A Confident Information First Principle for Parametric Reduction and Model Selection of Boltzmann Machines

Journal Article

A Confident Information First Principle for Parameter Reduction and Model Selection of Boltzmann Machines

Xiaozhao Zhao, Yuexian Hou, Dawei Song, and Wenjie Li

***Abstract*—Typical dimensionality reduction (DR) methods are data-oriented, focusing on directly reducing the number of random variables (or features) while retaining the maximal variations in the high-dimensional data. Targeting unsupervised situations, this paper aims to address the problem from a novel perspective and considers model-oriented dimensionality reduction in parameter spaces of binary multivariate distributions. Specifically, we propose a general parameter reduction criterion, called Confident-Information-First (CIF) principle, to maximally preserve confident parameters and rule out less confident ones. Formally, the confidence of each parameter can be assessed by its contribution to the expected Fisher information distance within a geometric manifold over the neighbourhood of the underlying real distribution. Then we demonstrate two implementations of CIF in different scenarios. First, when there are no observed samples, we revisit the Boltzmann Machines (BM) from a model selection perspective and theoretically show that both the fully visible BM (VBM) and the BM with hidden units can be derived from the general binary multivariate distribution using the CIF principle. This finding would help us uncover and formalize the essential parts of the target density that BM aims to capture and the non-essential parts that BM should discard. Second, when there exist observed samples, we apply CIF to the model selection for BM, which is in turn made adaptive to the observed samples. The sample-specific CIF is a heuristic method to decide the priority order of parameters, which can improve the search efficiency without degrading the quality of model selection results as shown in a series of density estimation experiments.**

***Index Terms*—Information Geometry, Boltzmann Machine, Parametric Reduction, Fisher Information**

I. Introduction

RECENTLY, deep learning models (e.g., Deep Belief Networks (DBN) [1], Stacked Denoising Auto-encoder [2], Deep Boltzmann Machine (DBM) [3], etc.) have drawn increasing attention due to their impressive empirical performance in various application areas, such as computer vision [4] [5] [6], natural language processing [7], information retrieval [8] [9] and other classification problems [10] [11] [12]. Despite of these practical successes, there have been debates on the fundamental principle that governs the design and training of those deep architectures. In most situations, searching the parameter space for deep learning models is difficult. To tackle this difficulty, *unsupervised pre-training* has been introduced as an important process. In [13], it has been empirically shown that the unsupervised pre-training could fit the network parameters in a region of the parameter space that could well capture the data distribution, thus alleviating generalization error of the trained deep architectures.

From the density estimation point of view, the unsupervised learning can be interpreted as an attempt to discover a set of parameters for a generative model that describes the underlying distribution of the observed data. In real-world applications, the datasets are often high-dimensional and we would need a model with high-dimensional parameter space in order to effectively depict the underlying distribution. However, when the model becomes excessively complex, the model would easily overfit the limited training data. This leads to the issue of *model selection, i.e., selecting a subset of parameters in an attempt to create a model of optimal complexity for the given data* (a comprehensive review of model selection can be found in [14] and [15]).

Targeting unsupervised situations, this paper aims to address this issue by considering model-oriented dimensionality reduction in parameter spaces of binary multivariate distributions. Since Boltzmann machines (BM) are fundamental building blocks for a number of widely used deep architectures (e.g., DBN and DBM), we will focus on a formal analysis of the essential parts of the target density (a multivariate binary distribution) that the BM aims to capture in term of model selection.

Here, finding the essential parts for the underlying distribution is the main objective of model selection, which has two implications: (1) the choice of the number of free parameters and (2) the choice of parameters given that number. The former has been studied by classical model selection approaches (such as Akaike information criterion (AIC) [16] and Bayesian information criterion (BIC) [17]), where the number of free parameters is adopted as a model complexity measure. However, the model complexity is a complicated quantity which should take into account the effectiveness of the selected free parameters (e.g., as illustrated in Section 3.5.3 of Bishop's book [18] and the curvature measure [19]). Therefore, the latter problem (i.e., choice of parameters given a model complexity) is also of great significance for the model selection purpose and should be thoroughly studied.

X. Zhao, Y. Hou, D. Song are with School of Computer Science and Technology, Tianjin University, Tianjin, 300072, China. e-mail: (0.25eye@gmail.com; yxhou@tju.edu.cn; dawei.song2010@gmail.com). D. Song is also with the School of Computing and Communications, The Open University, Milton Keynes, UK. W. Li is with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China. e-mail: (cswjli@comp.polyu.edu.hk). This work is funded in part by the Chinese 863 Program (grant No. 2015AA015403), the Key Project of Tianjin Natural Science Foundation (grant No. 15JCZDJC31100), the Tianjin Younger Natural Science Foundation (Grant no: 14JCQNJC00400) and the Major Project of Chinese National Social Science Fund (grant No. 14ZDB153). Authors of Correspondence: Y. Hou, D. Song and W. Li

Various model selection procedures have been proposed in the literature. In general, the model selection problem is tackled by the combination of a model selection criterion that allows comparison of alternative models and a search strategy that allows us to find the optimal model according to that criterion. Typical model selection criteria include: P-value [20] that is based on the hypothesis test of the likelihood ratio [21] between two nested models; the information-theoretic approaches that estimate the expected information distance based on the empirical log-likelihood function when a certain model is used [14]. For the latter, the expected information distance can be estimated by different techniques, such as bootstrap methods using bootstrap samples [22], cross-validation methods that average the log-likelihoods on validation samples [23], and asymptotic methods aiming to find an (asymptotically) unbiased estimator for the expected information distance that takes into account the log-likelihood of the data under a candidate model and a penalty related to the effective number of parameters in the model (well-known examples are the AIC that penalizes the number of parameters at the rate of $O(1)$ and BIC that penalizes at the rate of $O(\log N)$, N is the number of samples). In addition to the above model selection criteria, the L_1 regularization methods can also be used for the model selection purpose by jointly minimizing the empirical error (or negative log-likelihood) and penalty [24] [25], where the parameters with significant absolute values after training are selected to create the optimal model.

As described earlier, the model selection problem in high-dimensional parameter space can be divided into *the choice of the number of free parameters* and *the choice of parameters given that number*. The classical model selection criteria can be seen as the integration of the two, formalized by a single optimization objective[1]. However, the existing model selection approaches are insufficient when the model's parameters do not have a priority order, which is the case for BM.

First, the lack of an explicit priority order of parameters would lead to serious efficiency issues. Assuming there exists a universal parametric probabilistic model S (with K free parameters) that is general enough to represent all system phenomena. Consequently, there exist (2^K) candidate sub-models, corresponding to different choices of parameters. When K is large, it is computationally intractable to exhaustively test all sub-models in classical model selection methods. By using a stepwise greedy search strategy [26], the number of tested sub-models can be reduced to $O(K^2)$, which is often time-consuming in practice. Although L_1 regularization could avoid the heavy combinatorial search, the regularization coefficient needs to be determined through cross validation, which may also be time consuming in practice.

Second, for BM, the lack of an explicit priority order of parameters would lead to a comparison among non-nested models. For P-value, since the likelihood ratio tests exist only for nested models, the tests of hypotheses within a dataset are not independent (called multiple testing problem). As a result, it is difficult to make inferences [26]. For information criteria, there exist considerable controversies about their applicability for non-nested models. In [26], the authors argued that the information criteria make no assumptions about nested candidate models and the order in which the information criterion is computed over the set of models is not relevant. While, in [27], the author stated that the differences in AIC for pairs of nested models can be much more precisely estimated than those for some non-nested pairs, and the sampling error can make comparisons of AIC meaningless unless the differences are large.

In summary, although the classical methods have demonstrated a sound theoretical basis for choosing suitable model complexity, they suffer from limitations in parameter choice when the model's parameters do not have a priority order, especially for models involving a large number of parameters.

In this paper, we focus on addressing the parameter choice problem, for density estimation on the parameter space of multivariate distributions under a given number of free parameters. Note that the parameter reduction is related but different from traditional dimensionality reduction. Traditional dimensionality reduction techniques are often data-oriented and focus on directly reducing the number of random variables (features) while retaining the maximal variations in the high-dimensional data, e.g., feature selection [28] [29] and feature extraction [30] [31]. Generally speaking, they cannot be applied to the parametric reduction problem that this paper aims to tackle.

We propose a general parameter reduction criterion, namely the Confident-Information-First (CIF) principle, to maximally preserve confident parameters and rule out less confident ones. Formally, the confidence of each parameter can be assessed by its contribution to the expected Fisher information distance within a geometric manifold over the neighbourhood of the underlying real distribution, based on the theoretical framework of information geometry (IG) [32].

A. Motivation of CIF under IG

In IG, the general model S can be seen as a K-dimensional manifold and the goal of the parametric reduction is to derive a lower-dimensional sub-model M by reducing the number of free parameters in S. M is a smoothed submanifold of S. The number of free parameters in M is restricted to be a constant k ($k << K$). A more detailed introduction to IG is given in Section II.

The major difficulty in the parametric reduction procedure is the choice of parameters to keep or to remove. In this paper, we propose to reduce parameters such that the original geometric structure of S can be preserved as much as possible after projecting on the submanifold M.

Let $p_t, p_s \in S$ be the true distribution and the sampling distribution (maybe perturbed from p_t by sampling bias or noises) respectively. It can be assumed that the true distribution p_t is located somewhere in a ε-sphere surface B_s centered at p_s, i.e., $B_s = \{p_t \in S | D(p_t, p_s) = \varepsilon\}$, where $D(\cdot, \cdot)$ denotes some distance measure on the manifold S, and ε is a small number. This assumption is made without losing generality,

[1]For example, AIC [16] aims to reach a trade-off between the model complexity and its fitness to the sample data by minimizing the equation $AIC = -2 \log L + 2k$, where $\log L$ is the maximum log-likelihood and k is the number of the selected parameters.

