Entropy **2014**, *16*, 3670-3688; doi:10.3390/e16073670

OPEN ACCESS

entropy

ISSN 1099-4300

www.mdpi.com/journal/entropy

Article

Extending the Extreme Physical Information to Universal Cognitive Models via a Confident Information First Principle

Xiaozhao Zhao [1], Yuexian Hou [1,2,*], Dawei Song [1,3] and Wenjie Li [2]

[1] School of Computer Science and Technology, Tianjin University, Tianjin 300072, China; E-Mails: 0.25eye@gmail.com (X.Z.); dawei.song2010@gmail.com (D.S.)

[2] Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong, China; E-Mail: cswjli@comp.polyu.edu.hk

[3] Department of Computing and Communications, The Open University, Milton Keynes MK76AA, UK

* Author to whom correspondence should be addressed; E-Mail: yxhou@tju.edu.cn; Tel.: +86-022-27406538.

Received: 25 March 2014; in revised form: 6 June 2014 / Accepted: 20 June 2014 / Published: 1 July 2014

Abstract: The principle of extreme physical information (EPI) can be used to derive many known laws and distributions in theoretical physics by extremizing the physical information loss K, *i.e.*, the difference between the observed Fisher information I and the intrinsic information bound J of the physical phenomenon being measured. However, for complex cognitive systems of high dimensionality (e.g., human language processing and image recognition), the information bound J could be excessively larger than I ($J \gg I$), due to insufficient observation, which would lead to serious over-fitting problems in the derivation of cognitive models. Moreover, there is a lack of an established exact invariance principle that gives rise to the bound information in universal cognitive systems. This limits the direct application of EPI. To narrow down the gap between I and J, in this paper, we propose a confident-information-first (CIF) principle to lower the information bound J by preserving confident parameters and ruling out unreliable or noisy parameters in the probability density function being measured. The confidence of each parameter can be assessed by its contribution to the expected Fisher information distance between the physical phenomenon and its observations. In addition, given a specific parametric representation, this contribution can often be directly assessed by the Fisher information, which establishes a connection with the inverse variance of any unbiased estimate for the parameter via the Cramér–Rao bound. We then consider the dimensionality reduction in the parameter spaces of binary multivariate distributions. We show that the single-layer

Boltzmann machine without hidden units (SBM) can be derived using the CIF principle. An illustrative experiment is conducted to show how the CIF principle improves the density estimation performance.

Keywords: information geometry; Boltzmann machine; Fisher information; parametric reduction

1. Introduction

Information has been found to play an increasingly important role in physics. As stated in Wheeler [1]: "All things physical are information-theoretic in origin and this is a participatory universe...Observer participancy gives rise to information; and information gives rise to physics". Following this viewpoint, Frieden [2] unifies the derivation of physical laws in major fields of physics, from the Dirac equation to the Maxwell-Boltzmann velocity dispersion law, using the extreme physical information principle (EPI). More specifically, a variety of equations and distributions can be derived by extremizing the physical information loss K, *i.e.*, the difference between the observed Fisher information I and the intrinsic information bound J of the physical phenomenon being measured.

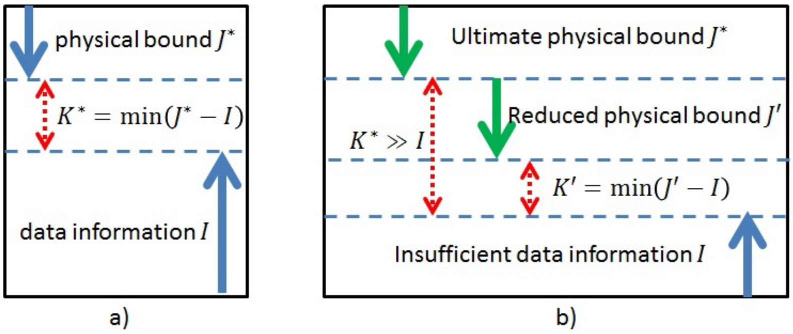
The first quantity, I, measures the amount of information as a finite scalar implied by the data with some suitable measure [2]. It is formally defined as the trace of the Fisher information matrix [3]. In addition to I, the second quantity, the information bound J, is an invariant that characterizes the information that is intrinsic to the physical phenomenon [2]. During the measurement procedure, there may be some loss of information, which entails $I = \kappa J$, where $\kappa \leq 1$ is called the efficiency coefficient of the EPI process in transferring the Fisher information from the phenomenon (specified by J) to the output (specified by I). For closed physical systems, in particular, any solution for I attains some fraction of J between $1/2$ (for classical physics) and one (for quantum physics) [4].

