [45].

With $\mathbf{Y}_v$ as ξ and $\mathcal{L}(\mathcal{X}_N, L, \mathbf{Y}_v, \Theta_q)$ as $Q(\xi)$, it follows from eq.(8) and eq.(10) that we approximately have

$$\begin{aligned}
&H_f(\mathcal{X}_N, L, \Theta, \mathbf{k}, \Xi) \approx \mathcal{L}(\mathcal{X}_N, L, \mathbf{Y}^*_{vL}(\Theta_q), \Theta_q) - 0.5d_{\mathbf{k}_Y}(L, \Theta_q) - Z(\Theta),\\
&d_{\mathbf{k}_Y}(L, \Theta) = Tr[\Sigma_L(\mathbf{Y}^*_{vL}, \Theta)H_L(\mathbf{Y}_v, \Theta_q)]_{\mathbf{Y}_v = \mathbf{Y}^*_{vL}(\Theta_q)},\\
&H_L(\mathbf{Y}_v, \Theta_q) = -\frac{\partial^2 \mathcal{L}(\mathcal{X}_N, L, \mathbf{Y}_v, \Theta_q)}{\partial \mathbf{Y}_v \partial \mathbf{Y}_v^T},\\
&\Sigma_L(\mathbf{Y}^*_{vL}, \theta_{y|x}) = \int [\mathbf{Y}_v - \mathbf{Y}^*_{vL}(\Theta_q)][\mathbf{Y}_v - \mathbf{Y}^*_{vL}(\Theta_q)]^T p(\mathbf{Y}_v|\mathcal{X}_N, L, \theta_{y|x})d\mathbf{Y}_v.
\end{aligned}$$

Following the discussion after eq.(10), see the Choice C in Tab.2, we consider

$$p(\mathbf{Y}_v|\mathcal{X}_N, L, \theta_{y|x}) = G(\mathbf{Y}_v|\mu(\mathcal{X}_N, \phi_{\mu,L}), \Sigma(\mathcal{X}_N, \phi_{\Sigma,L})). \quad (11)$$

Let $\mu(\mathcal{X}_N, \phi_{\mu,L}) = \mathbf{Y}^*_{vL}(\Theta_q)$ and $\Sigma(\mathcal{X}_N, \phi_{\Sigma,L}) = H_L^{-1}(\mathbf{Y}^*_{vL}(\Theta_q), \Theta_q)$, we have

$$\Sigma_L(\mathbf{Y}^*_{vL}, \theta_{y|x}) = H_L^{-1}(\mathbf{Y}^*_v(\Theta_q), \Theta_q),\ d_{\mathbf{k}_Y}(\Theta) = Tr[I_Y] = d_Y, \quad (12)$$

where d_Y is the dimension of $\mathbf{Y}_v$. Comparing with eq.(9), we can find that the only difference is this integer d_Y. This term is useful in Stage II of eq.(7) for making model selection on $\mathbf{k}$. However, the two problems mentioned after eq.(9) largely remain. Alternatively, we let $\Sigma(\mathcal{X}_N, \phi_{\Sigma,L}) = H_L^{-1}(\mathbf{Y}^*_v(\Theta_q), \Theta_q)$ but leave $\mu(\mathcal{X}_N, \phi_{\mu,L})$ to be a parametric function (e.g., a linear function or nonlinear function) with a unknown set $\phi_{\mu,L}$. Let $\mathbf{Y}_v - \mathbf{Y}^*_{vL} = \mathbf{Y}_v - \mu(\mathcal{X}_N, \phi_{\mu,L}) + \mu(\mathcal{X}_N, \phi_{\mu,L}) - \mathbf{Y}^*_{vL}$, we have

$$\Sigma(\mathbf{Y}^*_{vL}, \theta_{y|x}) = H_L^{-1}(\mathbf{Y}^*_{vL}, \Theta_q) + e(\mathbf{Y}^*_{vL}, \Theta)e^T(\mathbf{Y}^*_{vL}, \Theta),$$

$$d_{\mathbf{k}_Y}(L,\Theta) = d_Y + e^T(\mathbf{Y}^*_{vL},\Theta)H_L(\mathbf{Y}^*_{vL},\Theta_q)e(\mathbf{Y}^*_{vL},\Theta),$$
$$\mathbf{Y}^*_{vL} = \mathbf{Y}^*_{vL}(\Theta_q),\ e(\mathbf{Y}^*_{vL},\Theta) = \mu(\mathcal{X}_N,\phi_{\mu,L}) - \mathbf{Y}^*_{vL}. \tag{13}$$

In addition to d_Y, the second term in $d_{\mathbf{k}_Y}(L,\Theta)$ takes uncertainties in consideration. This term is updated via Θ_q and will gradually disappear as learning converges and $e(\mathbf{Y}^*_{vL},\Theta)$ tends to 0. Even without an analytical expression for $\mathbf{Y}^*_v(\Theta_q)$, we can get a value $\mathbf{Y}^*_{vL}$ via $\max_\Theta H(p\|q,\Theta)$ and then update Θ with the above $d_{\mathbf{k}_Y}(L,\Theta)$ in effect by certain extent. If $\mathbf{Y}^*_{vL}(\Theta_q)$ is obtained in a differentiable expression, a further improvement can be obtained via further taking $\nabla_{\Theta_q}\mathbf{Y}^*_{vL}(\Theta_q)$ in consideration through the chain rule.

When $\mathbf{Y}_v$ consists of vectors in binary variables. The above approach does not apply because we can not use eq.(10). In these cases, the integral over $\mathbf{Y}_v$ becomes summation, which can be computed but usually with a high computational complexity. Still, we can consider $p(\mathbf{Y}_v|\mathcal{X}_N,L,\theta_{y|x})$ in a parametric structure to facilitate the computation. An example is listed in Tab.2 and details will be further discussed in Sec.2.3.

We continue to proceed beyond the case of eq.(4) by considering $p(\Theta|\mathcal{X})$ in the Choice C of Tab.2 for a best Ying Yang harmony with not only $\Theta^*(\mathcal{X}_N)$ but also its corresponding second order statistics in consideration. With $p(\mathbf{X}|\theta_x)$ given empirically from $\mathcal{X}_N$, i.e., Choice A in Tab. 2, from eq.(3) we have $\int p(\Theta|\mathcal{X}_N,\Xi)H_f(\mathcal{X}_N,\Theta,\mathbf{k},\Xi)d\Theta$. Regarding Θ as ξ and $H_f(\mathcal{X}_N,\Theta,\mathbf{k},\Xi)$ as $Q(\xi)$, it follows again from eq.(10) that we approximately get eq.(5), that is

$$H(p\|q,\mathbf{k},\Xi) = H_f(\mathcal{X}_N,\Theta^*,\mathbf{k},\Xi) + \Delta(\Theta^*,\mathbf{k},\Xi),\ \Theta^* = \max_\Theta H_f(\mathcal{X}_N,\Theta,\mathbf{k},\Xi),$$
$$\Delta(\Theta^*,\mathbf{k},\Xi) = -0.5d_{\mathbf{k}},\ d_{\mathbf{k}} = Tr[\Sigma(\Theta^*)H_H(\Theta^*)], \tag{14}$$
$$\Sigma(\Theta^*) = \int(\Theta-\Theta^*)(\Theta-\Theta^*)^T p(\Theta|\mathcal{X}_N)d\Theta,\ H_H(\Theta) = -\frac{\partial^2 H_f(\mathcal{X}_N,\Theta,\mathbf{k},\Xi)}{\partial\Theta\partial\Theta^T}.$$

Further let $p(\Theta|\mathcal{X}_N) = G(\Theta|\Theta^*,H_H^{-1}(\Theta^*))$, we get that $d_{\mathbf{k}} = Tr[I]$ is the number n_f of free parameters in Θ [46, 47, 51]. It follows from eq.(14) that both Θ^* and $H_H(\Theta^*)$ depend $\mathcal{X}_N,\Xi$. Let $p(\Theta|\mathcal{X}_N) = G(\Theta|\mu(\mathcal{X}_N,\Xi),H_H^{-1}(\Theta^*))$ with $\mu(\mathcal{X}_N,\Xi)$ in a parametric function of $\mathcal{X}_N,\Xi$, similar to eq.(13) we also get

$$\Delta(\Theta^*,\mathbf{k},\Xi) = -0.5d_{\mathbf{k}},$$
$$d_{\mathbf{k}} = n_f + (\mu(\mathcal{X}_N,\Xi)-\Theta^*)^T H_H(\Theta^*)(\mu(\mathcal{X}_N,\Xi)-\Theta^*). \tag{15}$$

Revising the direction of thinking, put $p(\Theta|\mathcal{X}_N) = G(\Theta|\mu(\mathcal{X}_N,\Xi),H_H^{-1}(\Theta^*))$ into eq.(3) we can also consider

$$H(p\|q,\mathbf{k},\Xi) = \int G(\Theta|\mu(\mathcal{X}_N,\Xi),H_H^{-1}(\Theta^*))H_f^-(\mathcal{X}_N,\Theta,\mathbf{k},\Xi)d\Theta + \Delta(\Theta^*,\mathbf{k},\Xi),$$
$$\Delta(\Theta^*,\mathbf{k},\Xi) = \int G(\Theta|\mu(\mathcal{X}_N,\Xi),H_H^{-1}(\Theta^*))\ln q(\Theta|\Xi_q)d\Theta,$$
$$H_f^-(\mathcal{X}_N,\Theta,\mathbf{k},\Xi) = \sum_L \int p(L|\mathcal{X}_N,\theta_{y|x})p(\mathbf{Y}_v|\mathcal{X}_N,L,\theta_{y|x})\mathcal{L}(\mathcal{X}_N,L,\mathbf{Y}_v,\Theta_q)d\mathbf{Y}_v,$$

where $\mathcal{L}(\mathcal{X}_N,L,\mathbf{Y}_v,\Theta_q)$ is still given by eq.(8). There may be two ways to handle the integral in the first term. One is considering several typical structures on which the integral can be handled, with details referred to Sec.2.3. The other

way is similar to eq.(10). Considering a Taylor expansion of $Q(\xi)$ around μ up to the second order, we approximately have

$$\int G(\xi|\mu,\Sigma)Q(\xi)d\xi \approx Q(\xi)_{\xi=\mu} + \frac{1}{2}Tr[\Sigma\partial^2 Q(\xi)/\partial\xi\partial\xi^T]_{\xi=\mu}. \tag{16}$$