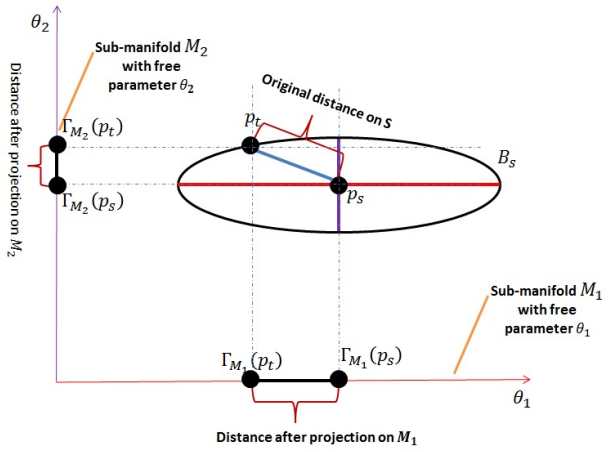

Fig. 1. Illustration of parametric reduction: Let S be a two-dimensional manifold with two free parameters θ_1 and θ_2. M_1 with a free parameter θ_1 and M_2 with a free parameter θ_2 are the submanifolds of S. As an illustration in Euclidean space, we show B_s (on which the true distribution p_t is located on) as the surface of a hyper-ellipsoid centered at sample distribution p_s, which is determined by the Fisher-Rao metric. Only part of the original distance between p_t and p_s ($p_t, p_s \in S$) can be preserved after they are projected onto a submanifold M. The preferred M is the one that maximally preserves the original distance after projection. Note that the scale of the distances in Fig 1 are shown as a demo, and are not exactly proportional to the real Riemann distances induced by Fisher-Rao metric

since the ε is a small variable. For a distribution p, the best approximation of p on M is the point q that belongs to M and is the closest to p in terms of the distance measure, i.e., $q = \arg\min_{q' \in M} D(q', p)$, which is defined as the projection of p onto M (denoted as $\Gamma_M(p)$).

Then, the parametric reduction can be defined as an optimization problem to maximally preserve the expectation of the squared Fisher information distance with respect to the constraint on the number of free parameters, when projecting distributions of S onto some submanifold M:

$$\begin{aligned} \underset{M}{\text{maximize}} \quad & E_{B_s}[D^2(\Gamma_M(p_t), \Gamma_M(p_s))] \\ \text{subject to} \quad & M \text{ has } k \text{ free parameters} \end{aligned} \tag{1}$$

where $p_t \in B_s$ and the expectation is taken on the surface B_s.

Here, the Fisher information distance (FID), i.e., the Riemannian distance induced by the Fisher-Rao metric [33], is adopted as the distance measure between two distributions. It has been shown that the Fisher-Rao metric could uniquely meet a set of natural axioms for distribution distance metric [32] [34] [35], e.g., the reparametrization invariant property and the monotonicity with respect to the random maps on variables. Let ξ be the distribution parameters. For two close distributions p_1 and p_2 with parameters ξ_1 and ξ_2, the Fisher information distance between p_1 and p_2 is:

$$D(p_1, p_2) = \sqrt{(\xi_1 - \xi_2)^T G_\xi (\xi_1 - \xi_2)} \tag{2}$$

where G_ξ represents the Fisher information matrix [32].

The rationality of maximally preserving the Fisher information distance can also be interpreted from the maximum-likelihood (ML) estimation's point of view. Let $\hat{\xi}$ be the ML estimators for ξ. The asymptotic normality of ML estimation implies that the density of $\hat{\xi}$ is the normal distribution with mean ξ and covariance Σ, i.e.,

$$f(\hat{\xi}) \sim \mathcal{N}(\xi, \Sigma) = \frac{1}{Z} exp\{-\frac{1}{2}(\xi - \hat{\xi})^T \Sigma^{-1} (\xi - \hat{\xi})\} \tag{3}$$

where the inverse of Σ can be asymptotically estimated using the Fisher information matrix G_ξ, as suggested by the Cramér-Rao bound [36]. From the Fisher information distance given in Equation 2, the exponent part of Equation 3 is just the opposite of the half squared Fisher information distance between two distributions p and $\hat{p}$ determined by the close parameters ξ and $\hat{\xi}$, respectively. Hence a larger Fisher information distance means a lower likelihood. It turns out that, in density estimation, maximally preserving the expected Fisher information distance after the projection Γ_M (Equation 1) is equivalent to maximally preserving the likelihood-structure among close distributions. In supervised learning tasks (e.g., classification), maximally preserving FID can also effectively preserve the likelihood-structure among different class densities (the generative distributions of classes), which helps prevent the interference of sample noises. Recall that sample noises always reduce the FID among class densities in a statistical sense, leading to a reduced discrimination marginality between two class densities. Hence, for noisy data, the model that maximally preserving FID will have a better discriminative capability.

To solve the optimization problem in Equation 1, we propose a parameter reduction criterion called the *Confident-Information-First* (CIF) principle, described as follows. For multivariate binary distributions, the squared Fisher information distance $D^2(p_t, p_s)$ can be decomposed into the distances of two orthogonal parts by using the dually orthogonal coordinates in IG [32]:

$$D^2(p_t, p_s) = D^2(\Gamma_M(p_t), \Gamma_M(p_s)) + D^2(\Gamma_{\overline{M}}(p_t), \Gamma_{\overline{M}}(p_s))$$

Here, the submanifold $\overline{M}$ is complementary to M, i.e., the free parameters in $\overline{M}$ are non-free in M, and vice versa. Thus, it is possible to categorize the system parameters of S into two sets, *i.e.*, the set of parameters in M with "major" variations and the set of parameters in $\overline{M}$ with "minor" variations, based on their contributions to the whole information distance. The former refers to parameters that are important for reliably distinguishing the true distribution from the sampling distribution, thus considered as "confident". Meanwhile, the parameters in $\overline{M}$ with minor contributions can be regarded as less reliable. Hence, the CIF principle can be stated as parametric reduction that preserves the confident parameters and rules out less confident parameters.

B. Outline and Contributions

The CIF principle is a general parametric choice framework that can be used to determine the optimal choice of submodel's free parameters. There exist different implementations of CIF with respect to different underlying distributions and/or their coordinate representations. The implementation of CIF can be analogized to some general inference objectives, e.g., the maximum likelihood (ML), which can achieve solutions in different senses (e.g., globally optimal or locally optimal) with

respect to different application contexts. In this paper, we demonstrate two kinds of implementation of the CIF principle in two different application scanarios.

First, when p_s is unknown, i.e., there is no observed information, the CIF principle could be used to derive an optimal probabilistic model in an expectation sense. If the l-mixed ζ-coordinates $[\zeta]_l$ (Equation 8) is adopted, we theoretically show that CIF could lead to an optimal submanifold M for $k = \sum_{i=1}^{l} C_n^i$ in term of the optimization problem in Equation 1 (see Section III). For $l = 2$, we further show that the optimal submanifold M derived by CIF is exactly the Boltzmann machine (see Section IV). Note that CIF may be used to derive other useful models if different coordinate systems are employed.

Second, when p_s is known, the CIF principle is applied in the model selection for BM as a heuristic method to decide the priority order of parameters, which can improve the search efficiency without degrading the quality of model selection results. For a given number of free parameters k, we can directly deduce a set of preferred parameters from all k-dimensional submodels by selecting the top-k confident parameters, where the confidences can be pre-calculated for efficiency. Then, by integrating the information criteria (such as AIC) with CIF, we can alleviate the possible bias of the model complexity when AIC is used to compare non-nested models and improve the computational efficiency of the model selection procedure. In this case, we only need to calculate the AIC for K sub-models to obtain a solution. Moreover, we have further developed separate hypothesis tests on the confidences of parameters to decide whether certain parameters should be preserved or not (see Section V-A), where the significance level α used in a hypothesis test can be considered as a heuristic threshold for the confidences of parameters.

II. Theoretical Foundations of IG

In this section, we introduce and develop the theoretical foundations of IG [32] for the manifold S of binary multivariate distributions with a given number of variables n, i.e., the open simplex of all probability distributions over binary vector $x \in \{0,1\}^n$. This will lay the foundation for our theoretical deviation of CIF.

A. Notations for Manifold S

In IG, a parametric family of probability distributions is regarded as a differentiable manifold equipped with certain coordinate systems. In the case of binary multivariate distributions, there exist four commonly used coordinate systems [32] [37]: p-coordinates, η-coordinates, θ-coordinates and the mixed ζ-coordinates. The ζ-coordinates is very important for the analysis in this paper.

For the p-coordinates $[p]$, the probability distribution over n binary variables can be completely specified by a vector of $2^n - 1$ positive numbers, corresponding to the probability of any $2^n - 1$ exclusive states of x. For example, when $n = 2$, the p-coordinates could be $[p] = (p_{01}, p_{10}, p_{11})$. Note that all probability terms need to be strictly positive as required by IG [32]. For simplicity, the p-coordinates are indexed by capital letters $I, J, \ldots$, where an index I denotes a subset of $\{1, 2, \ldots, n\}$. p_I is defined as the probability that all variables included in I are one and the complemented variables equal to zero. For example, if $I = \{1, 3\}$ and $n = 3$, we have:

$$p_I = p_{101} = Prob(x_1 = 1, x_2 = 0, x_3 = 1)$$

Note that the null set $I = \{\}$ is also valid for the $[p]$ and $p_{0\ldots0}$ denotes the probability that all variables are zero.

The η-coordinates $[\eta]$ are defined by:

$$\eta_I = E[X_I] = Prob\{\prod_{i \in I} x_i = 1\} \tag{4}$$

where $X_I = \prod_{i \in I} x_i$ and $E[\cdot]$ denotes the expectation with respect to the probability distribution over x. The η-coordinates can be grouped by coordinate orders and denoted as $[\eta] = (\eta_i^1, \eta_{ij}^2, \ldots, \eta_{1,2\ldots n}^n)$, where the superscript represents the order number of the corresponding parameter. For example, η_{ij}^2 denotes the set of all η parameters with the order number two.