However, it is usually not the case in cognitive science. For complex cognitive systems (e.g., human language processing and image recognition), the target probability density function (pdf) being measured is often of high dimensionality (e.g., thousands of words in a human language vocabulary and millions of pixels in an observed image). Thus, it is infeasible for us to obtain a sufficient collection of observations, leading to excessive information loss between the observer and nature. Moreover, there is a lack of an established exact invariance principle that gives rise to the bound information in universal cognitive systems. This limits the direct application of EPI in cognitive systems.

In terms of statistics and machine learning, the excessive information loss between the observer and nature will lead to serious over-fitting problems, since the insufficient observations may not provide necessary information to reasonably identify the model and support the estimation of the target pdf in complex cognitive systems. Actually, a similar problem is also recognized in statistics and machine learning, known as the model selection problem [5]. In general, we would require a complex model with a high-dimensional parameter space to sufficiently depict the original high-dimensional observations. However, over-fitting usually occurs when the model is excessively complex with respect to the given observations. To avoid over-fitting, we would need to adjust the complexity of the models to the available amount of observations and, equivalently, to adjust the information bound J corresponding to the observed information I.

In order to derive feasible computational models for cognitive phenomenon, we propose a confident-information-first (CIF) principle in addition to EPI to narrow down the gap between I and J (thus, a reasonable efficiency coefficient κ is implied), as illustrated in Figure 1. However, we do not intend to actually derive the distribution laws by solving the differential equations of the extremization of the new information loss K'. Instead, we assume that the target distribution belongs to some general multivariate binary distribution family and focus on the problem of seeking a proper information bound with respect to the constraint of the parametric number and the given observations.

Figure 1. (**a**) The paradigm of the extreme physical information principle (EPI) to derive physical laws by the extremization of the information loss K^* ($K^* = J/2$ for classical physics and $K^* = 0$ for quantum physics); (**b**) the paradigm of confident-information-first (CIF) to derive computational models by reducing the information loss K' using a new physical bound J'.



The key to the CIF approach is how to systematically reduce the physical information bound for high-dimensional complex systems. As stated in Frieden [2], the information bound J is a functional form that depends upon the physical parameters of the system. The information is contained in the variations of the observations (often imperfect, due to insufficient sampling, noise and intrinsic limitations of the "observer"), and can be further quantified using the Fisher information of system parameters (or coordinates) [3] from the estimation theory. Therefore, the physical information bound J of a complex system can be reduced by transforming it to a simpler system using some parametric reduction approach. Assuming there exists an ideal parametric model S that is general enough to represent all system phenomena (which gives the ultimate information bound in Figure 1), our goal is to adopt a parametric reduction procedure to derive a lower-dimensional sub-model M (which gives the reduced information bound in Figure 1) for a given dataset (usually insufficient or perturbed by noises) by reducing the number of free parameters in S.

Formally speaking, let $q(\xi)$ be the ideal distribution with parameters ξ that describes the physical system and $q(\xi + \Delta\xi)$ be the observations of the system with some small fluctuation $\Delta\xi$ in parameters. In [6], the averaged information distance $I(\Delta\xi)$ between the distribution and its observations, the so-called shift information, is used as a disorder measure of the fluctuated observations to reinterpret the EPI principle. More specifically, in the framework of information geometry, this information distance could also be assessed using the Fisher information distance induced by the Fisher–Rao metric, which can be decomposed into the variation in the direction of each system parameter [7]. In principle, it is possible to divide system parameters into two categories, *i.e.*, the parameters with notable variations and the parameters with negligible variations, according to their contributions to the whole information distance. Additionally, the parameters with notable contributions are considered to be confident, since they are important for reliably distinguishing the ideal distribution from its observation distributions. On the other hand, the parameters with negligible contributions can be considered to be unreliable or noisy. Then, the CIF principle can be stated as the parameter selection criterion that maximally preserves the Fisher information distance in an expected sense with respect to the constraint of the parametric number and the given observations (if available), when projecting distributions from the parameter space of S into that of the reduced sub-model M. We call it the distance-based CIF. As a result, we could manipulate the information bound of the underlying system by preserving the information of confident parameters and ruling out noisy parameters.