With $G(\Theta|\mu(\mathcal{X}_N,\Xi), H_H^{-1}(\Theta^*))$ as ξ and $H_f^-(\mathcal{X}_N,\Theta,\mathbf{k},\Xi)$ as $Q(\xi)$, we get

$$\begin{aligned}&\int G(\Theta|\mu(\mathcal{X}_N,\Xi), H_H^{-1}(\Theta^*))H_f^-(\Theta,\mathbf{k},\Xi)d\Theta = H_f^-(\mathcal{X}_N,\mu(\mathcal{X}_N,\Xi),\mathbf{k},\Xi)\\ &+\frac{1}{2}Tr[H_H^{-1}(\Theta^*)\{\partial^2 H_f^-(\Theta,\mathbf{k},\Xi)/\partial\Theta\partial\Theta^T\}]_{\Theta=\mu(\mathcal{X}_N,\Xi)}.\end{aligned} \tag{17}$$

For the second term of $H(p\|q,\mathbf{k},\Xi)$, i.e., $\Delta(\Theta^*,\mathbf{k},\Xi)$, we may handle it by either observing the specific structure of $q(\Theta|\Xi_q)$ or using eq.(16). By the latter, we reach $H(p\|q,\mathbf{k},\Xi)$ in eq.(14) again but with

$$\Delta(\Theta^*,\mathbf{k},\Xi) = -0.5d_{\mathbf{k}},\ d_{\mathbf{k}} = Tr[H_H^{-1}(\Theta^*)H_H(\mu(\mathcal{X}_N,\Xi))], \tag{18}$$

which becomes $d_{\mathbf{k}} = n_f$ when $\mu(\mathcal{X}_N,\Xi)$ reaches Θ^*.

Generally, it is difficult to consider $p(\Theta|\mathcal{X})$ given by a Bayesian structure, i.e., the Choices B in Tab.2. In some cases, we may divide the set Θ into two parts Θ' and Θ'' and assume that $p(\Theta|\mathcal{X}) = p(\Theta'|\mathcal{X})p(\Theta'|\mathcal{X})$ and $q(\Theta|\Xi_q) = q(\Theta'|\Xi_q')q(\Theta''|\Xi_q'')$. For one part Θ'', we may handle $\int p(\Theta''|\mathcal{X})\ln q(\Theta''|\Xi_q)d\Theta''$ analytically. Then, the remining parts are handled as before.

The last but not least, we further proceed to the case that $p(\mathbf{X}|\theta_x) = p_h(\mathbf{X})$ is given by Choice B in Tab. 2, with an extra unknown input h in consideration. Let $\mathbf{X}$ to be replaced by $\mathbf{X}, h$ and notice that only $\mathbf{X}$ relates to h, we have

$$\begin{aligned}&p(\mathbf{X},h) = p(\mathbf{X}|h)p(h),\ p(\mathbf{X}|h) = p_h(\mathbf{X}),\ p(\mathbf{R}|\mathbf{X},h) = p(\mathbf{R}|\mathbf{X}),\\ &q(\mathbf{X},h|\mathbf{R}) = q(h|\mathbf{X},\mathbf{R})q(\mathbf{X}|\mathbf{R}),\ q(h|\mathbf{X},\mathbf{R}) = q(h|\mathbf{X}),\end{aligned} \tag{19}$$

with $q(\mathbf{R})$ remains unchanged. Put it into eq.(3), we have

$$H(p\|q,\mathbf{k},\Xi) = \int p(h)p(\Theta|\mathbf{X},\Xi)H_f(\mathbf{X},\Theta,h,\mathbf{k},\Xi)d\Theta d\mathbf{X}dh \tag{20}$$

Maximizing $H(p\|q,\mathbf{k},\Xi)$ with $p(h)$ free of constraint leads to

$$\begin{aligned}&p(h) = \delta(h-h^*),\ h^* = arg\max_h H_h(p\|q,\mathbf{k},\Xi),\\ &H_h(p\|q,\mathbf{k},\Xi) = \int p(\Theta|\mathbf{X},\Xi)H_f(\mathbf{X},\Theta,h,\mathbf{k},\Xi)d\Theta.\end{aligned} \tag{21}$$

Equivalently, h^* is also given by $h^* = arg\max_h H_f(\mathbf{X},\Theta,h,\mathbf{k},\Xi)$. Moreover, it further follows from eq.(16) that we have

$$\begin{aligned}&H_f(\mathbf{X},\Theta,h,\mathbf{k},\Xi) = H_f(\mathcal{X}_N,\Theta,\mathbf{k},\Xi) + 0.5h^2Tr[\Sigma(\mathcal{X}_N)] - Z(h),\\ &Z(h) = -\ln q(h|\mathcal{X}_N),\ \Sigma(\mathbf{X}) = \frac{\partial^2\sum_L p(L|\mathcal{X}_N,\theta_{y|x})\ln q(\mathbf{X}|\mathbf{Y}_v,L,\theta_{x|y})}{\partial\mathbf{X}\partial\mathbf{X}^T}.\end{aligned}$$

An example of $q(h|\mathcal{X}_N)$ is obtained from $p_h(\mathbf{X})$ by eq.(29) in the next subsection.

Therefore, we can modify the two stage implementation by eq.(6) and eq.(7) with all the appearances of Θ replaced by $\{\Theta,h\}$ and $H_f(\mathcal{X}_N,\Theta,\mathbf{k},\Xi)$ by $H_f(\mathcal{X}_N,\Theta,h,\mathbf{k},\Xi)$. Recalling the discussion after eq.(7), the performance of automatic model selection on $\mathbf{k}_Y$ by eq.(6) will deteriorate as the sample size N reduces. With an approriate h^* learned together with Θ^*, a considerable improvement can be obtained to reduce this deterioration, as verified by experiments [35].

2.3 Ying machine : distributed log-quadratic inner structures

We proceed to typical scenarios of $q(\mathbf{X}, \mathbf{R})$. To accommodate the world knowledge and the dependences underlying $\mathcal{X}_N$ appropriately, there are several issues to be considered, consisting of inner representation with a wide coverage of typical learning tasks, computational feasibility, scalability of complicated problems, and structural scale reducibility that does not impede automatic model selection.

y	$\{l\}$	$\{y\}$	$\{y,l\}$
$q(y\|\theta_y)$	$q(l)=\alpha_l \geq 0, \sum_{l=1}^{k}\alpha_l = 1$	$q(y)=\prod_{j=1}^{m} q(y^{(j)})$	$q(y,l)=q(y\|l)\alpha_l, q(y\|l)=\prod_{j=1}^{m_l} q(y^{(j)}\|l)$

(a)

Case	(1) Gaussian	(2) Bernoulli	(3) $y^{(j)}=\{y_g^{(j)}, i^{(j)}\}$	(4) Mixture
$q(y^{(j)})$	$G(y^{(j)}\|0,\lambda^{(j)})$	$q_j^{y^{(j)}}(1-q_j)^{1-y^{(j)}}$	$G(y_g^{(j)}\|v^{(j,i)},\lambda^{(j,i)})\beta^{(j,i)}$	$\sum_{i=1}^{\kappa^{(j)}}\beta^{(j,i)}q^{(i)}(y_g^{(j)})$
$q(y^{(j)}\|l)$	$G(y^{(j)}\|0,\lambda_l^{(j)})$	$q_{l,j}^{y^{(j)}}(1-q_{l,j})^{1-y^{(j)}}$	$G(y_g^{(j)}\|v_l^{(j,i)},\lambda_l^{(j,i)})\beta_l^{(j,i)}$	$\sum_{i=1}^{\kappa^{(j)}}\beta_l^{(j,i)}q_l^{(i)}(y_g^{(j)})$
Note: for Type (3) we have $q(y^{(j)}\|l)=q(y_g^{(j)},i^{(j)}\|l)=q(y_g^{(j)}\|i^{(j)},l)q(i^{(j)}\|l)$ with $q(y_g^{(j)}\|i^{(j)},l)=G(y_g^{(j)}\|v_l^{(j,i)},\lambda_l^{(j,i)})$, $q(i^{(j)}\|l)=\beta_l^{(j,i)}\geq 0$, $\sum_{i=1}^{\kappa_l^{(j)}}\beta_l^{(j,i)}=1$.				