The θ-coordinates (natural coordinates) $[\theta]$ are defined by:

$$\log p(x) = \sum_{I \subseteq \{1,2,\ldots,n\}, I \neq NullSet} \theta^I X_I - \psi(\theta) \tag{5}$$

where $\psi(\theta) = \log(\sum_x exp\{\sum_I \theta^I X_I(x)\})$ is the cumulant generating function and its value equals to $-\log Prob\{x_i = 0, \forall i \in \{1, 2, ..., n\}\}$. By solving the linear system 5, we have $\theta^I = \sum_{K \subseteq I} (-1)^{|I-K|} log(p_K)$. The θ-coordinate is denoted as $[\theta] = (\theta_1^i, \theta_2^{ij}, \ldots, \theta_n^{1,\ldots,n})$, where the subscript represents the order number of the corresponding parameter. Note that the order indices locate differently in $[\eta]$ and $[\theta]$ by convention in [33].

There exists bijective map between coordinate systems $[\eta]$ and $[\theta]$, formally given by the Legendre transformation:

$$\theta^I = \frac{\partial \phi(\eta)}{\partial \eta_I}, \eta_I = \frac{\partial \psi(\theta)}{\partial \theta^I} \tag{6}$$

where $\psi(\theta)$ is specified in Equation 5 and $\phi(\eta) = \sum_x p(x;\eta) \log p(x;\eta)$ denotes the negative entropy. It can be shown that $\psi(\theta)$ and $\phi(\eta)$ satisfy the following equation [32]:

$$\psi(\theta) + \phi(\eta) - \sum \theta^I \eta_I = 0 \tag{7}$$

The l-mixed ζ-coordinates $[\zeta]_l$ are defined by:

$$[\zeta]_l = [\eta^{l-}, \theta_{l+}] = (\eta_i^1, \eta_{ij}^2, \ldots, \eta_{i,j,\ldots}^l, \theta_{l+1}^{i,j,\cdots}, \ldots, \theta_n^{1,\ldots,n}) \tag{8}$$

where $l \in \{1, ..., n-1\}$. The former part η^{l-} consists of η-coordinates with order less or equal to l and the latter part θ_{l+} consists of θ-coordinates with order greater than l.

B. Fisher Information Matrix for Parametric Coordinates

For a general coordinate system $[\xi]$, the i-th row and j-th column element of the Fisher information matrix G_ξ for $[\xi]$ is defined as the covariance of the scores of ξ_i and ξ_j [36]:

$$g_{ij} = E[\frac{\partial \log p(x;\xi)}{\partial \xi_i} \cdot \frac{\partial \log p(x;\xi)}{\partial \xi_j}]$$

where $E[\cdot]$ denotes the expectation with respect to the probability distribution over x. The Fisher information measures

the amount of information in the data that a statistic carries about the unknown parameters [38]. The Fisher information matrix is of vital importance to our analysis, since the inverse of the matrix gives an asymptotically tight lower bound to the covariance matrix of any unbiased estimators for the parameters under consideration [36]. Another important concept related to our analysis is the orthogonality defined by Fisher information. Two coordinate parameters ξ_i and ξ_j are called orthogonal if and only if their Fisher information vanishes, *i.e.*, $g_{ij} = 0$, meaning that their influences on the log likelihood function are uncorrelated.

According to [32], the Fisher information for $[\theta]$ can be rewritten as $g_{IJ} = \frac{\partial^2 \psi(\theta)}{\partial \theta^I \partial \theta^J}$, and for $[\eta]$, it is $g^{IJ} = \frac{\partial^2 \phi(\eta)}{\partial \eta_I \partial \eta_J}$. Let $G_\theta = (g_{IJ})$ be the Fisher information matrix for $[\theta]$ and $G_\eta = (g^{IJ})$ for $[\eta]$. As shown in [32], G_θ and G_η are mutually inverse matrices, *i.e.*, $\sum_J g^{IJ} g_{JK} = \delta^I_K$, where $\delta^I_K = 1$ if $I = K$ and zero if $I \neq K$. Next, we will develop the two propositions to compute G_θ and G_η generally. Note that Proposition 2.1 is a generalization of Theorem 2 in [33].

Proposition 2.1: The Fisher information between two parameters θ^I and θ^J in $[\theta]$, is given by:

$$g_{IJ}(\theta) = \eta_{I \cup J} - \eta_I \eta_J \tag{9}$$

Proof in Appendix A.

Proposition 2.2: The Fisher information between two parameters η_I and η_J in $[\eta]$, is given by:

$$g^{IJ}(\eta) = \sum_{K \subseteq I \cap J} (-1)^{|I-K|+|J-K|} \cdot \frac{1}{p_K} \tag{10}$$

where $|\cdot|$ denotes the cardinality operator.

Proof in Appendix B.

We take the probability distribution with three variables for example. Based on Equation 10, the Fisher information between η_I and η_J can be calculated, e.g., $g^{IJ} = \frac{1}{p_{000}} + \frac{1}{p_{010}}$ if $I = \{1, 2\}$ and $J = \{2, 3\}$, $g^{IJ} = -(\frac{1}{p_{000}} + \frac{1}{p_{010}} + \frac{1}{p_{100}} + \frac{1}{p_{110}})$ if $I = \{1, 2\}$ and $J = \{1, 2, 3\}$, and etc.

The next proposition shows that the Fisher information matrix G_ζ for the $[\zeta]_l$ can be calculated based on G_η and G_θ.

Proposition 2.3: The Fisher information matrix G_ζ of $[\zeta]_l$ is given by:

$$G_\zeta = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \tag{11}$$

where $A = ((G^{-1}_\eta)_{I_\eta})^{-1}$, $B = ((G^{-1}_\theta)_{J_\theta})^{-1}$, G_η and G_θ are the Fisher information matrices of $[\eta]$ and $[\theta]$, respectively, I_η is the index set of the parameters shared by $[\eta]$ and $[\zeta]_l$, *i.e.*, $\{\eta^1_i, ..., \eta^l_{i,j,...}\}$, and J_θ is the index set of the parameters shared by $[\theta]$ and $[\zeta]_l$, *i.e.*, $\{\theta^{i,j,\cdots}_{l+1}, \ldots, \theta^{1,\ldots,n}_n\}$.

Proof in Appendix C.

III. The General CIF Principle

The general manifold S of all probability distributions over binary vector $x \in \{0, 1\}^n$ could be exactly represented using $2^n - 1$ parametric coordinates. Given a target distribution $q(x) \in S$, we would like to realize it by a distribution in some lower-dimensional submanifold M. This is defined as the parametric reduction problem for multivariate binary distributions.

In this section, we will formally illuminate the general CIF for parametric reduction. Assume we can construct a coordinate system whose parameters entail a natural hierarchy according to their confidences, meaning that high confident parameters are significantly distinguished from and orthogonal to lowly confident ones. Then we could implement the CIF conveniently by assigning the lowly confident parameters to neutral values and keeping the high confident parameters unchanged. As described in Section I, we should assess the confidence of parameters according to their contributions to the expected information distance. Therefore, the choice of coordinates is crucial for the CIF principle. This strategy is infeasible in terms of p-coordinates, η-coordinates or θ-coordinates for two main reasons: first, the orthogonality does not hold in those coordinates, and hence we can NOT safely prune some parameters without affecting the values of others; second, in those coordinate systems, highly confident parameters cannot be significantly distinguished from lowly ones. In this section, we will focus on the l-mixed-coordinates $[\zeta]_l$ and show how $[\zeta]_l$ meets all requirements of CIF.

We will first show that $[\zeta]_l$ meets the requirements of CIF in typical distributions that generate real-world datasets, so as to grasp an intuitive picture for the general CIF strategy. Then we will prove that CIF could lead to an optimal submanifold w.r.t. the parametric reduction problem in Equation 1, in general cases.

A. The CIF in Typical Distributions

In this section, we consider the typical situation in real-world data collections to facilitate our analysis, where there are only a small fraction of all system states are frequent and meaningful patterns [39]. More formally, the underlying distributions $q(x)$ is assumed to have at least $(2^n - 2^{n/2})$ p-coordinates of the scale ϵ, where ϵ is a sufficiently small value. Therefore, the residual at most $2^{n/2}$ p-coordinates are all of scale $\Theta(1/2^{(n/2)})$, and their sum approximates 1.

Next, for the true distribution $q(x)$, we introduce a small perturbation Δp to the p-coordinates $[p]$, and the perturbed distribution is denoted as $q'(x)$. For those p-coordinates with a sufficiently small value, we assume that the scale of the fluctuation Δp_I for p_I is proportional to $a \cdot p_I$, where a is a small constant coefficient. For p-coordinates that are significantly larger than zero, the scale of the fluctuation Δp_I for p_I is assumed to be proportional to the standard deviation of the estimate of the corresponding p_I. According to the Cramér–Rao bound theory, the standard deviation can be approximated by the inverse of the square root of the Fisher information. After some algebra, we can hence assume the perturbation Δp_I to be $a\sqrt{p_I}$.

In our previous work [40], we have analyzed the l-mixed-coordinates $[\zeta]_l = (\eta^{l-}; \theta_{l+})$, where $l = 2$. The incremental of mixed-coordinates caused by perturbation Δp is denoted as $\Delta\zeta_q = (\Delta\eta^{2-}; \Delta\theta_{2+})$. By decomposing the squared Fisher

TABLE I
SIMULATION ON THE FID PRESERVED BY $[\zeta]_{l_t}$ ($l = 2$)

n	ratio of preserved parameters	ratio of preserved FID	
		mean	standard deviation
3	0.857	0.9972	0.0055
4	0.667	0.9963	0.0043
5	0.484	0.9923	0.0054
6	0.333	0.9824	0.0112
7	0.220	0.9715	0.0111

information distance $D^2(q, q') = (\Delta\zeta_q)^T G_\zeta \Delta\zeta_q$ into the direction of each coordinate in $[\zeta]_l$, we have shown that the scale of the Fisher information distance in each coordinate of η^{l-} is significantly larger than that of θ_{l+} (refer to our previous paper for more details [40]).