In this paper, the CIF principle is analyzed in the multivariate binary distribution family in the mixed-coordinate system [8]. It turns out that, in this problematic configuration, the confidence of a parameter can be directly evaluated by its Fisher information, which also establishes a connection with the inverse variance of any unbiased estimate for the parameter via the Cramér–Rao bound [3]. Hence, the CIF principle can also be interpreted as the parameter selection procedure that keeps the parameters with reliable estimates and rules out unreliable or noisy parameters. This CIF is called the information-based CIF. Note that the definition of confidence in distance-based CIF depends on both Fisher information and the scale of fluctuation, and the confidence in the information-based CIF (*i.e.*, Fisher information) can be seen as a special case of confidence measure with respect to certain coordinate systems. This simplification allows us to further apply the CIF principle to improve existing learning algorithms for the Boltzmann machine.

The paper is organized as follows. In Section 2, we introduce the parametric formulation for the general multivariate binary distributions in terms of information geometry (IG) framework [7]. Then, Section 3 describes the implementation details of the CIF principle. We also give a geometric interpretation of CIF by showing that it can maximally preserve the expected information distance (in Section 3.2.1), as well as the analysis on the scale of the information distance in each individual system parameter (in Section 3.2.2). In Section 4, we demonstrate that a widely used cognitive model, *i.e.*, the Boltzmann machine, can be derived using the CIF principle. Additionally, an illustrative experiment is conducted to show how the CIF principle can be utilized to improve the density estimation performance of the Boltzmann machine in Section 5.

2. The Multivariate Binary Distributions

Similar to EPI, the derivation of CIF depends on the analysis of the physical information bound, where the choice of system parameters, also called "Fisher coordinates" in Frieden [2], is crucial. Based on information geometry (IG) [7], we introduce some choices of parameterizations for binary multivariate distributions (denoted as statistical manifold S) with a given number of variables n, *i.e.*, the open simplex of all probability distributions over binary vector $x \in \{0, 1\}^n$.

2.1. Notations for Manifold S

In IG, a family of probability distributions is considered as a differentiable manifold with certain parametric coordinate systems. In the case of binary multivariate distributions, four basic coordinate systems are often used: p-coordinates, η-coordinates, θ-coordinates and mixed-coordinates [7,9]. Mixed-coordinates is of vital importance for our analysis.

For the p-coordinates $[p]$ with n binary variables, the probability distribution over 2^n states of x can be completely specified by any $2^n - 1$ positive numbers indicating the probability of the corresponding exclusive states on n binary variables. For example, the p-coordinates of $n = 2$ variables could be $[p] = (p_{01}, p_{10}, p_{11})$. Note that IG requires all probability terms to be positive [7].

For simplicity, we use the capital letters $I, J, \ldots$ to index the coordinate parameters of probabilistic distribution. To distinguish the notation of Fisher information (conventionally used in literature, e.g., data information I and information bound J in Section 1) from the coordinate indexes, we make explicit explanations when necessary from now on. An index I can be regarded as a subset of $\{1, 2, \ldots, n\}$. Additionally, p_I stands for the probability that all variables indicated by I equal to one and the complemented variables are zero. For example, if $I = \{1, 2, 4\}$ and $n = 4$, then $p_I = p_{1101} = Prob(x_1 = 1, x_2 = 1, x_3 = 0, x_4 = 1)$. Note that the null set can also be a legal index of the p-coordinates, which indicates the probability that all variables are zero, denoted as $p_{0\ldots0}$.