(b)

Type	A	B	C	$q(x\|y,\theta_{x\|y})=G(x\|\mu_y,\Sigma_y)$	Case	(1)	(2)	(3)	(4)
y	$\{l\}$	$\{y\}$	$\{y,l\}$		Noises relates to	none of $\{y,l\}$	only $\{l\}$	only $\{y\}$	both $\{y,l\}$
μ_y	μ_l	$Ay+\mu$	$A_l y+\mu_l$		Σ_y	Σ	Σ_l	Σ_y	$\Sigma_{y,l}$

(c)

Table 3 Typical structures of $q(y)=q(y|\theta_y)$ and $q(x|y)=q(x|y,\theta_{x|y})=G(x|\mu_y,\Sigma_y)$

For many learning tasks, these issues are collectively considered in a class of structures for Ying machine, namely distributed log-quadratic inner structures. As already discussed in Sec.2.1, we consider a distributed inner representation $\mathbf{Y}=\{\mathbf{Y}_v,\mathbf{L}\}$, i.e., vector based inner representations $\mathbf{Y}_v=\{y_v\}$ are distributively and collaboratively described in a collection $\mathbf{L}=\{\ell\}$ with the help of $q(\mathcal{X}_N|\mathbf{Y}_v,\mathbf{L},\theta_{x|y})$ and $q(\mathbf{Y}_v|\theta_y)$ in the following log-quadratic structures :

$$\begin{aligned}&\mathcal{L}(\mathcal{X}_N,\mathbf{L},\mathbf{Y},\Theta_q)=\ln[q(\mathcal{X}_N|\mathbf{Y}_v,\mathbf{L},\theta_{x|y})q(\mathbf{Y}_v|\mathbf{L},\theta_y)q(\mathbf{L})]\\&=Tr[\Phi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q)\mathbf{Y}_v\mathbf{Y}_v^T+\Psi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q)\mathbf{Y}_v+\phi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q)]+\ln q(\mathbf{L}),\end{aligned}\tag{22}$$

where $\Phi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q),\Psi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q),\ \phi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q)$ are in given expressions. Eq.(8) becomes

$$H_f(\mathcal{X}_N,\Theta,\mathbf{k},\Xi)=\sum_{\mathbf{L}}p(\mathbf{L}|\mathcal{X}_N,\theta_{y|x})\ln q(\mathbf{L})+\sum_{\mathbf{L}}p(\mathbf{L}|\mathcal{X}_N,\theta_{y|x})Tr[\phi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q)]+$$

$$\sum_{\mathbf{L}} p(\mathbf{L}|\mathcal{X}_N, \theta_{y|x}) Tr[\Psi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q)\mu(\mathcal{X}_N, \theta_{y|x}) + \Phi_{\mathcal{X}_N}(\Theta_{\mathbf{L}}^q)\Gamma_{\mathbf{L}}(\mathcal{X}_N, \theta_{y|x})] - Z(\Theta),$$
$$\mu(\mathcal{X}_N, \theta_{y|x}) = \int \mathbf{Y}_v p(\mathbf{Y}_v|\mathcal{X}_N, \mathbf{L}, \theta_{y|x}) d\mathbf{Y}_v,$$
$$\Gamma_{\mathbf{L}}(\mathcal{X}_N, \theta_{y|x}) = \int \mathbf{Y}_v \mathbf{Y}_v^T p(\mathbf{Y}_v|\mathcal{X}_N, \mathbf{L}, \theta_{y|x}) d\mathbf{Y}_v. \quad (23)$$

y	$\{l\}$	$\{y_v\}$	$\{y_v, l\}$
$p(y\|x,\theta_{y\|x})$	$p(l\|x)=\alpha_l(x)\geq 0,\ \alpha_l(x)=\varsigma_l(x)\Big/\sum_{l=1}^{k}\varsigma_l(x)$	$p(y_v\|x)$	$p(y_v,l\|x)=p(y_v\|x,l)\alpha_l(x)$

(A)

Case	(1) Gaussian	(2) Bernoulli	(3) $y=[y_g,\{i^{(j)}\}]$
$p(y\|x)$	$G(y\|\zeta(x),\Gamma)$	$\prod_{j=1}^{m}\varsigma_j^{y^{(j)}}(x)[1-\varsigma_j(x)]^{1-y^{(j)}}$	$G(y_g\|\varsigma(x),\Gamma)\prod_{j=1}^{m}\beta(y_g^{(j)})$
$p(y\|x,l)$	$G(y\|\zeta_l(x),\Gamma_l)$	$\prod_{j=1}^{m_l}\varsigma_{j,l}^{y^{(j)}}(x)[1-\varsigma_{j,l}(x)]^{1-y^{(j)}}$	$G(y_g\|\varsigma_l(x),\Gamma_l)\prod_{j=1}^{m_l}\beta_l(y_g^{(j)})$
For case (3), $p(y\|x,l)=p(y_g,\{i^{(j)}\}\|x,l)=p(y_g\|x,l)\prod_{j=1}^{m_l}p(i^{(j)}\|y_g^{(j)},l)$			
$p(y_g\|x,l)=G(y_g\|\varsigma_l(x),\Gamma_l),\ p(i^{(j)}\|y_g^{(j)},x,l)=p(i^{(j)}\|y_g^{(j)},l)=\beta_l(y_g^{(j)})$			
Γ , Γ_l given by eq.(24) in the i.i.d. special case.			
$\beta_l(y_g^{(j)})=exp(\pi_l^{(j,i)})\Big/\sum_{i=1}^{\kappa_l^{(j)}}exp(\pi_l^{(j,i)})$			
(a) Posteriori $\pi_l^{(j,i)}=ln[G(y_g^{(j)}\|\nu_l^{(j,i)},\lambda_l^{(j,i)})\beta_l^{(j,i)}]$		(b) $\pi_l^{(j,i)}$ or $\beta_l(y_g^{(j)})$ is a free constant	(c) Parametric $\pi_l^{(j,i)}=b_0+b_1y_g^{(j)}+b_2y_g^{(j)2}$

(B)

$\varsigma_l(x)=[\varsigma_{1,l}(x),\ldots,\varsigma_{m_l,l}(x)]^{\mathrm{T}},\ \varsigma(x)=\varsigma_l(x)$ at $k=1$, $o_{j,l}(x)=\beta_l x^T H_{jl}x+w_{jl}^T x+c_{jl}$, see eq.(27)		
(1) Gaussian	(2) Bernoulli	(3) Multinomial $\{l\}$
$\varsigma_l(x)=o_l$ with $U_{j,l}=0,\forall j$	$\varsigma_{j,l}(x)=s(o_{j,l}),\ 0\leq s(r)\leq 1$ is a sigmoid function, e.g., $(1+e^{-r})^{-1}$	$\varsigma_{j,l}(x)=s(o_{j,l})$ with $m_l=1,\forall l$ $0\leq s(r)$ is monotonic, e.g., e^{-r}

(C)

Table 4 Typical parametric structures of $p(y|x,\theta_{y|x})$

As a result, the integral over $\mathbf{Y}_v$ for $H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi)$ has been turned into the integrals for getting the first order and second order statistics of $p(\mathbf{Y}_v|\mathcal{X}_N, \mathbf{L}, \theta_{y|x})$. Let $p(\mathbf{Y}_v|\mathcal{X}_N, \mathbf{L}, \theta_{y|x}) = G(\mathbf{Y}_v|\mathbf{Y}_{vL}^*(\Theta_q), H_L^{-1}(\mathbf{Y}_{vL}^*(\Theta_q))$, we have $\mu(\mathcal{X}_N, \theta_{y|x}) = \mathbf{Y}_{vL}^*(\Theta_q)$ and $\Gamma_{\mathbf{L}}(\mathcal{X}_N, \theta_{y|x}) = H_L^{-1}(\mathbf{Y}_{vL}^*(\Theta_q) + \mathbf{Y}_{vL}^*(\Theta_q)\mathbf{Y}_{vL}^{*T}(\Theta_q)$ and thus it returns back to the situation same as that in eq.(11) and eq.(12). From eq.(11) with $\Sigma(\mathcal{X}_N, \phi_{\Sigma,L}) = H_L^{-1}(\mathbf{Y}_{vL}^*(\Theta_q), \Theta_q)$, we also have

$$\mu(\mathcal{X}_N, \theta_{y|x}) = \mu(\mathcal{X}_N, \phi_{\mu,L}),$$
$$\Gamma_{\mathbf{L}}(\mathcal{X}_N, \theta_{y|x}) = H_L^{-1}(\mathbf{Y}^*(\Theta_q), \Theta_q) + \mu(\mathcal{X}_N, \phi_{\mu,L})\mu^T(\mathcal{X}_N, \phi_{\mu,L}), \quad (24)$$

and thus encounter a situation equivalent to that by eq.(11) and eq.(13).

Beyond those cases based on eq.(10), eq.(23) is also applicable to the cases that $\mathbf{Y}_v$ consists of binary or discrete variables. It is applicable to any structures in the log-quadratic form by eq.(22) or in this form approximately. Beyond eq.(11), we consider $p(\mathbf{Y}_v|\mathcal{X}_N, \mathbf{L}, \theta_{y|x})$ in a structure that the integrals for $\mu(\mathcal{X}_N, \theta_{y|x})$ and $\mathit{\Gamma}_{\mathbf{L}}(\mathcal{X}_N, \theta_{y|x})$ in eq.(23) can be solved analytically or computed efficiently. Moreover, instead of specifying the entire $p(\mathbf{Y}_v|\mathcal{X}_N, \mathbf{L}, \theta_{y|x})$, we can merely design $\mu(\mathcal{X}_N, \theta_{y|x})$ and $\mathit{\Gamma}_{\mathbf{L}}(\mathcal{X}_N, \theta_{y|x})$ in certain pre-specified parametric functions that are analytically computable.

The above discussions apply to the general scenarios that there maybe also temporal or even graphical dependence among that elements of $\mathbf{X} = \{x\}$. Correspondingly, we consider certain structure among the elements of $\mathbf{Y} = \{y_v, \ell\}$ to accommodate this dependence. E.g., further details about temporal dependences are referred to [59, 50]. To get further insights, here we focus on the cases that the elements of $\mathbf{X} = \{x\}$ are independently and identically distributed (i.i.d.), and thus the elements of $\mathbf{Y} = \{y_v, \ell\}$ are i.i.d. That is, we consider

$$\begin{aligned} &q(\mathbf{Y}|\theta_y) = \prod q(y_v, \ell|\theta_y),\ q(y_v, \ell|\theta_y) = q(y_v|\ell, \theta_{y,\ell})q(\ell), \\ &p(\mathbf{Y}|\mathbf{X}, \theta_{y|x}) = \prod p(y_v, \ell|x, \theta_{y|x}),\ q(\mathbf{X}|\mathbf{Y}, \theta_{x|y}) = \prod q(x|y_v, \ell, \theta_{x|y}), \end{aligned} \quad (25)$$

with $p(y_v, \ell|x, \theta_{y|x}) = p(y_v|x, \ell, \theta_{y|x,\ell})p(\ell|x, \theta_{\ell|x})$.