As described above, the confidences of coordinate parameters (measured by the decomposed Fisher information distance) in $[\zeta]_l$ entail a natural hierarchy. Moreover, the parameters in $[\eta^{l-}]$ are orthogonal to the ones in $[\theta_{l+}]$ [37]. Additionally, those low confident parameters $[\theta_{l+}]$ have the neutral value of zero. Hence, we can perform parametric reduction in $[\zeta]_l$ using the CIF principle by setting low confident parameters to be 0 and reconstructing the distribution according to the new coordinates. The resulting submanifold M tailored by CIF becomes $[\zeta]_{l_t} = (\eta^1_i, ..., \eta^l_{ij...k}, 0, \ldots, 0)$, which is called $[\zeta]_{l_t}$ the l-tailored-mixed-coordinates.

To verify our theoretical analysis, we conduct a simulation on the ratio of FID that is preserved by the l-tailored-mixed-coordinates $[\zeta]_{l_t}$ ($l = 2$) w.r.t. the original mixed-coordinates $[\zeta]$. We also show the corresponding ratio of preserved parameters: number of free parameters / total number of parameters. First we randomly select real distribution p_t with n variables, where the distribution satisfies the basic assumption that we make in the beginning of this section. The $2^{n/2}$ significant p-coordinates are generated based on Jeffery prior (the Dirichlet distribution with alpha parameters set to 0.5), and the left p-coordinates are set to a small constant. Then, we generate the sample distribution p_s based on random samples drawn from the real distribution. Last, we calculate the FID between p_t and p_s in terms of the $[\zeta]$ and $[\zeta]_{l_t}$ respectively. The result is shown in Table I. We can see that the $[\zeta]_{l_t}$ can indeed preserve most of the FID, which is consistent with our theoretical analysis.

B. The CIF Leads to an Optimal Submanifold M

Let B_q be a ε-sphere surface centered at $q(x)$ on manifold S, *i.e.*, $B_q = \{q' \in S | KL(q, q') = \varepsilon\}$, where $KL(\cdot, \cdot)$ denotes the KL divergence and ε is small. Let $q'(x)$ be a close neighbor of $q(x)$, which is uniformly sampled from the surfaceB_q, as shown in Figure 2. Recall that, for a small ε, the KL divergence can be approximated by half of the squared Fisher information distance. Thus, using the parameterization of $[\zeta]_l$, B_q is indeed the surface of a hyper-ellipsoid determined by G_ζ with the center at $q(x)$. The next proposition shows that the general CIF would lead to an optimal submanifold M, which could maximally preserve the expected squared Fisher information distance. Note that the expectation is computed w.r.t the uniform density on B_q.

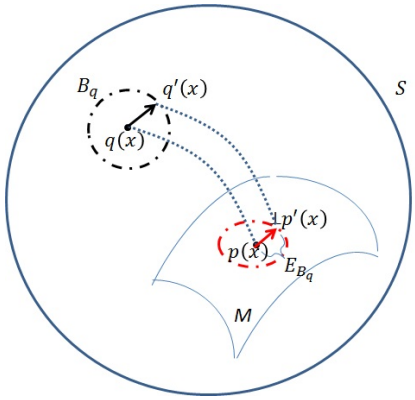

Fig. 2. By projecting a point $q(x)$ on S to a submanifold M, the l-tailored mixed-coordinates $[\zeta]_{l_t}$ gives a desirable M that maximally preserves the expected Fisher information distance when projecting a ε-neighborhood centered at $q(x)$ onto M.

Proposition 3.1: Consider the manifold S in l-mixed-coordinates $[\zeta]_l$. Let k be the number of free parameters in the l-tailored-mixed-coordinates $[\zeta]_{l_t}$, i.e., $k = C^1_n + C^2_n \cdots + C^l_n$. Then, among all k-dimensional submanifolds of S, the submanifold determined by $[\zeta]_{l_t}$ can maximally preserve the expected squared Fisher information distance induced by the Fisher–Rao metric.

Proof Assume $I_T = \{i_1, i_2, \ldots, i_k\}$ is the index of k coordinates that we choose to form the tailored submanifold T in the mixed-coordinates $[\zeta]$. This proof comes in three parts:

1) According to the fundamental analytical properties of the surface of the hyper-ellipsoid, we show that there exists a strict positive monotonicity between the expected information distance for T and the sum of eigenvalues of the sub-matrix $(G_\zeta)_{I_T}$;
2) To maximize the preserved information distance, we should choose I_T to preserve the top-k eigenvalues of G_ζ, i.e., $\lambda_1, \ldots, \lambda_k$;
3) Actually $I_T = \{\eta^{l-}\}$ gives the maximum Fisher information distance.

See Appendix D for the detailed proof.

IV. INTERPRETATION OF BOLTZMANN MACHINE USING CIF

In previous section, an implementation of CIF is uncovered in the $[\zeta]_l$ coordinates for multivariate binary distributions. Next, when l equals to 2, we show that the optimal submanifold M derived by CIF is exactly the Boltzmann machine. This can help us uncover and formalize the essential parts of the target density that BM aims to capture and the non-essential parts that BM discards.

A. Introduction to the Boltzmann Machines

In general, a BM [41] is defined as a stochastic neural network consisting of visible units $x \in \{0, 1\}^{n_x}$ and hidden units $h \in \{0, 1\}^{n_h}$.

The energy function is defined as follows:

$$E_{BM}(x, h; \xi) = -\frac{1}{2} x^T U x - \frac{1}{2} h^T V h - x^T W h - b^T x - d^T h \quad (12)$$

where $\xi = \{U, V, W, b, d\}$ are the free parameters. U, V and W represent the visible-visible, hidden-hidden and visible-hidden interactions respectively. b and d represent the visible and hidden self-connections respectively. The diagonals of U and V are set to zero. The joint Boltzmann distribution of x and h can be expressed as below:

$$p(x, h; \xi) = \frac{1}{Z} exp\{-E_{BM}(x, h; \xi)\} \tag{13}$$

where Z is a normalization factor.

1) The Coordinates for Boltzmann Machines:

Let B be the set of Boltzmann distributions realized by BM. Actually, B is a submanifold of the general manifold S_{xh} over $\{x, h\}$. From Equation (13) and (12), we can see that $\xi = \{U, V, W, b, d\}$ plays the role of B's coordinates in θ-coordinates (Equation 5) as follows:

$$\begin{aligned}
\theta_1 \quad &: \quad \theta_1^{x_i} = b_{x_i}, \theta_1^{h_j} = d_{h_j} (\forall x_i \in x, h_j \in h) \\
\theta_2 \quad &: \quad \theta_2^{x_i x_j} = U_{x_i, x_j}, \theta_2^{x_i h_j} = W_{x_i, h_j}, \\
&\quad \theta_2^{h_i h_j} = V_{h_i, h_j}, (\forall x_i, x_j \in x; h_i, h_j \in h) \\
\theta_{2+} \quad &: \quad \theta_m^{x_i \ldots x_j h_u \ldots h_v} = 0, m > 2, \\
&\quad (\forall x_i, \ldots, x_j \in x; h_u, \ldots, h_v \in h)
\end{aligned} \tag{14}$$

So the θ-coordinates for BM is given by:

$$[\theta]_{BM} = (\underbrace{\theta_1^{x_i}, \theta_1^{h_j}}_{1-order}, \underbrace{\theta_2^{x_i x_j}, \theta_2^{x_i h_j}, \theta_2^{h_i h_j}}_{2-order}, \underbrace{0, \ldots, 0}_{orders>2}). \tag{15}$$

The fully visible BM (VBM) and restricted BM (RBM) are special cases of the general BM. Since VBM has no hidden units ($n_h = 0$) and all the visible units are connected to each other, the parameters of VBM are $\xi_{vbm} = \{U, b\}$ and $\{V, W, d\}$ are all set to zero. For RBM, it has connections only between hidden and visible units. Thus, the parameters of RBM are $\xi_{rbm} = \{W, b, d\}$ and $\{U, V\}$ are set to zero.

2) The Gradient-based Learning of BM:

Given the sample $\underline{x}$ that generated from the underlying distribution, the *maximum-likelihood* (ML) is a commonly used gradient ascent method for training BM in order to maximize the log-likelihood $\log p(\underline{x}; \xi)$ of the parameters ξ [42]. Based on Equation (13), the log-likelihood is given as follows:

$$\log p(\underline{x}; \xi) = log \sum_h e^{-E(\underline{x}, h; \xi)} - log \sum_{x', h'} e^{-E(x', h'; \xi)}$$

Differentiating the log-likelihood, the gradient with respect to ξ is as follows:

$$\begin{aligned}
\frac{\partial \log p(\underline{x}; \xi)}{\partial \xi} \quad &= \quad \sum_h p(h|\underline{x}; \xi) \frac{\partial [-E(\underline{x}, h; \xi)]}{\partial \xi} \\
&\quad - \sum_{x', h'} p(h'|x'; \xi) \frac{\partial [-E(x', h'; \xi)]}{\partial \xi}
\end{aligned} \tag{16}$$

where $\frac{\partial E(x, h; \xi)}{\partial \xi}$ can be easily calculated from Equation (12). Then we can obtain the stochastic gradient using Gibbs sampling [43] in two phases: sample $\underline{h}$ given $\underline{x}$ for the first term, called the positive phase, and sample $(\underline{x}', \underline{h}')$ from the stationary distribution $p(x', h'; \xi)$ for the second term, called the negative phase. Now with the resulting stochastic gradient estimation, the learning rule is to adjust ξ by:

$$\begin{aligned}
\Delta \xi_{ml} \quad &= \quad \varepsilon \frac{\partial \log p(\underline{x}; \xi)}{\partial \xi} \\
&\propto \quad -\langle \frac{\partial E(\underline{x}, \underline{h}; \xi)}{\partial \xi} \rangle_0 + \langle \frac{\partial E(\underline{x}', \underline{h}'; \xi)}{\partial \xi} \rangle_\infty
\end{aligned} \tag{17}$$

where ε is the learning rate, $\langle \cdot \rangle_0$ denotes the average using the sample data and $\langle \cdot \rangle_\infty$ denotes the average with respect to the stationary distribution $p(x, h; \xi)$ after the corresponding Gibbs sampling phases.