Another coordinate system often used in IG is η-coordinates, which is defined by:

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

where the value of X_I is given by $\prod_{i \in I} x_i$ and the expectation is taken with respect to the probability distribution over x. Grouping the coordinates by their orders, the η-coordinate system is denoted as $[\eta] = (\eta_i^1, \eta_{ij}^2, \ldots, \eta_{1,2\ldots n}^n)$, where the superscript indicates 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) are defined by:

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

where $\psi(\theta) = \log(\sum_x exp\{\sum_I \theta^I X_I(x)\})$ is the cumulant generating function and its value equals to $-\log Prob\{x_i = 0, \forall i \in \{1, 2, ..., n\}\}$. The θ-coordinate is denoted as $[\theta] = (\theta_1^i, \theta_2^{ij}, \ldots, \theta_n^{1,\ldots,n})$, where the subscript indicates the order number of the corresponding parameter. Note that the order indices locate at different positions in $[\eta]$ and $[\theta]$ following the convention in Amari *et al.* [8].

The relation between coordinate systems $[\eta]$ and $[\theta]$ is bijective. More formally, they are connected by the Legendre transformation:

$$\theta^I = \frac{\partial \phi(\eta)}{\partial \eta_I}, \eta_I = \frac{\partial \psi(\theta)}{\partial \theta^I} \tag{3}$$

where $\psi(\theta)$ is given in Equation (2) and $\phi(\eta) = \sum_x p(x; \eta) \log p(x; \eta)$ is the negative of entropy. It can be shown that $\psi(\theta)$ and $\phi(\eta)$ meet the following identity [7]:

$$\psi(\theta) + \phi(\eta) - \sum \theta^I \eta_I = 0 \tag{4}$$

Next, we introduce mixed-coordinates, which is important for our derivation of CIF. In general, the manifold S of probability distributions could be represented by the l-mixed-coordinates [8]:

$$[\zeta]_l = (\eta_i^1, \eta_{ij}^2, \dots, \eta_{i,j,\dots,k}^l, \theta_{l+1}^{i,j,\dots,k}, \dots, \theta_n^{1,\dots,n}) \tag{5}$$

where the first part consists of η-coordinates with order less or equal to l (denoted by $[\eta^{l-}]$) and the second part consists of θ-coordinates with order greater than l (denoted by $[\theta_{l+}]$), $l \in \{1, ..., n-1\}$.

2.2. 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 for $[\xi]$ (denoted by G_ξ) is defined as the covariance of the scores of $[\xi_i]$ and $[\xi_j]$ [3], *i.e.*,

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

under the regularity condition for the pdf that the partial derivatives exist. The Fisher information measures the amount of information in the data that a statistic carries about the unknown parameters [10]. The Fisher information matrix is of vital importance to our analysis, because the inverse of Fisher information matrix gives an asymptotically tight lower bound to the covariance matrix of any unbiased estimate for the considered parameters [3]. 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.

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 $g^{IJ} = \frac{\partial^2 \phi(\eta)}{\partial \eta_I \partial \eta_J}$ [7]. Let $G_\theta = (g_{IJ})$ and $G_\eta = (g^{IJ})$ be the Fisher information matrices for $[\theta]$ and $[\eta]$, respectively. It can be shown that G_θ and G_η are mutually inverse matrices, *i.e.*, $\sum_J g^{IJ} g_{JK} = \delta_K^I$, where $\delta_K^I = 1$ if $I = K$ and zero otherwise [7]. In order to generally compute G_θ and G_η, we develop the following Propositions 1 and 2. Note that Proposition 1 is a generalization of Theorem 2 in Amari *et al.* [8].

Proposition 1. *The Fisher information between two parameters θ^I and θ^J in $[\theta]$, is given by:*

$$g_{IJ}(\theta) = \eta_{I \bigcup J} - \eta_I \eta_J \tag{6}$$

Proof. in Appendix A. □

Proposition 2. *The Fisher information between two parameters η_I and η_J in $[\eta]$, is given by:*

$$g^{IJ}(\eta) = \sum_{K \subseteq I \cap J} (-1)^{|I-K|+|J-K|} \cdot \frac{1}{p_K} \tag{7}$$

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

Proof. in Appendix B. □

Based on the Fisher information matrices G_η and G_θ, we can calculate the Fisher information matrix G_ζ for the l-mixed-coordinate system $[\zeta]_l$, as follows:

Proposition 3. *The Fisher information matrix G_ζ of the l-mixed-coordinates $[\zeta]_l$ is given by:*

$$G_\zeta = \begin{pmatrix} A & 0 \\ 0 & B \end{pmatrix} \tag{8}$$

where $A = ((G_\eta^{-1})_{I_\eta})^{-1}$, $B = ((G_\theta^{-1})_{J_\theta})^{-1}$, 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., $\{\eta_i^1, ..., \eta_{i,j,...,k}^l\}$, and J_θ is the index set of the parameters shared by $[\theta]$ and $[\zeta]_l$, i.e., $\{\theta_{l+1}^{i,j,...,k}, \dots, \theta_n^{1,...,n}\}$.

Proof. in Appendix C. □

3. The General CIF Principle

In this section, we propose the CIF principle to reduce the physical information bound for high-dimensionality systems. Given a target distribution $q(x) \in S$, we consider the problem of realizing it by a lower-dimensionality submanifold. This is defined as the problem of parametric reduction for multivariate binary distributions. The family of multivariate binary distributions has been proven to be useful when we deal with discrete data in a variety of applications in statistical machine learning and artificial intelligence, such as the Boltzmann machine in neural networks [11,12] and the Rasch model in human sciences [13,14].

Intuitively, if we can construct a coordinate system so that the confidences of its parameters entail a natural hierarchy, in which high confident parameters are significantly distinguished from and orthogonal to lowly confident ones, then we can conveniently implement CIF by keeping the high confident parameters unchanged and setting the lowly confident parameters to neutral values. Therefore, the choice of coordinates (or parametric representations) in CIF is crucial to its usage. This strategy is infeasible in terms of p-coordinates, η-coordinates or θ-coordinates, since the orthogonality condition cannot hold in these coordinate systems. In this section, we will show that the l-mixed-coordinates $[\zeta]_l$ meets the requirement of CIF.

In principle, the confidence of parameters should be assessed according to their contributions to the expected information distance between the ideal distribution and its fluctuated observations. This is called the distance-based CIF (see Section 1). For some coordinated systems, e.g., the mixed-coordinate system $[\zeta]_l$, the confidence of a parameter can also be directly evaluated by its Fisher information. This is called the information-based CIF (see Section 1). The information-based CIF (*i.e.*, Fisher information) can be seen as an approximation to distance-based CIF, since it neglects the influence of parameter scaling to the expected information distance. However, considering the standard mixed-coordinates $[\zeta]_l$ for the manifold of multivariate binary distributions, it turns out that both distance-based CIF and information-based CIF entail the same submanifold M (refer to Section 3.2 for detailed reasons).

For the purpose of legibility, we will start with the information-based CIF, where the parameter's confidence is simply measured using its Fisher information. After that, we show that the information-based CIF leads to an optimal submanifold M, which is also optimal in terms of the more rigorous distance-based CIF.

3.1. The Information-Based CIF Principle

In this section, we will show that the l-mixed-coordinates $[\zeta]_l$ meet the requirement of the information-based CIF. According to Proposition 3 and the following Proposition 4, the confidences of coordinate parameters (measured by Fisher information) in $[\zeta]_l$ entail a natural hierarchy: the first part of high confident parameters $[\eta^{l-}]$ are separated from the second part of low confident parameters $[\theta_{l+}]$. Additionally, those low confident parameters $[\theta_{l+}]$ have the neutral value of zero.

Proposition 4. *The diagonal elements of A are lower bounded by one, and those of B are upper bounded by one.*

Proof. in Appendix D. □

Moreover, the parameters in $[\eta^{l-}]$ are orthogonal to the ones in $[\theta_{l+}]$, indicating that we could estimate these two parts independently [9]. Hence, we can implement the information-based CIF for parametric reduction in $[\zeta]_l$ by replacing low confident parameters with neutral value zero and reconstructing the resulting distribution. It turns out that the submanifold of S tailored by information-based CIF becomes $[\zeta]_{l_t} = (\eta^1_i, ..., \eta^l_{ij...k}, 0, \ldots, 0)$. We call $[\zeta]_{l_t}$ the l-tailored-mixed-coordinates.