Shown in Tab. 3 and Tab.4 are several typical structures, covering several typical learning tasks [49]. All of them satisfy the format by eq.(23), except the case (d) for $q(y|\theta_y)$. Even for this exceptional case as well as other structures that fail to satisfy the format by eq.(23), we can still approximately use a Taylor expansion of $\ln[q(x|y, \theta_{x|y})q(y|\theta_y)]$ or $\ln q(y^{(j)}|\ell)$ with respect to y up to the second order. Readers are also referred to [53, 46, 47] for various other choices.

In the i.i.d. cases by eq.(25) and with the structures given in Tabs. 3 & 4, we can get a further insight on eq.(23) via a more detailed expression. Considering $q(y|\theta_y)$ at its Case (1) & Case (2) in Tab.3(b), we write eq.(23) into

$$\begin{aligned} &H_f(\mathcal{X}_N, \mathit{\Theta}, \mathbf{k}, \mathit{\Xi}) = \sum_t \sum_\ell p(\ell|x_t, \theta_{y|x})[\ln \alpha_\ell + \phi(x_t, \mathit{\Theta}_q^\ell)] - Z(\mathit{\Theta}) + \\ &\sum_t \sum_\ell p(\ell|x_t, \theta_{y|x}) Tr\{\mathit{\Psi}(x_t, \mathit{\Theta}_q^\ell)\mu_\ell(x_t) + \mathit{\Phi}(x_t, \mathit{\Theta}_q^\ell)[\mathit{\Gamma}_\ell(x_t) + \mu_\ell(x_t)\mu_\ell^T(x_t)]\}, \\ &\mu_\ell(x) = \textstyle\int y p(y|x, \ell)dy,\ \mathit{\Gamma}_\ell(x) = \int (y - \mu_\ell(x))(y - \mu_\ell(x))^T p(y|x, \ell)dy, \end{aligned} \quad (26)$$

The analytical expressions for $\mu_\ell(x)$ and $\mathit{\Gamma}_\ell(x)$, as well as the corresponding $\phi(x_t, \mathit{\Theta}_q^\ell)$, $\mathit{\Psi}(x_t, \mathit{\Theta}_q^\ell)$, and $\mathit{\Phi}(x_t, \mathit{\Theta}_q^\ell)$ are given in Tab.5.

Moreover, $p(\ell|x_t, \theta_{y|x})$ is given by Tab.4(a) with $\zeta_\ell(x)$ in the choice (3) of Tab.4(c), which is a simplification of $p(\mathbf{L}|\mathcal{X}_N, \theta_{y|x})$ in the choice B of Tab.2. Also, we can get the simplified counterpart of the choice C of Tab.2 as follows

$$p(\ell|x_t, \theta_{y|x}) = e^{-o_\ell(x_t)} / \sum_{j=1}^k e^{-o_j(x_t)},\ o_\ell(x) = \beta_\ell x^T H_\ell x + b_\ell^T x + c_\ell,$$

$$H_\ell = \partial^2 \ln \int q(x|y_v, \ell, \theta_{x|y}) q(y_v|\ell, \theta_y) dy_v / \partial x \partial x^T. \quad (27)$$

Furthermore, we are ready to elaborate the issue of structural scale reducibility, mentioned several times previously but without a further interpretation yet. As discussed after eq.(3), maximizing $H(p\|q, \mathbf{k}, \Xi)$ will push $\mathbf{k}$ as least as possible. Similarly, $\max_{\Theta,h} H_f(\mathcal{X}_N, \Theta, h, \mathbf{k}, \Xi)$ will push the entropy of $q(\mathcal{X}_N|\mathbf{Y}_v, \mathbf{L}, \theta_{x|y})$ $q(\mathbf{Y}_v|\mathbf{L}, \theta_y) q(\mathbf{L})$ as least as possible. One part of this entropy is contributed from the representation scale $\mathbf{k}_Y$ of $\mathbf{Y} = \{\mathbf{Y}_v, \mathbf{L}\}$. Interestingly, each integer of $\mathbf{k}_Y$ is associated with one or several parameters in Θ_q, and the least complexity nature will push these parameters towards 0 if the corresponding integer represents a redundant scale part. We say that $q(\mathcal{X}_N|\mathbf{Y}_v, \mathbf{L}, \theta_{x|y}) q(\mathbf{Y}_v|\mathbf{L}, \theta_y) q(\mathbf{L})$ has scale reducibility if there is no constraint to impede or block these parameters to be pushed towards 0.

$p(y\|x,l)$	$G(y\|\zeta_l(x),\Sigma_l)$	$\prod_{j=1}^{m_l} \varsigma_{j,l}^{y^{(j)}}(x)[1-\varsigma_{j,l}(x)]^{1-y^{(j)}}$
$\mu_l(x)$	$\zeta_l(x)$	$\varsigma_l(x) = [\varsigma_{1,l}(x), \ldots, \varsigma_{m_l,l}(x)]^{\mathrm{T}}$
$\Sigma_l(x)$	Σ_l	$diag[\varsigma_{1,l}(x) - \varsigma_{1,l}^2(x), \ldots, \varsigma_{m_l,l}(x) - \varsigma_{m_l,l}^2(x)]^{\mathrm{T}}$

(b)

$q(x\|y,l) = G(x\|A_l y + \mu_l, \Sigma_l)$		
$q(y\|l)$	Gaussian $G(y\|0,\Lambda_l)$	Bernoulli $\prod_{j=1}^{m_l} q_{l,j}^{y^{(j)}}(1-q_{l,j})^{1-y^{(j)}}$
$\Phi(x,\theta_q^l)$	$-0.5(\Lambda_l^{-1} + \Sigma_{A,l}^{-1})$	$-0.5\Sigma_{A,l}^{-1}$
$\Psi(x,\theta_q^l)$	$e_l(x)$	$[ln\frac{q_{l,1}}{1-q_{l,1}}, \ldots, ln\frac{q_{l,m_l}}{1-q_{l,m_l}}]^{\mathrm{T}} + e_l(x)$
$-\phi(x,\theta_q^l)$	$-0.5[D_{\Sigma_l}(x) + \ln\|\Lambda_l\| + m_l \ln(2\pi)]$	$-0.5 D_{\Sigma_l}(x) + \sum_{j=1}^{m_l}(1-q_{l,j})$
$e_l(x) = (x-\mu_l)\Sigma_l^{-1}A_l, \Sigma_{A,l}^{-1} = A_l^T \Sigma^{-1} A_l$, $D_{\Sigma_l}(x) = (x-\mu_l)\Sigma_l^{-1}(x-\mu_l) + \ln\|\Sigma_l\| + d\ln(2\pi)$		

(a)

Table 5 Major terms of $H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi)$ with $p(y|x,\ell)$ in Tab. 4, $q(x|y,\theta_{x|y})$ and $q(y|\theta_y)$ at Case (1) & Case (2) in Tab. 3

For the structures in eq.(25), $\mathbf{k}_Y$ consists of k and $\{m_\ell\}_{\ell=1}^k$. We observe that pushing one $q(\ell) = \alpha_\ell$ towards zero is equivalent to reducing the scale k to $k-1$, with all the corresponding structures discarded in effect. Moreover, pushing the variance of $q(y^{(j)}|\ell)$ towards 0 means that those of the j-th dimension can be discarded. We say that $q(\ell)$ is scale reducible if no constraint prevents $q(\ell)$ to become 0 for every ℓ, and that $q(y_v|\ell, \theta_{y,\ell})$ is scale reducible if no constraint prevents $q(y^{(j)}|\ell, \theta_{y,\ell})$ to become zero for every j, ℓ. Thus, $q(y|\theta_y)$ is scale reducible when both $q(\ell)$ and $q(y_v|\ell, \theta_{y,\ell})$ are scale reducible. If there are extra parts of structure was allocated in a scale reducible $q(y|\theta_y)$, $\max_{\Theta,h} H_f(\mathcal{X}_N, \Theta, h, \mathbf{k}, \Xi)$ will drive those extra parameters towards zero. i.e., automatic model selection on $\mathbf{k}_Y$ is made during parameter learning on Ξ^*, Θ^* by eq.(6). Taking $H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi)$ in eq.(26) as an example, maximizing the term $\sum_t \sum_\ell p(\ell|x_t, \theta_{y|x}) \ln \alpha_\ell$ of $H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi)$ becomes equivalently

maximizing $N\sum_\ell \alpha_\ell \ln \alpha_\ell$ with $\alpha_\ell = \sum_t p(\ell|x_t, \theta_{y|x})/N$, which tends to push α_ℓ towards zero when it is extra.

The last, we consider the details of $q(\Theta|\Xi)$ in Tab.2. The simplest case is the choice 1, i.e., ignoring its role simply by letting $Z(\Theta) = -\ln q(\Theta) = 0$. Moreover, we may further divided each Θ_ξ into independent groups $q(\Theta_\xi) = \prod_j q(\Theta_\xi^{(j)})$, and let $q(\Theta_\xi^{(j)})$ in the following choices [25, 21, 32]:

(1) The simplest way is to consider the so called improper priori, e.g., a uniform improper priori

$$q(\Theta_\xi^{(j)}) \propto constant, \tag{28}$$

for real a parameter. For a scalar $\gamma > 0$, we consider $q(\gamma)$ by a Jeffrey improper priori $q(\gamma) \propto 1/\gamma$. When Σ is a $d \times d$ covariance matrix, a Jeffreys improper priori is $q(\Sigma) \propto 1/|\Sigma|^{0.5(d+1)}$.