To avoid the difficulty of computing the log-likelihood gradient, the Contrastive divergence (CD) [42] optimizes a different objective function based on KL divergence, shown as follows:

$$\begin{aligned}
\Delta \xi_{cd} \quad &= \quad -\varepsilon \frac{\partial (KL(p_0||p) - KL(p_m||p))}{\partial \xi} \\
&\propto \quad -\langle \frac{\partial E(\underline{x}, \underline{h}; \xi)}{\partial \xi} \rangle_0 + \langle \frac{\partial E(\underline{x}', \underline{h}'; \xi)}{\partial \xi} \rangle_m
\end{aligned} \tag{18}$$

where $KL(\cdot||\cdot)$ denotes the KL divergence, p_0 represents the sample distribution, and p_m denotes the distribution sampled from the Markov chain after running m steps (begins with the sample data). CD can be seen as an approximation to ML by replacing the last expectation $\langle \cdot \rangle_\infty$ with $\langle \cdot \rangle_m$.

B. The Fully Visible Boltzmann Machine

Consider the parametric reduction on the manifold S over $\{x\}$ and result with a k-dimensional submanifold M of S, where $k \ll 2^{n_x} - 1$. M is set to be the same dimensionality as VBM, i.e., $k = \frac{n_x(n_x+1)}{2}$, then all candidate M are comparable to the submanifold M_{vbm} endowed by VBM.

In the following corollary, we theoretically show that the CIF principle leads to the optimal submanifold that is exactly the submanifold specified by Bolzmann machines.

Corollary 4.1: Given the general manifold S in 2-mixed-coordinates $[\zeta]_2$, VBM (with coordinates $[\zeta]_{2_t}$) defines an k-dimensional submanifold of S that can maximally preserve the expected squared Fisher information distance induced by Fisher-Rao metric.

Proof in Appendix E.

To learn such $[\zeta]_{2_t}$, we need to learn the parameters ξ of VBM such that its stationary distribution preserves the same coordinates $[\eta^{2-}]$ as target distribution $q(x)$. Actually, this is exactly what traditional gradient-based learning algorithms intend to do. Next proposition shows that the ML learning of VBM is equivalent to learn the coordinates $[\zeta]_{2_t}$: 1) high-order (θ_{2+}) components are set to zero; 2) low-order (η_i^1, η_{ij}^2) components are unchanged comparing to the original distribution.

Proposition 4.2: Given the target distribution $q(x)$ with 2-mixed coordinates:

$$[\zeta]_2 = (\eta_i^1, \eta_{ij}^2, \theta_{2+})$$

the coordinates of the stationary distribution of VBM trained by ML are uniquely given by $[\zeta]_{2_t}$:

$$[\zeta]_{2_t} = (\eta_i^1, \eta_{ij}^2, \theta_{2+} = 0)$$

Proof in Appendix F.

C. The Boltzmann Machine with Hidden Units

In previous section, the CIF principle is applied to models without hidden units and leads to VBM by preserving the 1-order and 2-order η-coordinates. In this section, we will investigate the cases where hidden units are introduced.

Let S_{xh} be the manifold of distributions over the joint space of visible units x and hidden units h. A general BM produces a stationary distribution $p(x, h; \xi) \in S_{xh}$ over $\{x, h\}$. Let B denote the submanifold of S_{xh} with probability distributions $p(x, h; \xi)$ realizable by BM.

Given any target distribution $q(x)$, only the marginal distribution of BM over the visible units are specified, leaving the distributions on hidden units vary freely. Let H_q be the submanifold of S_{xh}, consisting of probability distributions $q(x, h)$ that have the same marginal distribution as $q(x)$ and the conditional distribution $q(h|x)$ of hidden units is realised by the BM's activation functions with some parameter ξ_{bm}.

Then, the best BM is the one that minimizes the distance between B and H_q. Due to the existence of hidden units, the solution may not be unique. In this section, the training process of BM is analysed in terms of manifold projection (described in Section I), following the framework of the learning rule proposed in [33]. And we will show that the learning of BM with hidden units can be interpreted using CIF.

1) The Iterative Projection Learning for BM:

The learning algorithm using iterative manifold projection is first proposed in [33] and theoretically compared to EM (Expectation and Maximization) algorithm in [44]. The learning of RBM can be implemented by the following iterative projection process: Let ξ_p^0 be the initial parameters of BM and $p_0(x, h; \xi_p^0)$ be the corresponding stationary distribution.

For $i = 0, 1, 2, \ldots,$

1) Put $q_{i+1}(x, h) = \Gamma_H(p_i(x, h; \xi_p^i))$
2) Put $p_{i+1}(x, h; \xi_p^{i+1}) = \Gamma_B(q_{i+1}(x, h))$

where $\Gamma_H(p)$ denotes the projection of $p(x, h; \xi_p)$ to H_q, and $\Gamma_B(q)$ denotes the projection of $q(x, h)$ to B. The iteration ends when we reach the fixed points of the projections p^* and q^*, that is $\Gamma_H(p^*) = q^*$ and $\Gamma_B(q^*) = p^*$. The iterative projection process is illustrated in Figure 3. The convergence property of this iterative algorithm is guaranteed using the following proposition:

Proposition 4.3: The monotonic relation holds in the iterative learning algorithm:

$$D[q_{i+1}, p_i] \geq D[q_{i+1}, p_{i+1}] \geq D[q_{i+2}, p_{i+1}] \qquad (19)$$

where the equality holds only for the fixed points of the projections.

Proof in Appendix G.

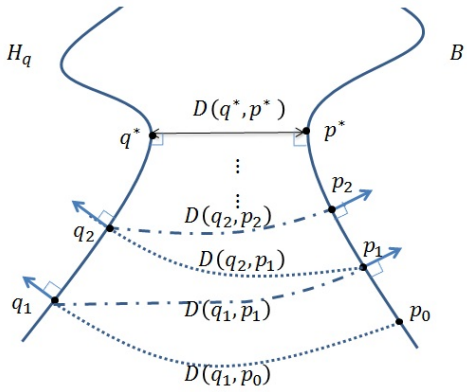

Fig. 3. The iterative learning for BM: in searching for the minimum distance between H_q and B, we first choose an initial BM p_0 and then perform projections $\Gamma_H(p)$ and $\Gamma_B(q)$ iteratively, until the fixed points of the projections p^* and q^* are reached. With different initializations, the iterative projection algorithm may end up with different local minima on H_q and B, respectively.

Next two propositions show how the projection $\Gamma_H(p)$ and $\Gamma_B(q)$ are obtained.

Proposition 4.4: Given a distribution $p(x, h; \xi_p) \in B$, the projection $\Gamma_H(p) \in H_q$ that gives the minimum divergence $D(H_q, p(x, h; \xi_p))$ from H_q to $p(x, h; \xi_p)$ is the $q(x, h; \xi_{bm}) \in H_q$ that satisfies $\xi_{bm} = \xi_p$.

Proof in Appendix H.

Proposition 4.5: Given $q(x, h; \xi_q) \in H_q$ with mixed coordinates: $[\zeta^{xh}]_q = (\eta_{x_i}^1, \eta_{h_j}^1, \eta_{x_i x_j}^2, \eta_{x_i h_j}^2, \eta_{h_i h_j}^2, \theta_{2+})$, the coordinates of the learnt projection $\Gamma_B(q) \in B$ are uniquely given by the tailored mixed coordinates:

$$[\zeta^{xh}]_{\Gamma_B(q)} = (\eta_{x_i}^1, \eta_{h_j}^1, \eta_{x_i x_j}^2, \eta_{x_i h_j}^2, \eta_{h_i h_j}^2, \theta_{2+} = 0) \qquad (20)$$

Proof This proof comes in three parts:

1) the projection $\Gamma_B(q)$ of $q(x, h)$ on B is unique;
2) this unique projection $\Gamma_B(q)$ can be achieved by minimizing the divergence $D[q(x, h), B]$ using gradient descent method;
3) The mixed coordinates of $\Gamma_B(q)$ is exactly the one given in Equation (20).

See Appendix I for the detailed proof.

2) The Interpretation for BM Learning via CIF:

The iterative projection learning (IP) gives us an alternative way to investigate the learning process of BM. Based on the CIF principle in Section III, we can see that the process of the projection $\Gamma_B(q_i)$ can be interpreted using CIF, i.e., highly confident coordinates $[\eta_{x_i}^1, \eta_{h_j}^1, \eta_{x_i x_j}^2, \eta_{x_i h_j}^2, \eta_{h_i h_j}^2]$ of q_i are preserved while lowly confident coordinates $[\theta_{2+}]$ are set to neutral value zero, given in Equation 20.

In summary, the essential parts of the real distribution that can be learnt by BM (with and without hidden units) are exactly the confident coordinates indicated by the CIF principle. As a special kind of BM, the commonly used RBM can be analysed similarly.

V. Experimental Study on Sample-specific CIF

In previous sections, when p_s is unknown and the l-mixed ζ-coordinates $[\zeta]_l$ is adopted, we have shown that there exist globally optimal solutions of CIF for $k = \sum_{i=1}^{l} C_n^i$ in terms of the optimization problem in Equation 1. For $l = 2$, we further show that the optimal submanifold M derived by CIF is exactly the BM.

Given specific samples, *can CIF further recognize less-confident parameters in BM and reduce them properly*? For VBM, we will investigate how to use CIF to modify the topology of VBM by reducing less confident connections among visible units with respect to given samples. We will empirically investigate this sample-specific CIF principle as a heuristic method to decide the priority order of parameters, so as to improve the search efficiency for the model selection of BM without degrading the quality of model selection results.