To grasp an intuitive picture for the CIF strategy and its significance w.r.t mixed-coordinates, let us consider an example with $[p] = (p_{001} = 0.15, p_{010} = 0.1, p_{011} = 0.05, p_{100} = 0.2, p_{101} = 0.1, p_{110} = 0.05, p_{111} = 0.3)$. Then, the confidences for coordinates in $[\eta]$, $[\theta]$ and $[\zeta]_2$ are given by the diagonal elements of the corresponding Fisher information matrices. Applying the two-tailored CIF in mixed-coordinates, the loss ratio of Fisher information is 0.001%, and the ratio of the Fisher information of the tailored parameter (θ^{123}_3) to the remaining η parameter with the smallest Fisher information is 0.06%. On the other hand, the above two ratios become 7.58% and 94.45% (in η-coordinates) or 12.94% and 92.31% (in θ-coordinates), respectively. We can see that $[\zeta]_2$ gives us a much better way to tell apart confident parameters from noisy ones.

3.2. The Distance-Based CIF: A Geometric Point-of-View

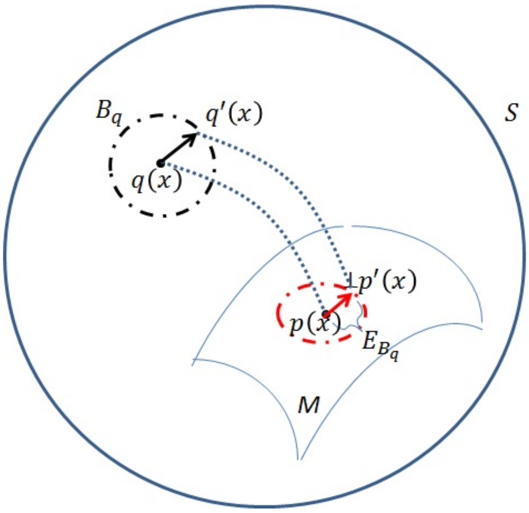
In the previous section, the information-based CIF entails a submanifold of S determined by the l-tailored-mixed-coordinates $[\zeta]_{l_t}$. A more rigorous definition for the confidence of coordinates is the distance-based confidence used in the distance-based CIF, which relies on both of the coordinate's Fisher information and its fluctuation scaling. In this section, we will show that the the submanifold M determined by $[\zeta]_{l_t}$ is also an optimal submanifold M in terms of the distance-based CIF. Note that, for other coordinate systems (e.g., arbitrarily rescaling coordinates), the information-based CIF may not entail the same submanifold as the distance-based CIF.

Let $q(x)$, with coordinate ζ_q, denote the exact solution to the physical phenomenon being measured. Additionally, the act of observation would cause small random perturbations to $q(x)$, leading to some observation $q'(x)$ with coordinate $\zeta_q + \Delta\zeta_q$. When two distributions $q(x)$ and $q'(x)$ are close, the divergence between $q(x)$ and $q'(x)$ on manifold S could be assessed by the Fisher information distance: $D(q, q') = (\Delta\zeta_q \cdot G_\zeta \cdot \Delta\zeta_q)^{1/2}$, where G_ζ is the Fisher information matrix and the perturbation $\Delta\zeta_q$ is small. The Fisher information distance between two close distributions $q(x)$ and $q'(x)$ on manifold S is the Riemannian distance under the Fisher–Rao metric, which is shown to be the square root of the twice of the Kullback–Leibler divergence from $q(x)$ to $q'(x)$ [8]. Note that we adopt the Fisher information distance as the distance measure between two close distributions, since it is shown to be the unique metric meeting a set of natural axioms for the distribution metrics [7,15,16], e.g., the invariant property with respect to reparametrizations and the monotonicity with respect to the random maps on variables.

Let M be a smooth k-dimensionality submanifold in S ($k < 2^n - 1$). Given the point $q(x) \in S$, the projection [8] of $q(x)$ on M is the point $p(x)$ that belongs to M and is closest to $q(x)$ with respect to the Kullback–Leibler divergence (K-L divergence) [17] from the distribution $q(x)$ to $p(x)$. On the submanifold M, the projections of $q(x)$ and $q'(x)$ are $p(x)$ and $p'(x)$, with coordinates ζ_p and $\zeta_p + \Delta\zeta_p$, respectively, shown in Figure 2.