(2) For different parameters, we consider different specific distribution. For examples, a gamma distribution for a scalar $\gamma > 0$, an inverted Wishart distribution for Σ of a $d \times d$ covariance matrix, a Beta/Dirichlet distribution for the proportional parameters $\{\alpha_\ell\}$, and a uniform improper priori by eq.(28) or a Gaussian distribution for real parameters in a vector or matrix.

(3) Another general way for getting an improper priori is the choice 3 in Tab.2, which was developed in [46, 47, 51]. Here we explain its rationale. Given a density $p(u|\theta)$, we have $\int p(u|\theta)du = 1$ that does not depend on θ for an infinite size of samples. This is no longer true for $s = \sum_{t=1}^N p(u_t|\theta)$ by considering a finite sample size of samples $\{u_t\}_{t=1}^N$. This s actually varies with θ and imposes an implicit distribution θ. Considering a priori $q(\theta) \propto \frac{1}{s}$ can balance off this unnecessary bias. E.g., for h in eq.(19) we have

$$q(h) \propto [\sum_{t=1}^N \sum_{\tau=1}^N G(x_t|x_\tau, h^2 I)/N]^{-1}. \tag{29}$$

Other examples are referred to [46, 47, 51], e.g., eqn.(11) for $q(\Theta)$ in [47].

3 Best Harmony vs Best Matching: Relations to Others

3.1 Special cases: relations to existing approaches

It is interesting to further observe how the best harmony learning degenerates as a BYY system degenerates to a conventional model $q(\mathbf{X}|\Theta)$. We consider $\mathbf{R} = \{\Theta\}$ without an inner representation part $\mathbf{Y}$, which leads us back to Fig.2(c), and simplifies $H(p\|q) = H(p\|q, \mathbf{k}, \Xi)$ in eq.(3) into

$$H(p\|q) = \int p(\Theta|\mathbf{X})p(\mathbf{X}) \ln [q(\mathbf{X}|\Theta)q(\Theta)]d\mathbf{X}d\Theta. \tag{30}$$

For a $p(\Theta|\mathbf{X})$ free of structure and $p(\mathbf{X})$ of the choice (A) in Tab.2, maximizing $H(p\|q)$ with respect to $p(\Theta|\mathbf{X})$ leads to the MB type Bayesian learning in Tab.1, i.e., $\max_\Theta \ln [q(\mathcal{X}_N|\Theta)q(\Theta)]$, while $J(\mathbf{k})$ in eq.(7) becomes

$$\mathbf{k}^* = arg\min_{\mathbf{k}} J(\mathbf{k}), J(\mathbf{k}) = -\max_{\Theta} \ln [q(\mathcal{X}_N|\Theta)q(\Theta)] + 0.5d_{\mathbf{k}}, \tag{31}$$

which is a Bayesian learning based extension of AIC. For a non-informative $q(\Theta)$, it further degenerates to exactly AIC [1, 2]. Moreover, for a general case with $p(\mathbf{X}, h)$ by eq.(19), it follows from eq.(22) that eq.(30) is extended into

$$\begin{aligned} &H(p\|q) = \int p(h)p(\Theta|\mathcal{X}_N)H_h(p\|q,\Theta)d\Theta \approx \max_{\Theta,h} H_h(p\|q,\Theta) - 0.5d_{\mathbf{k}}, \\ &H_h(p\|q,\Theta) = \ln\left[q(\mathcal{X}_N|\Theta)q(\Theta)\right] + 0.5h^2 Tr[\Sigma(\mathcal{X}_N)] - Z(h), \\ &Z(h) = -\ln q(h|\mathcal{X}_N), \Sigma(\mathbf{X}) = \partial^2 \ln q(\mathbf{X}|\Theta)/\partial\mathbf{X}\partial\mathbf{X}^T. \end{aligned} \tag{32}$$

With $p(\Theta|\mathbf{X})$ in a given structure, the BYY harmony learning is different from the conventional Bayesian learning. E.g., we consider $p(\Theta|\mathbf{X})$ with the BI structure in Tab.1 and rewrite eq.(30) into

$$H(p\|q) = \int p(\Theta|\mathbf{X})p(\mathbf{X}) \ln p(\Theta|\mathbf{X})d\mathbf{X}\Theta + \int p(\mathbf{X}) \ln q(\mathcal{X}|S)d\mathbf{X}. \tag{33}$$

Particularly, for $p(\mathbf{X})$ of the choice (A) in Tab.2, it further becomes

$$H(p\|q) = \int p(\Theta|\mathcal{X}_N) \ln p(\Theta|\mathcal{X}_N)d\Theta + \ln q(\mathcal{X}_N|S). \tag{34}$$

The maximization of its second term is exactly the MI (marginal likelihood) choice in Tab.1. As already discussed in Section 1, it has been previously studied under various names [34, 23, 21, 22]. The first term in eq.(34) is the negative entropy of $p(\Theta|\mathcal{X}_N)$ and its maximization is seeking an inverse inference $\mathcal{X}_N \to \Theta$ with a least uncertainty. More generally, it follows from eq.(3) that we get an extension of eq.(34) as follows:

$$\begin{aligned} &H(p\|q) = \int p(\mathbf{R}|\mathcal{X}_N) \ln p(\mathbf{R}|\mathcal{X}_N)d\mathbf{R} + \ln q(\mathcal{X}_N|S), \\ &p(\mathbf{R}|\mathcal{X}) = q(\mathcal{X}|\mathbf{R})q(\mathbf{R})/q(\mathcal{X}|S), \quad q(\mathcal{X}|S) = \int q(\mathcal{X}|\mathbf{R})q(\mathbf{R})d\mathbf{R}. \end{aligned} \tag{35}$$

Even generally, we also let S to be included in the inner representation $\mathbf{R}$, and get a further generalization of eq.(30) as follows:

$$H(p\|q) = \sum_S p(S|\mathbf{X})p(\mathbf{X}) \ln\left[q(\mathbf{X}|S)q(S)\right]. \tag{36}$$

When $p(S|\mathbf{X})$ is free of structure, maximizing $H(p\|q)$ with respect to $p(S|\mathbf{X})$ leads to $\max_S \ln\left[q(\mathcal{X}_N|S)q(S)\right]$ for model selection, i.e., the BI choice in Tab.1. In the special case that $q(S)$ is equal for each candidate S, it further degenerates to $\max_S \ln q(\mathcal{X}_N|S)$, i.e., the ML choice in Tab.1. Also, a generalized counterpart of eq.(34) becomes

$$H(p\|q) = \sum_S p(S|\mathcal{X}_N) \ln p(S|\mathcal{X}_N) + \ln q(\mathcal{X}_N), \quad q(\mathcal{X}_N) = \sum_S q(\mathcal{X}_N|S)q(S).$$

3.2 Best Harmony versus Best Matching

For a BYY system, in addition to making the best harmony learning by eq.(3), an alternative has also been proposed and studied in [64] under the name of

Bayesian Kullback Ying Yang (BKYY) learning that performs the following *best matching* principle:

$$\min KL(p\|q),\ KL(p\|q) = \int p(\mathbf{R}|\mathbf{X})p(\mathbf{X}) \ln \frac{p(\mathbf{R}|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X}|\mathbf{R})q(\mathbf{R})} d\mathbf{X}d\mathbf{R} \tag{37}$$
$$= \int p(\Theta|\mathbf{X})\{\int p(\mathbf{Y}|\mathbf{X},\theta_{y|x})p(\mathbf{X}) \ln \frac{p(\Theta|\mathbf{X})p(\mathbf{Y}|\mathbf{X},\theta_{y|x})p(\mathbf{X})}{q(\mathbf{X}|\mathbf{Y},\theta_{x|y})q(\mathbf{Y}|\theta_y)q(\Theta)} d\mathbf{X}d\mathbf{Y}\}d\Theta,$$

which reaches to the best matching $KL(p\|q) = 0$ at $p(\mathbf{R}|\mathbf{X})p(\mathbf{X}) = q(\mathbf{X}|\mathbf{R})q(\mathbf{R})$.

As a BYY system degenerates to a conventional model $q(\mathbf{X}|\Theta)$, the above eq.(37) is simplified into the following counterpart of eq.(30):

$$\min KL(p\|q),\ KL(p\|q) = \int p(\Theta|\mathbf{X})p(\mathbf{X}) \ln \frac{p(\Theta|\mathbf{X})p(\mathbf{X})}{q(\mathbf{X}|\Theta)q(\Theta)} d\mathbf{X}d\Theta. \tag{38}$$

When $p(\Theta|\mathbf{X})$ is free of structure, minimizing $KL(p\|q)$ with respect to $p(\Theta|\mathbf{X})$ leads to $p(\Theta|\mathbf{X}) = q(\mathbf{X}|\Theta)q(\Theta)/q(\mathcal{X}|S)$ and $q(\mathcal{X}|S) = \int q(\mathbf{X}|\Theta)q(\Theta)\mu(d\Theta)$. As a result, eq.(38) becomes

$$\min KL(p\|q),\ KL(p\|q) = \int p(\mathbf{X}) \ln [p(\mathbf{X})/q(\mathcal{X}|S)] d\mathbf{X}, \tag{39}$$

where $p(\mathbf{X})$ is an input irrelevant to $q(\mathcal{X}|S)$ and $q(\Theta)$. For $p(\mathbf{X})$ of the choice (A) in Tab.2, eq.(39) further becomes equivalent to the MI (marginal likelihood) choice in Tab.1. For a general case with $p(\mathbf{X}, h)$ by eq.(19), eq.(39) provides its data smoothing version with not only Θ but also h learned.