A. The Sample-specific CIF: Adaptive Model Selection of BM

The data constrains the state of knowledge about the unknown distribution. Let $q(x)$ denote the sampling distribution (representing the data). In order to force the estimate of our probabilistic model (denoted as $p(x;\xi)$) to meet the data, we could incorporate the data into CIF by recognizing the confidence of parameters ξ in terms of $q(x)$. Then, parametric reduction procedure can be further applied to modify the topology of VBM adaptively according to the data, as shown in Algorithm 1 and explained as in the following.

Algorithm 1 Adaptive Network Design for BM

Input: Samples $D = \{d_1, d_2, \ldots, d_N\}$; Significance level α; Nodes $V = \{x_1, x_2, \ldots, x_n\}$; Edges $U = \{U_{ij}, \forall x_i, x_j\}$;
Output: Set of confident edges $U_{conf} \subset U$

```
U_conf ← {};
for U_ij ∈ U do
    Estimate marginal distribution p(x_i, x_j) from samples
    ** parameterize to ζ-coordinates: [ζ] **
    η_i ← E_p[x_i]; η_j ← E_p[x_j]
    θ^ij ← log p_00 − log p_01 − log p_10 + log p_11
    [ζ] ← {η_i, η_j, θ^ij}

    ** Fisher information of θ^ij in [ζ] **
    g ← (1/p_00 + 1/p_01 + 1/p_10 + 1/p_11)^{-1}

    ** confidence of θ^ij in [ζ] **
    ρ_ij ← θ^ij · g · θ^ij

    ** hypothesis test: ρ_ij = 0 against ρ_ij ≠ 0 **
    π ← cdf_{χ²(1)}(N ρ_ij)
    if (1 − π) · 2 < α then
        ** reject null hypothesis: ρ_ij = 0 **
        U_conf ← U_conf ∪ {U_ij}
    end if
end for
return U_conf
```

As a graphical model, the VBM comprises a set of vertices $V = \{x_1, x_2, \ldots, x_n\}$ together with a set of connections $U = \{U_{ij}, \forall x_i, x_j, i \neq j\}$. The confidence for each connection parameter U_{ij} can be assessed by the parameter choice criterion in CIF, i.e., the contribution to the Fisher information distance. Based on the Theorem 1 in [33], U_{ij} could be expressed as follows:

$$U_{ij} = \log \frac{p(x_i = x_j = 1|A) \cdot p(x_i = x_j = 0|A)}{p(x_i = 1, x_j = 0|A) \cdot p(x_i = 0, x_j = 1|A)}$$

where the relation hold for any conditions A on the rest variables. However, it is often infeasible for us to calculate the exact value of U_{ij} because of data sparseness. To tackle this problem, we propose to approximate the value of U_{ij} by using the marginal distribution $p(x_i, x_j)$ to avoid the effect of condition A. To estimate $p(x_i, x_j)$, we need to go through all N samples and count the number of samples for each assignment of x_i and x_j. For example, $p(x_i = 0, x_j = 0) = \frac{count(x_i=0, x_j=0)}{N}$, and etc.

Let $[\zeta]_{ij} = (\eta_i, \eta_j, \theta^{ij})$ be the mixed-coordinates for the marginal distribution $p(x_i, x_j)$ of VBM. Note that each θ^{ij} corresponds to one connection U_{ij}. Since θ^{ij} is orthogonal to η_i and η_j, the Fisher information distance between two distributions can be decomposed into two independent parts: the information distance contributed by $\{\eta_i, \eta_j\}$ and $\{\theta^{ij}\}$. For the purpose of parameter reduction, we consider the two close distributions p_1 and p_2 with coordinates $\zeta_1 = \{\eta_i, \eta_j, \theta^{ij}\}$ and $\zeta_2 = \{\eta_i, \eta_j, 0\}$ respectively. The confidence of θ^{ij}, denoted as $\rho(\theta^{ij})$, can be estimated by its contribution to Fisher information distance between p_1 and p_2:

$$\rho(\theta^{ij}) = (\zeta_1 - \zeta_2)^T G_\zeta (\zeta_1 - \zeta_2) = \theta^{ij} \cdot g_\zeta(\theta^{ij}) \cdot \theta^{ij} \quad (21)$$

where G_ζ is the Fisher information matrix in Proposition 2.3 and $g_\zeta(\theta^{ij})$ is the Fisher information for θ^{ij}. Note that the second equality holds since θ^{ij} is orthogonal to η_i and η_j.

For a given number of free parameters k, we can directly decide the optimal subset of connections for VBM by selecting the top-k confident parameters. In principle, this solution could be globally optimal in the sense of preserving the Fisher-Rao distance, since the orthogonality of the fractional mixed coordinates [45]. However, a global solution is time-consuming. In this paper, we use a greedy heuristic to 'approximate' the global solution. Our heuristic separately evaluates the confidence of each parameter U_{ij} w.r.t the marginal distribution $p(x_i, x_j)$, instead of the whole joint distribution $p(x_1, x_2, \ldots, x_n)$. Then, to decide whether the Fisher information distance in the coordinate direction of θ^{ij} is significant or negligible, we set up the hypothesis test for ρ, i.e., null hypothesis $\rho = 0$ versus alternative $\rho \neq 0$. Based on the analysis in [46], we have $N\rho \sim \chi^2(1)$ asymptotically, where the $\chi^2(1)$ is chi-square distribution with degree of freedom 1 and N is the sampling number. In this way, by setting the significant level α, we can simply determine the threshold for ρ, i.e., only those parameters with confidence value greater than the threshold are selected. For example, when $\alpha = 0.05$ and $N\rho > 5.024$, we can ensure that the Fisher information distance in the direction of θ^{ij} (w.r.t., the marginal distribution $p(x_i, x_j)$) is significant with at least 95% confidence. This model selection method is shown in Algorithm 1, called the CIF+α.

B. Experiments with VBM

In this section, we investigate the density estimation performance of CIF-based model selection methods for VBM, i.e., CIF+AIC and CIF+α. Two comparative model selection methods are used:

- StepwiseAIC: the backward elimination approach is adopted to select the subset of connections with the optimal AIC: starting with all candidate connections; testing AIC after the deletion of each connection; deleting the connection with the lowest AIC; repeating this process until AIC stops decreasing.
- L_1-norm: a full VBM is trained to minimize the L_1-regularization objective:$-\log L(\xi) + \lambda\|\xi\|_1$, where ξ is the set of parameters, $\log L(\xi)$ denotes the log-likelihood, λ is the regularization coefficient and $\|\cdot\|_1$ represents the L_1 norm. The optimal coefficient λ_{opt} is chosen via 5-fold cross-validation. Then, those connections with significant absolute values, i.e., $|\xi| > \epsilon$, are selected. Here ϵ is certain small positive threshold.

In this paper, we adopt the AIC as $-2\log L(\xi_k)+2k$, where k is the number of selected parameters and $L(\xi_k)$ represent the maximum likelihood of samples. For L_1-norm, the threshold ϵ is searched in the interval from 10^{-2} to 10^{-4}, and it generally achieves good performance at 10^{-3}.

1) Experiment on artificial dataset - density estimation:

The artificial binary dataset is generated as follows: we first randomly select the target distribution $q(x)$, which is randomly chosen from the open probability simplex over the n random variables using the Jefferys prior [47]. Since the distribution over n binary variables belongs to the family of multinomial distribution, we adopt the Jefferys prior for the p-coordinates of $q(x)$, i.e., the Dirichlet distribution with all alpha parameters set to 0.5. Then, the dataset with N samples are generated from $q(x)$. For computation simplicity, the artificial dataset is set to be 10-dimensional. The CD learning algorithm is used to train the VBMs.

The Full-VBM, i.e., the VBM with full connections are used as baseline. The KL divergence is adopted to measure the performances of the VBMs trained by various algorithms. For each sample size N, 20 distributions are randomly generated and the averaged KL divergence is reported. Note that this experiment studies on the case that the variable number n=10, which is relatively small. Because it is convenient to evaluate the KL divergence analytically and hence study the proposed algorithms in more details. Changing the number of variables only offers a trivial influence for experimental results since we obtained similar observations on various variable numbers (not reported here).

Results and Summary: The averaged KL divergences between VBM and the real distribution are shown in Figure 4. We can see that all model selection methods could improve density estimation results of VBM, especially when the sample size is small (N=100 to 900). With relatively large samples, the effect of parameter reduction gradually becomes marginal.

For the two methods that use AIC criterion, CIF+AIC and StepwiseAIC achieve similar performance, as shown in Figure 4. The StepwiseAIC is a strong baseline since it extensively searches candidate VBM topologies in a stepwise approach. The result indicates that CIF could generally produce the set of connections which are also preferred in the StepwiseAIC. To show this, we make comparison between the topology of VBM selected by both methods in the first row of Figure 5. For all sample sizes (from 100 to 3000), we can see that the majority of preserved connections are shared for both methods. We also show how the AIC score changes when different number of connections are preserved in the second row of Figure 5. Both methods achieve optimal AIC performance at the (almost) same k. In terms of time complexity, CIF+AIC is a linear time algorithm while StepwiseAIC is $O(K^2)$.

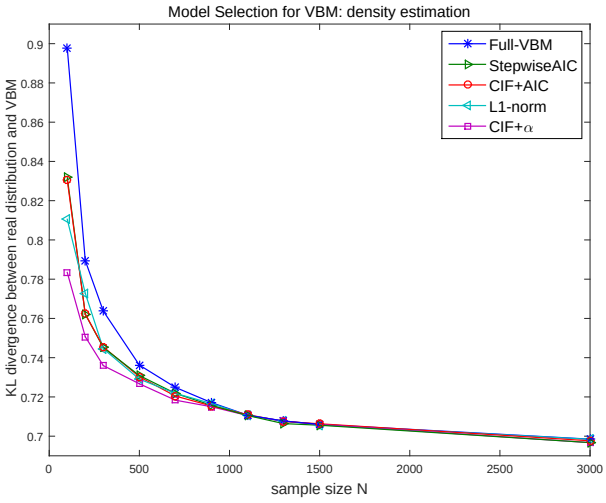

Fig. 4. The averaged density estimation results of VBM for artificial dataset

For the two methods that use thresholds to control the choice of parameters, CIF+α shows better performance than L_1-norm when sample size is small. We check the connection weights in L_1-norm, and our experiment shows a significant consistency between CIF+α and L_1-norm. That is, the parameters selected by CIF+α would also have higher significant absolute values in L_1-norm. This observation can be regarded as a mutual-confirmation of the CIF principle and the L_1-regularization method. Note that, there is a lack of an analytical method to adaptively set the threshold ϵ in L_1-norm for different sample sizes and parameters. That maybe the main reason that the more principled CIF+α achieves superior performance.