Let the preserved Fisher information distance be $D(p, p')$ after projecting on M. In order to retain the information contained in observations, we need the ratio $\frac{D(p,p')}{D(q,q')}$ to be as large as possible in the expected sense, with respect to the given dimensionality k of M. The next two sections will illustrate that CIF leads to an optimal submanifold M based on different assumptions on the perturbations $\Delta\zeta_q$.

Figure 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.



3.2.1. Perturbations in Uniform Neighborhood

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 K-L divergence and ε is small. Additionally, $q'(x)$ is a neighbor of $q(x)$ uniformly sampled on B_q, as illustrated in Figure 2. Recall that, for a small ε, the K-L divergence can be approximated by half of the squared Fisher information distance. Thus, in the parameterization of $[\zeta]_l$, B_q is indeed the surface of a hyper-ellipsoid (centered at $q(x)$) determined by G_ζ. The following proposition shows that the general CIF would lead to an optimal submanifold M that maximally preserves the expected information distance, where the expectation is taken upon the uniform neighborhood, B_q.

Proposition 5. *Consider the manifold S in l-mixed-coordinates $[\zeta]_l$. Let k be the number of free parameters in the l-tailored-mixed-coordinates $[\zeta]_{l_t}$. Then, among all k-dimensional submanifolds of S,* *the submanifold determined by* $[\zeta]_{l_t}$ *can maximally preserve the expected information distance induced by the Fisher–Rao metric.*

Proof. in Appendix E. □

3.2.2. Perturbations in Typical Distributions

To facilitate our analysis, we make a basic assumption on the underlying distributions $q(x)$ that at least $(2^n - 2^{n/2})$ p-coordinates are of the scale ϵ, where ϵ is a sufficiently small value. Thus, residual p-coordinates (at most $2^{n/2}$) are all significantly larger than zero (of scale $\Theta(1/2^{(n/2)})$), and their sum approximates one. Note that these assumptions are common situations in real-world data collections [18], since the frequent (or meaningful) patterns are only a small fraction of all of the system states.

Next, we introduce a small perturbation Δp to the p-coordinates $[p]$ for the ideal distribution $q(x)$. The scale of each fluctuation Δp_I is assumed to be proportional to the standard variation of corresponding p-coordinate p_I by some small coefficients (upper bounded by a constant a), which can be approximated by the inverse of the square root of its Fisher information via the Cramér–Rao bound. It turns out that we can assume the perturbation Δp_I to be $a\sqrt{p_I}$.

In this section, we adopt the l-mixed-coordinates $[\zeta]_l = (\eta^{l-}; \theta_{l+})$, where $l = 2$ is used in the following analysis. Let $\Delta\zeta_q = (\Delta\eta^{2-}; \Delta\theta_{2+})$ be the incremental of mixed-coordinates after the perturbation. The squared Fisher information distance $D^2(p, p') = \Delta\zeta_q \cdot G_\zeta \cdot \Delta\zeta_q$ could be decomposed into the direction of each coordinate in $[\zeta]_l$. We will clarify that, under typical cases, the scale of the Fisher information distance in each coordinate of θ_{l+} (reduced by CIF) is asymptotically negligible, compared to that in each coordinate of η^{l-} (preserved by CIF).

The scale of squared Fisher information distance in the direction of η_I is proportional to $\Delta\eta_I \cdot (G_\zeta)_{I,I} \cdot \Delta\eta_I$, where $(G_\zeta)_{I,I}$ is the Fisher information of η_I in terms of the mixed-coordinates $[\zeta]_2$. From Equation (1), for any I of order one (or two), η_I is the sum of 2^{n-1} (or 2^{n-2}) p-coordinates, and the scale is $\Theta(1)$. Hence, the incremental $\Delta\eta^{2-}$ is proportional to $\Theta(1)$, denoted as $a \cdot \Theta(1)$. It is difficult to give an explicit expression of $(G_\zeta)_{I,I}$ analytically. However, the Fisher information $(G_\zeta)_{I,I}$ of η_I is bounded by the (I, I)-th element of the inverse covariance matrix [19], which is exactly $1/g^{I,I}(\theta) = \frac{1}{\eta_I - \eta_I^2}$ (see Proposition 3). Hence, the scale of $(G_\zeta)_{I,I}$ is also $\Theta(1)$. It turns out that the scale of squared Fisher information distance in the direction of η_I is $a^2 \cdot \Theta(1)$.