Alternatively we may also consider $\min_{q(\Theta)} KL(p\|q)$ when $q(\Theta)$ is free of constraint, which leads to $q(\Theta) = p(\Theta|\mathbf{X})$ and

$$\min KL(p\|q),\ KL(p\|q) = \int p(\Theta|\mathbf{X})p(\mathbf{X}) \ln [p(\mathbf{X})/q(\mathbf{X}|\Theta)] d\mathbf{X}d\Theta. \tag{40}$$

When $p(\mathbf{X})$ is an input irrelevant to $q(\mathbf{X}|\Theta)$, it is equivalent to

$$\max \int p(\Theta|\mathbf{X})p(\mathbf{X}) \ln q(\mathbf{X}|\Theta) d\mathbf{X}d\Theta, \tag{41}$$

which further becomes $\max \int p(\Theta|\mathcal{X}_N) \ln q(\mathcal{X}_N|\Theta)d\Theta$ for $p(\mathbf{X})$ of the choice (A) in Tab.2. Its maximization with a structural free $p(\Theta|\mathbf{X})$ leads to the classical ML learning again. Moreover, in help of eq.(10), we are again lead to eq.(31), i.e., the ML learning based AIC [1, 2].

Next, we return to eq.(37) with its inner representation $\mathbf{Y}$ in consideration. When $p(\mathbf{Y}|\mathbf{X},\theta_{y|x})$ is free, $\min_{p(\mathbf{Y}|\mathbf{X},\theta_{y|x})} KL(p\|q)$ leads to eq.(38) again with

$$\begin{aligned} &p(\mathbf{Y}|\mathbf{X},\theta_{y|x}) = q(\mathbf{X}|\mathbf{Y},\theta_{x|y})q(\mathbf{Y}|\theta_y)/q(\mathbf{X}|\Theta), \\ &q(\mathbf{X}|\Theta) = \int q(\mathbf{X}|\mathbf{Y},\theta_{x|y})q(\mathbf{Y}|\theta_y)d\mathbf{Y}. \end{aligned} \tag{42}$$

In other words, we can integrate over the effect of inner representation $\mathbf{Y}$ to get $q(\mathbf{X}|\Theta)$ and then handle it by eq.(38).

On the other hand, $\min_{q(\Theta)} KL(p\|q)$ with a free $q(\Theta)$ results in $q(\Theta) = p(\Theta|\mathbf{X})$ and also

$$\min KL(p\|q) = \int p(\Theta|\mathbf{X}) KL(p\|q,\Theta)d\Theta \geq \min_f KL(p\|q,\Theta).$$

$$KL(p\|q,\Theta)=\int p(\mathbf{Y}|\mathbf{X},\theta_{y|x})p(\mathbf{X})\ln\frac{p(\mathbf{Y}|\mathbf{X},\theta_{y|x})p(\mathbf{X})}{q(\mathbf{X}|\mathbf{Y},\theta_{x|y})q(\mathbf{Y}|\theta_y)}d\mathbf{X}d\mathbf{Y}. \quad (43)$$

This $\min_f KL(p\|q,\Theta)$ was originally proposed in 1995 under the name Bayesian Kullback Ying Yang (BKYY) learning [64]. From $\min_{p(\mathbf{Y}|\mathbf{X},\theta_{y|x})} KL(p\|q,\Theta)$, we are lead to the above discussed eq.(42) and eq.(41) again.

The difference between the best Ying Yang matching by eq.(37) and the best Ying Yang harmony learning by eq.(3) can be better understood from the following relation:

$$KL(p\|q)=\int p(\mathbf{R}|\mathbf{X})p(\mathbf{X})\ln[p(\mathbf{R}|\mathbf{X})p(\mathbf{X})]d\mathbf{X}d\mathbf{R}-H(p\|q). \quad (44)$$

In addition to maximizing $H(p\|q)$, minimizing $KL(p\|q)$ also includes minimizing the first term that is the negative entropy of the Yang representation, which cancels out the least complexity nature that was discussed after eq.(3). Therefore, the best Ying Yang harmony learning by eq.(3) considerably outperforms the best Ying Yang matching learning by eq.(37) for learning tasks that need model selection, as already verified by a number of experimental comparisons and applications [35].

3.3 BKYY learning, Helmholtz machine, and variational approach

Aiming at avoiding the integral $q(\mathbf{X}|\Theta)=\int q(\mathbf{X}|\mathbf{Y},\theta_{x|y})q(\mathbf{Y}|\theta_y)d\mathbf{Y}$ for the ML learning, maximizing the likelihood function is suggested to be replaced by maximizing one of its lower bound via the *Helmholtz free energy or called variational free energy* [11, 24], by which $\max_\Theta q(\mathbf{X}|\Theta)$ is replaced by maximizing

$$\begin{aligned}
&F=-\int p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})\ln[p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})/q(\mathcal{X}_N|\mathbf{Y},\Theta)q(\mathbf{Y}|\theta_y)]d\mathbf{Y}\\
&=-\int p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})\ln\frac{p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})}{q(\mathbf{Y}|\mathcal{X}_N,\Theta)}d\mathbf{Y}+\ln q(\mathcal{X}_N|\Theta)\le\ln q(\mathcal{X}_N|\Theta),\\
&q(\mathbf{Y}|\mathcal{X}_N,\Theta)=q(\mathcal{X}_N|\mathbf{Y},\theta_{x|y})q(\mathbf{Y}|\theta_y)/q(\mathcal{X}_N|\Theta). \quad (45)
\end{aligned}$$

Instead of computing $q(\mathcal{X}_N|\Theta)$ and $q(\mathbf{Y}|\mathcal{X}_N,\Theta)$, a parametric model is considered for $p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})$, and learning is made for determining the unknown parameters $\theta_{y|x}$ together with Θ via maximizing F.

In fact, maximizing F by eq.(45) is equivalent to $\min_f KL(p\|q,\Theta)$ by eq.(43) with $p(\mathbf{X})$ in the choice (A) of Tab.2. In other words, the two approaches coincide in this situation, though they were motivated from two different perspectives. Maximizing F by eq.(45) directly aims at approximating the ML learning on $q(\mathcal{X}_N|\Theta)$, with an approximation gap to trade off computational efficiency via a pre-specified parametric $p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})$. This gap disappears if $p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})$ is able to reach the posteriori $q(\mathbf{Y}|\mathcal{X}_N,\Theta)$. Instead, minimizing $KL(p\|q,\Theta)$ by eq.(43) is not motivated from a purpose of approximating the ML learning though it was also shown in [64] that $\min_{p(\mathbf{Y}|\mathbf{X},\theta_{y|x})} KL(p\|q,\Theta)$ for a $p(\mathbf{Y}|\mathbf{X},\theta_{y|x})$ free of constraint makes $\min_f KL(p\|q,\Theta)$ become the ML learning with $p(\mathbf{X})$ in the choice (A) of Tab.2. The motivation is determining all the unknowns in the Ying-Yang pair to make the pair best matched. Beyond

becoming equivalent to the ML learning and approximating the ML learning, studies on $\min_f KL(p\|q,\Theta)$ by eq.(43) covers not only extensions to the general case with $p(\mathbf{X},h)$ by eq.(19), but also the problems of minimizing $KL(p\|q,\Theta)$ with respect to a free $q(\mathbf{X}|\mathbf{Y},\theta_{x|y})$, which leads to

$$\min \int p(\mathbf{Y}|\Theta_p) \ln \frac{p(\mathbf{Y}|\Theta_p)}{q(\mathbf{Y}|\theta_y)} \mu(d\mathbf{Y}),\ p(\mathbf{Y}|\Theta_p) = \int p(\mathbf{Y}|\mathbf{X},\theta_{y|x}) p(\mathbf{X}) \mu(d\mathbf{X}). \quad (46)$$

If $q(\mathbf{Y}|\theta_y)$ is independent among its components and $p(\mathbf{Y}|\mathbf{X},\theta_{y|x})$ has a post-linear structure, eq.(46) becomes equivalent to the minimum mutual information (MMI) base ICA learning [3]. The details are referred to [64, 56, 51, 52].

In the past decade, extensive studies have also been made under the name of variational approximation methods [20, 19], which further put the basic idea of the *Helmholtz free energy* [11, 24] in a general framework of approximation methods rooting from techniques in the calculus of variations and with a wide variety of uses [33]. The key idea is turning a complex problem into a simpler one, featured by a decoupling of the degrees of freedom in the original problem. This decoupling is achieved via an expansion of the problem to include additional parameters (called variational parameters), in help of convex duality [31]. The variational approximation method revisits the *Helmholtz free energy* approach under the formulation of probability theory, in a sense that $p(\mathbf{Y}|\mathcal{X}_N,\theta_{y|x})$ is used as an additional parameter to turn the problem of the integral $q(\mathbf{X}|\Theta) = \int q(\mathbf{X}|\mathbf{Y},\theta_{x|y}) q(\mathbf{Y}|\theta_y) d\mathbf{Y}$ into eq.(45).

3.4 A relationship map

A summary of the BYY learning related approaches is provided in Fig.6 under the principles of best harmony versus best matching, as well as their relations to typical learning approaches.

The common part of all the approaches is the shadowed center area, featured by using a probabilistic model to best match a data set $\mathcal{X}_N$ via determining three levels of its unknowns. The first two levels are the ML learning for unknown parameter learning and model selection shown in the ML row of Tab.1, which has been widely studied from various perspective as previously discussed in Sec. 1 [34, 23, 21, 22]. The third level is evaluating or selecting an appropriate meta structure $\aleph$ via $q(\mathcal{X}_N|\aleph)$, i.e., the second term in eq.(37), for which few studies have been made yet but deserve to explore.

Outbound from this shadowed center we have two directions. One is to the left-side. Priori probabilities are taken in consideration for determining three levels of its unknowns. The first two levels are the MB choices for parameter learning and model selection in Tab.1. As discussed in Sec. 1, studies have made under the name of Bayesian learning or Bayesian approach [25, 32], as well as MML [42]. The third level is again evaluating an appropriate meta structure $\aleph$ via $q(\mathcal{X}_N|\aleph)q(\aleph)$ with a priori $q(\aleph)$ in consideration. Moving forward even left, we are lead to those areas of the best Ying Yang harmony learning by eq.(3), which includes but goes beyond the areas of the ML and MB approaches, as already discussed in Sec.3.1.