We also compare CIF and L_1-norm when the number of free parameters k is fixed. For CIF, we simply sort the confidence of connections in descend order, select the top k connections as free parameters and set the weight of residual connections to zero. For L_1-norm, we first train the VBM using L_1-regularization and than select the k connections with top absolute weights. The result is shown in Figure 6. We can see that the CIF outperforms L_1-norm on most model complexities k, which indicates the effectiveness of CIF.

For the two CIF-based method, CIF+α shows better performance than CIF+AIC when sample size N is small and gradually achieves similar performance along with the increasing N. We observe that the CIF+α tends to preserve less connections

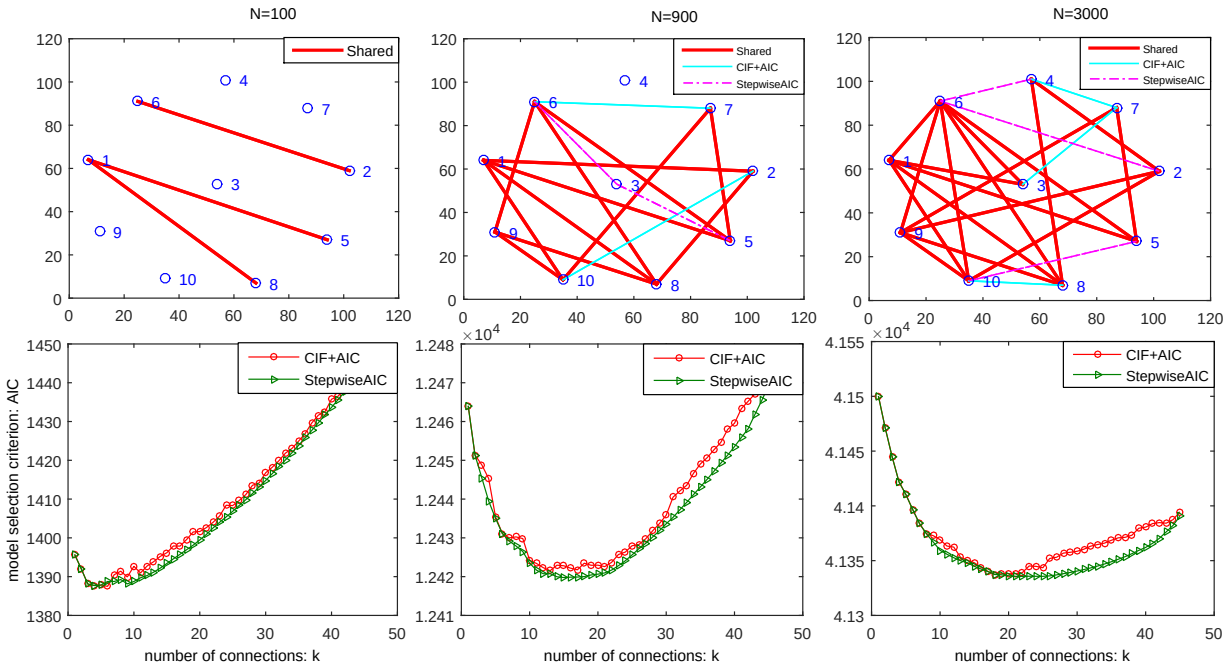

Fig. 5. Illustration on the behavior of CIF+AIC and StepwiseAIC for different sample sizes ($N = \{100, 900, 3000\}$). First row shows the topology of VBM selected by both methods; Second row compares the AIC criterion of both methods in different model complexities.

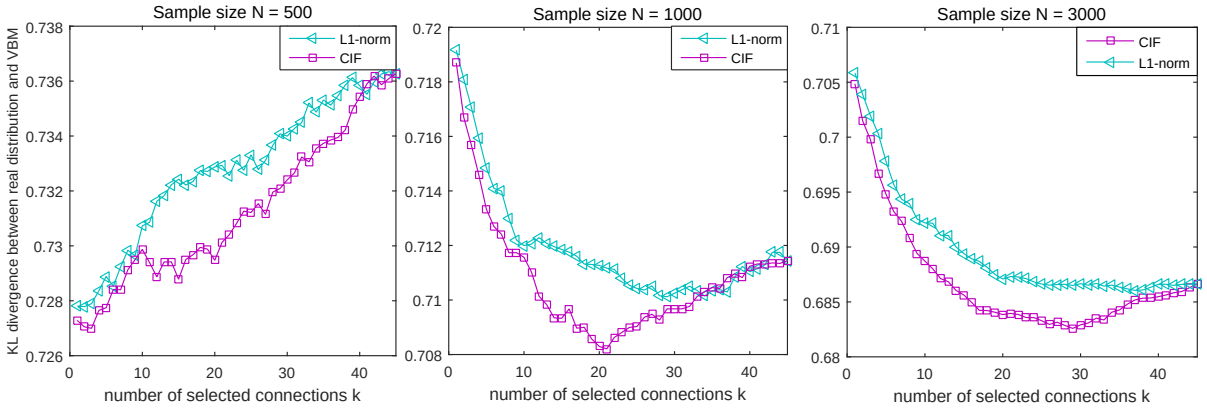

Fig. 6. The density estimation results of CIF and L_1-norm when the number of parameters is fixed for different sample sizes ($N = \{500, 1000, 3000\}$).

than CIF+AIC when the sample size is relatively small. This indicates that the AIC criterion may need to increase the penalty term for small sample sizes so as to achieve better performance, which will not be explored further in this paper.

2) Experiments on real datasets - information retrieval:

In this section, we empirically investigate how the CIF-based model selection algorithm works on real-world datasets in the context of information retrieval. In particular, we use the VBM to learn the underlying probability density of terms in the document [48], which is further used to rank document based on the query likelihood in information retrieval.

Experiments are conducted on three standard TREC collections: WSJ8792 (topics 151-200), ROBUST04 (topics 601-700) and WT10G (topics 501-550). The WSJ and ROBUST collections are relatively small and consist of news articles, science and technology reports and government documents, whereas WT10G is a larger Web collection. Collections are indexed by Indri 5.3 with Porter stemming and stopwords removed. For each topic, both the *title* field and *desc* field are used to generate queries.

We compare six document models: VBM (baseline), VBM+Smooth (use the smoothing approach in [48]), VBM+L_1-norm, VBM+StepwiseAIC, VBM+CIF (apply CIF+AIC or CIF+α). Mean average precision (MAP) is used as the evaluation metric, which is the mean of average precision scores over all the queries. The result is shown in Table II.

For short queries (*title* field), CIF-based VBM shows similar performance compare to the VBM without model selection. This is because short query usually contains only 1-4 terms

TABLE II
APPLY CIF ON DOCUMENT BOLTZMANN MACHINE IN IR

MAP	WSJ+*title*	ROBUST+*title*	WT10G+*title*
VBM	0.219	0.218	0.105
VBM+Smooth	**0.220**	**0.222**	0.108
L1-norm	0.214	0.220	0.108
StepwiseAIC	0.217	0.218	0.108
CIF+AIC	0.217	0.218	0.108
CIF+α	0.218	0.221	0.108
	WSJ+*desc*	ROBUST+*desc*	WT10G+*desc*
VBM	0.134	0.203	0.102
VBM+Smooth	0.150	0.217	0.101
L1-norm	0.155	0.239	0.101
StepwiseAIC	0.156	0.233	0.105
CIF+AIC	0.155	0.237	**0.108**
CIF+α	**0.159**	**0.242**	0.103

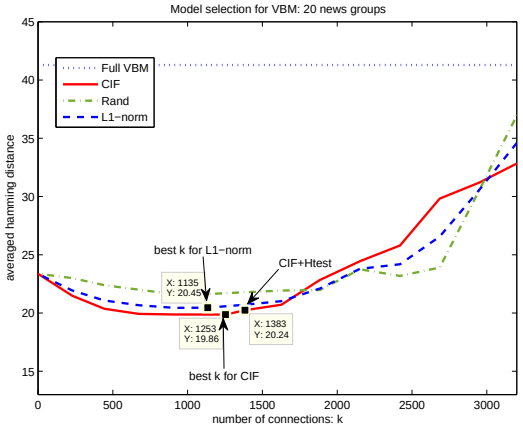

Fig. 7. Performance changes on real dataset w.r.t. number of free parameters

which leads to relatively simple VBMs w.r.t. training samples in one document, and there is no need to do model selection.

However, for long queries (*desc* field) with 8-16 terms, all model selection methods show significant improvements over VBM on WSJ and ROBUST04, except for WT10G. Note that the documents length of WT10G (about 380 words) are much longer than that of WSJ and ROBUST04 (about 250 words), leading to a larger training set. It maybe the main reason that the improvement becomes less significant on WT10G.

Comparing CIF+AIC with StepwiseAIC, CIF+AIC achieves similar or slightly better performance on all collections. This indicates that CIF could effectively determine the set of connections that are also preferred by StepwiseAIC. CIF+α outperforms L_1-norm on all collections, which shows that Algorithm 1 provides a reasonable way to balance between the model complexity and limited training samples.