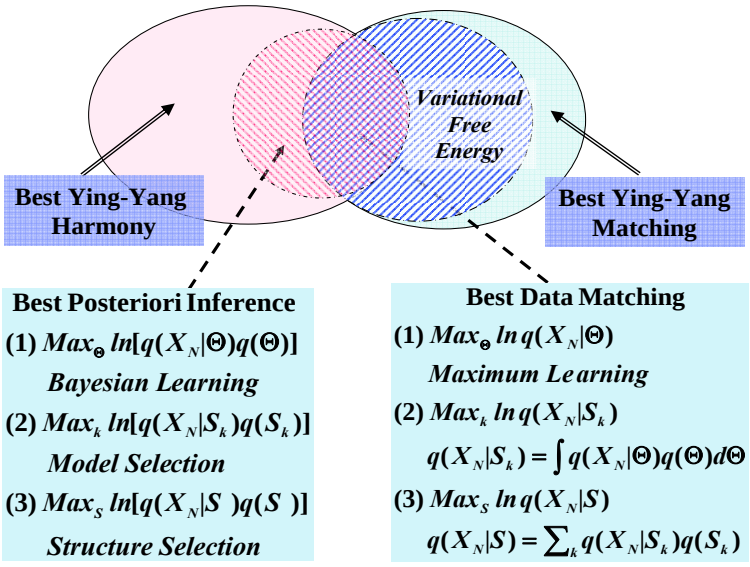

Fig. 6. Best harmony, Best matching, and Typical learning approaches

The second direction goes the right-side, the domain of the best Ying Yang matching by eq.(37). Out of the shadowed center, we enter the common area shared with the approach of *variational free energy* or the Helmholtz machine [11, 24]. Moving right still, we proceed beyond and lead to a number of other cases, as already discussed in Sec.3.2.

In addition to the mathematical relation by eq.(44), the difference between the best Ying Yang matching and the best Ying Yang harmony can also be understood from a best information transfer perspective and a projection geometry perspective. The details are referred to Section II(C) and Section III of [51], respectively. Other discussions on relations and differences are further referred to several recent papers [47, 46, 49].

4 Gaussian Manifold Based Systems, Typical Applications, and Concluding Remarks

One common structural feature shared by those structures in Tab.3 & Tab.4 is that each of them actually describes samples in the space of x via a number of Gaussian manifolds in certain organization. Shown in Tab.6 are three examples of mixtures of Gaussian manifolds, by combining $q(x|y) = q(x|y, \theta_{x|y}) = G(x|A_\ell y + \mu_\ell, \Sigma_\ell)$ with three different types of $q(y) = q(y|\theta_y)$. Taking the LFA case as an example, it follows from eq.(11), eq.(12), and eq.(13) that we have

$$H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi) = \sum_t H_f^{(t)}(\mathcal{X}_N, \Theta, h, \mathbf{k}, \Xi) + \ln q(\Theta|\Xi) + \ln q(h|\mathcal{X}_N),$$

$$H_f^{(t)}(\mathcal{X}_N, \Theta, h, \mathbf{k}, \Xi) = \sum_\ell p(\ell|x_t) \ln [G(x|A_\ell y(x_t) + \mu_\ell, \Sigma_\ell) G(y(x_t)|0, \Lambda_\ell)\alpha_\ell]$$
$$-0.5 \sum_\ell p(\ell|x_t)\{m_\ell + [y(x_t) - y*(x_t)]^T [\Lambda_\ell^{-1} + A_\ell^T \Sigma_\ell^{-1} A_\ell][y(x_t) - y*(x_t)]\},$$
$$-0.5h^2 \sum_\ell p(\ell|x_t) Tr[\Sigma_\ell^{-1}],\ \Theta = \{A_\ell^T A_\ell = I, \mu_\ell, \Sigma_\ell, \Lambda_\ell, \alpha_\ell, W_\ell, w_\ell, b_\ell, c_\ell\}_{\ell=1}^k,$$
$$y(x) = W_\ell x + w_\ell, \quad y^*(x) = [\Lambda_\ell^{-1} + A_\ell^T \Sigma_\ell^{-1} A_\ell]^{-1} A_\ell^T \Sigma_\ell^{-1}(x - \mu_\ell), \tag{47}$$
$$p(\ell|x_t, \theta_{y|x}) = \frac{e^{-o_\ell(x_t)}}{\sum_{j=1}^k e^{-o_j(x_t)}}, \quad o_\ell(x) = \beta_\ell x^T [\Sigma_\ell + A_\ell \Lambda_\ell A_\ell^T]^{-1} x + b_\ell^T x + c_\ell.$$

where $p(\ell|x_t, \theta_{y|x})$ is given by eq.(27) with $H_\ell = \Sigma_\ell + A_\ell \Lambda_\ell A_\ell^T$, and $q(h|\mathcal{X}_N)$ is given by eq.(29). From the above $H_f(\mathcal{X}_N, \Theta, \mathbf{k}, \Xi)$, we can develop a gradient based adaptive algorithm to implement learning by eq.(6). Moreover, we can also get $J(\mathbf{k})$ for Stage II in eq.(7)(e.g., see eqn.(86) in [46]).

$q(y\|\theta_y)$	$q(x \mid y, l, \theta_{x\|y,l}) = G(x \mid A_l y + \mu_l, \Sigma_l)$
$q(l) = \alpha_l \geq 0$	At $A_\ell = 0$ we get GM $q(x \mid \theta) = \sum_{\ell=1}^k \alpha_\ell G(x \mid \mu_\ell, \Sigma_\ell)$
$q(y) = \prod_{j=1}^m q_j^{y^{(j)}} (1-q_j)^{1-y^{(j)}}$	At k=1 we get BFA $q(x\|\theta) = \sum_y G(x\|Ay+\mu, \Sigma) \prod_{j=1}^m q_j^{y^{(j)}} (1-q_j)^{1-y^{(j)}}$
$q(y, l) = \alpha_l \prod_{j=1}^{m_l} G(y^{(j)}\|0, \lambda_l^{(j)})$	we get LFA $q(x \mid \theta) = \sum_{\ell=1}^k \alpha_\ell \int G(x\|A_l \vec{y} + \mu_l, \Sigma_l) \prod_{j=1}^{m_l} G(y^{(j)}\|0, \lambda_l^{(j)}) dy$

Table 6 Gaussian mixture (GM), Binary factor analysis (BFA), and Local factor analysis (LFA)

Using Gaussian manifold based systems, computational feasibility is guaranteed by the fact that the integral over y is analytically solvable, scalability of complicated problems is obtained via increasing the number of Gaussian manifolds in consideration, and coverage of typical learning tasks is achieved via different ways that organize Gaussian manifolds. Moreover, the scale $\mathbf{k}_Y$ of Gaussian manifold based systems is simply featured by the variance of a Gaussian variable in Y and the probability of a discrete variable in Y, which ensures scale reducibility of $q(\mathbf{Y}|\theta_y)$. Due to these natures, Gaussian manifold based systems have been applied to various tasks. Several examples are listed below:

- Cluster analysis, Gaussian mixture, and mixture of shape-structures (including lines, planes, curves, surfaces, and even complicated shapes) [63, 60, 57, 56, 55, 46].
- Factor analysis (FA) and local FA, including PCA, subspace analysis and local subspaces, etc [57, 36, 35, 18, 17].
- Independent subspace analysis, including independence components analysis (ICA), binary factor analysis (BFA), nonGaussian factor analysis (NFA), and LMSER, as well as three layer net [56, 54, 51, 53, 4].

- Independent state space analysis, including temporal factor analysis (TFA), independent hidden Markov model (HMM), temporal LMSER, and variants [61, 59, 56, 50].
- Combination of multiple inference, including multiple classifier combination, RBF nets, mixture of experts, etc [62, 60, 46].

Proposed firstly in 1995 [64], the BYY harmony learning has been systematically developed in the past decade. Studies have demonstrated the feasibility of using BYY system as a general framework for unifying a number of typical learning models and a promising direction of adopting best Ying-Yang harmony as a general theory for parameter learning and model selection. The BYY harmony learning leads to not only a criterion that outperforms existing typical model selection criteria in a two-phase implementation, but also automatic model selection during parameter learning with computing cost saved significantly. Readers are referred to [47, 46, 49] for a tutorial and recent systematic overviews and also to some earlier papers for a similar purpose [58, 51, 52]. Moreover, readers are referred to [61, 59, 50, 56] for the studies on the BYY harmony learning with temporal dependences taken in consideration.

Acknowledgement The work described in this paper was fully supported by a Hong Kon RGC grant Project No: CUHK4173/06E).

References

1. Akaike, H (1974), "A new look at the statistical model identification", *IEEE Tr. Automatic Control 19*, 714-723.
2. Akaike, H (1981), "Likelihood of a model and information criteria", *Journal of Econometrics 16*, 3-14.
3. Amari, S, Cichocki, A, & Yang, H (1996), "A new learning algorithm for blind signal separation", *Advances in NIPS 8* , MIT Press, 757-763.
4. An, YJ, et al, (2006), "A Comparative Investigation on Model Selection in Independent Factor Analysis" *J. Mathematical Modelling and Algorithms 5*, 447-473.
5. Barndorff-Nielson, OE (1978), *Methods of Information and Exponential Families*, Wiley, Chichester.
6. Bourlard, H & Kamp, Y (1988), "Auto-association by multilayer Perceptrons and singular value decomposition", Biological Cybernetics 59, 291-294.
7. Bozdogan, H (1987) "Model Selection and Akaike's Information Criterion: The general theory and its analytical extension", *Psychometrika 52*, 345-370.
8. Bozdogan, H & Ramirez, DE (1988), "FACAIC: Model selection algorithm for the orthogonal factor model using AIC and FACAIC", *Psychometrika 53(3)* , 407-415.
9. Brown, L (1986), *Fundamentals of Statistical Exponential Families*, Institute of Mathematical Statistics, Hayward, CA.
10. Cavanaugh, JE (1997), "Unifying the derivations for the Akaike and corrected Akaike information criteria", *Statistics & Probability Letters 33*, 201-208.
11. Dayan, P & Hinton, GE, Neal, RM, & Zemel, RS (1995), "The Helmholtz machine", *Neural Computation 7:5*, 889-904.
12. Gilks, WR, Richardson, S, & Spiegelhakter, DJ (1996), *Markov Chain Monte carlo in Practice*, London: Chapman and Hall.

13. Girosi, F et al, (1995) "Regularization theory and neural architectures", *Neural Computation 7*, 219-269.
14. Grossberg, S (1976), Adaptive patten classification and universal recording: I&II. *Biological Cybernetics 23*, 121-134 and 187-202.
15. Hinton, GE & Zemel, RS (1994), "Autoencoders, minimum description length and Helmholtz free energy", *Advances in NIPS 6*, 3-10.
16. Hinton, GE, Dayan, P, Frey, BJ, & Neal, RN (1995), "The wake-sleep algorithm for unsupervised learning neural networks", *Science 268*, 1158-1160.
17. Hu, XL & Xu, L (2006), "A comparative study on selection of cluster number and local subspace dimension in the mixture PCA models", *Lecture Notes in Computer Sciences, Vol. 3971*, 1214 - 1221, Springer Verlag.
18. Hu, XL & Xu, L (2004) "A comparative investigation on subspace dimension determination", *Neural Networks 17*, 1051–1059.
19. Jaakkola, TS (2001), "Tutoiral on variational approximation methods", in Opper & Saad, eds, *Advanced Mean Field methods: Theory & Pratice*, 129-160, MIT press.
20. Jordan, M, Ghahramani, Z, Jaakkola, T, & Saul, L (1999), "Introduction to variational methods for graphical models", *Machine Learning 37*, 183-233.
21. Kass, RE & Raftery, AE (1995), "Bayes Factors", *Journal of the American Statistical Association 90*, 773-795.
22. MacKay, DJC, (2003), *Information Theory, Inference, and Learning Algorithms*, Cambridge University Press.
23. Neath, AA & Cavanaugh, JE (1997), "Regression and time series model selection using variants of the Schwarz information criterion", *Communications in Statistics A 26*, 559-580.
24. Neal, R & Hinton, GE (1999), "A view of the EM algorithm that justifies incremental, sparse, and other variants", In M. I. Jordan (Ed.), *Learning in graphical models*, Cambridge, MA: MIT Press, pp355-368.
25. Press, SJ (1989), *Bayesian statistics: principles, models, and applications*, Factors", John Wiley & Sons, Inc., 1989.
26. Poggio, T & Girosi, F (1990), "Networks for approximation and learning", *Proc. of IEEE 78*, 1481-1497.
27. Redner, RA & Walker, HF (1984), "Mixture densities, maximum likelihood, and the EM algorithm", *SIAM Review 26*, 195-239.
28. Rissanen, J (1986), "Stochastic complexity and modeling", *Annals of Statistics 14(3)*, 1080-1100.
29. Rissanen, J (1989), *Stochastic Complexity in Statistical Inquiry*, World Scientific: Singapore.
30. Rivals, I & Personnaz, L (1999) "On Cross Validation for Model Selection", *Neural Computation 11*, 863-870.
31. Rockafellar, R (1972), *Convex Analysis*, Princeton University Press.
32. Ruanaidh O, & Joseph, JK (1996), *Numerical Bayesian methods applied to signal processing*, Springer-Verlag, New York, Inc., 1996.
33. Rustagi, J (1976), *Variational Method in Statistics*, New York: Academic Press.
34. Schwarz, G (1978), "Estimating the dimension of a model", *Annals of Statistics 6*, 461-464.
35. Shi, L (2008), Bayesian Ying-Yang harmony learning for local factor analysis: a comparative investigation, *Oppositional Concepts in Computational Intelligence*, Tizhoosh & Ventresca, eds, Springer-Verlag (Studies in CI), in press.
36. Shi, L & Xu, L (2006), "Local factor analysis with automatic model selection: a comparative study and digits recognition application", *Artificial Neural Networks - ICANN 2006: LNCA 4132*, Springer Berlin, 260-269.

37. Stone, M (1974), "Cross-validatory choice and assessment of statistical prediction", *J. Royal Statistical Society B 36*, 111-147.
38. Stone, M (1977), "An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion", *J. Royal Statistical Society B 39* (1), 44-47.
39. Stone, M (1978), "Cross-validation: A review", *Math. Operat. Statist. 9*, 127-140.
40. Tikhonov, AN & Arsenin, VY (1977), *Solutions of Ill-posed Problems*, Winston and Sons.
41. Vapnik, VN (1995), *The Nature Of Statistical Learning Theory*, Springer.
42. Wallace, CS & Boulton, DM (1968), "An information measure for classification", *Computer Journal 11*, 185-194.
43. Wallace, CS & Freeman, PR (1987), "Estimation and inference by compact coding", *J. of the Royal Statistical Society 49(3)*, 240-265.
44. Wang L & Feng, J (2005), "Learning Gaussian mixture models by structural risk minimization", *Proc. ICMLC05*, 4858-4863, 19-21 Aug. 2005, Guangzhou, China.
45. Xu, L (2008), "Machine learning problems from optimization perspective ", to appear on *Journal of Global Optimization*.
46. Xu, L (2007), "A unified perspective and new results on RHT computing, mixture based learning, and multi-learner based problem solving", *Pattern Recognition 40*, 2129-2153.
47. Xu, L (2007), "Bayesian Ying Yang learning", *Scholarpedia 2(3):1809*, http: // scholarpedia.org /article/ Bayesian Ying Yang Learning.
48. Xu, L (2007), "Rival penalized competitive learning", *Scholarpedia, 2(8):1810*, http: // scholarpedia.org /article/ Rival Penalized Competitive Learning.
49. Xu, L (2007), "A trend on regularization and model selection in statistical learning: a Bayesian Ying Yang learning perspective", In W. Duch & J. Mandziuk, eds, Challenges for Computational Intelligence, Springer-Verlag, pp365-406.
50. Xu, L (2004), "Temporal BYY encoding, Markovian state spaces, and space dimension determination", *IEEE Tr. Neural Networks 15*, 1276-1295.
51. Xu, L (2004), "Advances on BYY harmony learning: information theoretic perspective, generalized projection geometry, and independent factor auto-determination", *IEEE Tr. Neural Networks 15*, 885-902.
52. Xu, L (2004), "Bayesian Ying Yang learning : (I) & (II) ", *Intelligent Technologies for Information Analysis*, Zhong & Liu (eds), Springer, 615-706.
53. Xu, L (2004), "BI-directional BYY learning for mining structures with projected polyhedra and topological map", Invited talk, in *Proc. of FDM 2004: Foundations of Data Mining*, Lin, Smale, Poggio, & Liau (eds.), Brighton, UK, pp5-18.
54. Xu, L (2003), "BYY learning, regularized implementation, and model selection on modular networks with One hidden layer of binary units", *Neurocomputing 51*, 227-301.
55. Xu, L (2003), "Data smoothing regularization, multi-sets-learning, and problem solving strategies", *Neural Networks 15(5-6)*, 817-825.
56. Xu, L (2003), "Independent component analysis and extensions with noise and time: a Bayesian Ying-Yang learning perspective", *Neural Information Processing Letters and Reviews 1*, 1-52.
57. Xu, L (2002), "BYY harmony learning, structural RPCL, and topological self-organizing on unsupervised and supervised mixture models", *Neural Networks 15*, 1125-1151.
58. Xu, L (2002), "Bayesian Ying Yang harmony learning", *The Handbook of Brain Theory and Neural Networks*, Second edition, (MA Arbib, Ed.), Cambridge, MA: The MIT Press, pp1231-1237.

59. Xu, L (2001), "BYY harmony learning, independent state space and generalized APT financial analyses", *IEEE Tr. Neural Networks 12*, 822-849.
60. Xu, L (2001), "Best harmony, unified RPCL and automated model selection for unsupervised and supervised learning on Gaussian mixtures, ME-RBF models and three-layer nets", *Intl J. Neural Systems 11*, 3-69.
61. Xu, L (2000), "Temporal BYY learning for state space approach, hidden Markov model and blind source separation", *IEEE Tr. on Signal Processing 48*, 2132-2144.
62. Xu, L (1998), "RBF nets, mixture experts, and Bayesian Ying-Yang learning", *Neurocomputing 19(1-3)*, 223-257.
63. Xu, L (1997), "Bayesian Ying-Yang machine, clustering and number of clusters", *Pattern Recognition Letters 18(11-13)*, 1167-1178.
64. Xu, L (1995), "Bayesian-Kullback coupled YING-YANG machines: unified learnings and new results on vector quantization", *Proc. ICONIP95*, 977-988, Oct 30-Nov.3, 1995, Beijing.
65. Xu, L, Krzyzak, A & Oja, E (1992&93), "Rival penalized competitive learning for clustering analysis, RBF net and curve detection", *IEEE Tr. on Neural Networks 4*, 636-649. Its early version on *Proc. of 11th ICPR92*, Vol.I, pp.672-675.
66. Xu, L (1991&93) "Least mean square error reconstruction for self-organizing neural-nets", *Neural Networks*, 6: 627-648, 1993. Its early version on *Proc. IJCNN91'Singapore*, 2363-2373, 1991.

J.M. Zurada et al. (Eds.): WCCI 2008 Plenary/Invited Lectures, LNCS 5050, pp. 48–78, 2008.