3) Experiments on real datasets - density estimation:

In particular, we use the VBM to learn the underlying probability density over 100 terms of the *20 News Groups* binary dataset, with different model complexities. There are 18000 documents in 20 News Groups in total, which is partitioned into two set: train set (80%) and test set (20%). The learning rate for CD is manually tuned in order to converge properly and set to 0.01. Since it is infeasible to compute the KL devergence due to the high dimensionality, the averaged Hamming distance between the samples in the test dataset and those generated from VBM is used to evaluate the performance. Let $D = \{d_1, d_2, \ldots, d_N\}$ denote the dataset of N documents (each document d_i is a 100-dimensional binary vector). For the VBM to be evaluated, we first randomly generate N samples from its stationary distribution $p(x; \xi_{vbm})$, denoted as $V = \{v_1, v_2, \ldots, v_N\}$. Then the averaged hamming distance D_{ham} is calculated as follows:

$$D_{ham}[D, V] = \frac{\sum_{d_i \in D}(\min_{v_j \in V}(Ham[d_i, v_j])}{N}$$

where $Ham[d_i, v_j]$ is the number of positions where the corresponding values are different.

Due to high dimensionality, it is infeasible to compute the AIC criterion. Therefore, we compare four kinds of VBMs: 1) Full-VBM: the VBM with all connections; 2) Rand: randomly select k connections to build the VBM; 3) L_1-norm: k connections are selected in descend order based on their absolute values of weights trained with L_1-norm; 4) CIF: select the k connections in descend order based on their confidences (Equation 21). After training all VBMs on the training dataset, we evaluate the trained VBMs on the test dataset. We search for the best performances with respect to different model complexities, by varying k from 0 to 4950. We also mark the VBM that is automatically selected by CIF+α and the best k for L1-norm and CIF. The result is shown in Figure 7.

We can see that the parameter reduction criteria (CIF, L_1-norm, Rand) significantly improve the performance of the Full VBM, which indicates the existence of overfitting. Comparing the best performances, CIF significantly outperforms Rand and achieves slightly better performance than L_1-norm The performance of CIF+α is close to the optimal solution of CIF. These reflect the effectiveness of CIF to select suitable VBM with respect to given samples.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we studied the parametric reduction and model selection problem for Boltzmann machines from both theoretical and application perspectives, and proposed a Confident Information First (CIF) principle as a general framework for the parametric reduction to maximally preserve the confident parameters and ruling out less confident ones. On the theoretical side, we showed that CIF could lead to an optimal submanifold for binary multivariate distributions in term of Equation 1. Furthermore, we illustrated that the Boltzmann machines (with or without hidden units) can be derived from the general manifold based on the CIF principle. In the future, CIF could be the start of an information-oriented interpretation of deep learning models where BM is used as building blocks. For deep Boltzmann machine (DBM) [3], several layers of Restricted Boltzmann machines (RBM) compose a deep architecture in order to achieve a sufficiently abstract representation at a certain level. The CIF principle can be used to describe how the information flows in the transformations of representation across layers. Each layer of DBM determines a submanifold M of S, where M could maximally preserve the highly confident information on

parameters. Then the whole DBM can be seen as the process of repeatedly applying CIF in each layer, achieving a tradeoff between the abstractness of representation features and the intrinsic confidence of information preserved on parameters. The more detailed analysis and CIF-based designs on deep models will be left as future work.

On the application side, we proposed two sample-specific CIF-based model selection methods for BM, i.e., CIF+AIC and CIF+α, which can independently select model parameters, or integrate with information criterion, e.g., AIC, by providing a heuristic way to decide the priority order of parameters and improve the search efficiency without degrading the quality of model selection results. In the future, we plan to incorporate the CIF into deep learning models (e.g., DBM) to modify the network topology such that the most confident information in the data can be well captured.

REFERENCES

[1] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[2] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *J. Mach. Learn. Res.*, vol. 11, pp. 3371–3408, 2010.

[3] R. Salakhutdinov and G. E. Hinton, "An efficient learning procedure for deep boltzmann machines," *Neural Computing*, vol. 24, no. 8, pp. 1967–2006, 2012.

[4] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle, "Greedy layer-wise training of deep networks," in *NIPS'06*, Vancouver, British Columbia, Canada, 2006, pp. 153–160.

[5] M. Ranzato, C. Poultney, S. Chopra, and Y. LeCun, "Efficient learning of sparse representations with an energy-based model," in *NIPS'06*, 2006, pp. 1137–1144.

[6] S. Osindero and G. E. Hinton, "Modeling image patches with a directed hierarchy of markov random field," in *NIPS'07*, 2007, pp. 1121–1128.

[7] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *ICML'08*, 2008, pp. 160–167.

[8] R. Salakhutdinov and G. E. Hinton, "Using deep belief nets to learn covariance kernels for gaussian processes," in *NIPS'07*, 2007, pp. 1249–1256.

[9] R. Salakhutdinov and G. Hinton, "Semantic hashing," in *Workshop SIGIR'07*, 2007.

[10] H. Larochelle and Y. Bengio, "Classification using discriminative restricted boltzmann machines," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 536–543.

[11] T. Schmah, G. E. Hinton, S. L. Small, S. Strother, and R. S. Zemel, "Generative versus discriminative training of rbms for classification of fmri images," in *Advances in neural information processing systems*, 2008, pp. 1409–1416.

[12] J. Xu, H. He, and H. Man, "Dcpe co-training for classification," *Neurocomputing*, vol. 86, pp. 75–85, 2012.

[13] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" *Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.

[14] W. Zucchini, "An introduction to model selection," *Journal of mathematical psychology*, vol. 44, no. 1, pp. 41–61, 2000.

[15] E. J. Ward, "A review and comparison of four commonly used bayesian and maximum likelihood model selection tools," *Ecological Modelling*, vol. 211, no. 1, pp. 1–10, 2008.

[16] H. Akaike, "A new look at the statistical model identification," *Automatic Control, IEEE Transactions on*, vol. 19, no. 6, pp. 716–723, 1974.

[17] G. Schwarz, "Estimating the Dimension of a Model," *Annals of Statistics*, vol. 6, pp. 461–464, 1978.

[18] C. Bishop, "Pattern recognition and machine learning (information science and statistics)," 2007.

[19] Z. Lv, S. Luo, Y. Liu, and Y. Zheng, "Information geometry approach to the model selection of neural networks," in *Innovative Computing, Information and Control, 2006. ICICIC'06. First International Conference on*, vol. 3. IEEE, 2006, pp. 419–422.

[20] P. A. Murtaugh, "In defense of p values," *Ecology*, vol. 95, no. 3, pp. 611–617, 2014.

[21] J. Neyman and E. S. Pearson, "On the use and interpretation of certain test criteria for purposes of statistical inference: Part i," *Biometrika*, pp. 175–240, 1928.

[22] B. Efron, *Bootstrap methods: another look at the jackknife*. Springer, 1992.

[23] S. Arlot, A. Celisse *et al.*, "A survey of cross-validation procedures for model selection," *Statistics surveys*, vol. 4, pp. 40–79, 2010.

[24] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 267–288, 1996.

[25] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[26] K. P. Burnhan and D. R. Anderson, "Model selection and multimodel inference: a practical information-theoretic approach," *Journal of Wildlife Management*, vol. 67, no. 3, p. 606, 2002.

[27] B. D. Ripley, "Selecting amongst large classes of models," *Methods and models in statistics: In honor of Professor John Nelder, FRS*, pp. 155–170, 2004.

[28] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *The Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[29] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[30] I. Fodor, "A survey of dimension reduction techniques," Center for Applied Scientic Computing, Lawrence Livermore National Laboratory, United States, Tech. Rep., 2002.

[31] C. O. S. Sorzano, J. Vargas, and A. P. Montano, "A survey of dimensionality reduction techniques," *arXiv preprint arXiv:1403.2877*, 2014.

[32] S. Amari and H. Nagaoka, *Methods of Information Geometry*, ser. Translations of Mathematical Monographs. Oxford University Press, 1993.

[33] S. Amari, K. Kurata, and H. Nagaoka, "Information geometry of boltzmann machines," *IEEE Transactions on Neural Networks*, vol. 3, no. 2, pp. 260–271, 1992.

[34] P. Gibilisco, *Algebraic and geometric methods in statistics*. Cambridge University Press, 2010.

[35] N. N. Čencov, *Statistical decision rules and optimal inference*. American Mathematical Soc., 1982.

[36] C. R. Rao, "Information and the accuracy attainable in the estimation of statistical parameters," *Bulletin of calcutta mathematics society*, vol. 37, pp. 81–89, 1945.

[37] Y. Hou, X. Zhao, D. Song, and W. Li, "Mining pure high-order word associations via information geometry for information retrieval," *ACM TOIS*, vol. 31(3), 2013.

[38] R. E. Kass, "The geometry of asymptotic inference," *Statistical Science*, vol. 4, no. 3, pp. 188–219, 1989.

[39] P. Buhlmann and S. van de Geer, *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

[40] X. Zhao, Y. Hou, D. Song, and W. Li, "Extending the extreme physical information to universal cognitive models via a confident information first principle," *Entropy*, vol. 16, no. 7, pp. 3670–3688, 2014.

[41] D. H. Ackley, G. E. Hinton, and T. J. Sejnowski, "A learning algorithm for Boltzmann machines," *Cognitive Science*, vol. 9, pp. 147–169, 1985.

[42] M. A. Carreira-Perpinan and G. E. Hinton, "On contrastive divergence learning," *Artificial Intelligence and Statistics*, pp. 17–24, 2005.

[43] W. R. Gilks, S. Richardson, and D. Spiegelhalter, *Markov Chain Monte Carlo in Practice*. Chapman and Hall/CRC, 1996.

[44] S. Amari, "Information geometry of the em and em algorithms for neural networks," *Neural networks*, vol. 8, no. 9, pp. 1379–1408, 1995.

[45] X. Zhao, Y. Hou, Q. Yu, D. Song, and W. Li, "Understanding deep learning by revisiting boltzmann machines: An information geometry approach," *CoRR*, vol. abs/1302.3931, 2013.

[46] H. Nakahara and S. Amari, "Information geometric measure for neural spikes," *Neural Computation*, vol. 14, pp. 2269–2316, 2002.

[47] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[48] Q. Yu, P. Zhang, Y. Hou, D. Song, and J. Wang, "Document boltzmann machines for information retrieval," in *ECIR*, 2015, pp. 666–671.