Similarly, for the part θ_{2+}, the scale of squared Fisher information distance in the direction of θ^J is proportional to $\Delta\theta^J \cdot (G_\zeta)_{J,J} \cdot \Delta\theta^J$, where $(G_\zeta)_{J,J}$ is the Fisher information of θ^J in terms of the mixed-coordinates $[\zeta]_2$. The scale of θ^J is maximally $f(k)|log(\sqrt{\epsilon})|$ based on Equation (2), where k is the order of θ^J and $f(k)$ is the number of p-coordinates of scale $\Theta(1/2^{(n/2)})$ that are involved in the calculation of θ^J. Since we assume that $f(k) \leq 2^{(n/2)}$, the maximum scale of θ^J is $2^{(n/2)}|log(\sqrt{\epsilon})|$. Thus, the incremental $\Delta\theta^J$ is of a scale bounded by $a \cdot 2^{(n/2)}|log(\sqrt{\epsilon})|$. Similar to our previous deviation, the Fisher information $(G_\zeta)_{J,J}$ of θ^J is bounded by the (J, J)-th element of the inverse covariance matrix, which is exactly $1/g_{J,J}(\eta)$ (see Proposition 3). Hence, the scale of $(G_\zeta)_{J,J}$ is $(2^k - f(k))^{-1}\epsilon$. In summary, the scale of squared Fisher information distance in the direction of θ^J is bounded by the scale of $a^2 \cdot$

$\Theta(2^n\epsilon\frac{|log(\sqrt{\epsilon})|^2}{2^k-f(k)})$. Since ϵ is a sufficiently small value and a is constant, the scale of squared Fisher information distance in the direction of θ^J is asymptotically zero.

In summary, in terms of modeling the fluctuated observations of typical cognitive systems, the original Fisher information distance between the physical phenomenon ($q(x)$) and observations ($q'(x)$) is systematically reduced using CIF by projecting them on an optimal submanifold M. Based on our above analysis, the scale of Fisher information distance in the directions of $[\eta^{l-}]$ preserved by CIF is significantly larger than that of the directions $[\theta_{l+}]$ reduced by CIF.

4. Derivation of Boltzmann Machine by CIF

In the previous section, the CIF principle is uncovered in the $[\zeta]_l$ coordinates. Now, we consider an implementation of CIF when l equals to two, which gives rise to the single-layer Boltzmann machine without hidden units (SBM).

4.1. Notations for SBM

The energy function for SBM is given by:

$$E_{SBM}(x;\xi) = -\frac{1}{2}x^TUx - b^Tx \tag{9}$$

where $\xi = \{U, b\}$ are the parameters and the diagonals of U are set to zero. The Boltzmann distribution over x is $p(x;\xi) = \frac{1}{Z}exp\{-E_{SBM}(x;\xi)\}$, where Z is a normalization factor. Actually, the parametrization for SBM could be naturally expressed by the coordinate systems in IG (e.g., $[\theta] = (\theta_1^i = b_i, \theta_2^{ij} = U_{ij}, \theta_3^{ijk} = 0, ..., \theta_n^{1,2,...,n} = 0)$).

4.2. The Derivation of SBM using CIF

Given any underlying probability distribution $q(x)$ on the general manifold S over $\{x\}$, the logarithm of $q(x)$ can be represented by a linear decomposition of θ-coordinates, as shown in Equation (2). Since it is impractical to recognize all coordinates for the target distribution, we would like to only approximate part of them and end up with a k-dimensional submanifold M of S, where k ($\ll 2^n - 1$) is the number of free parameters. Here, we set k to be the same dimensionality as SBM, *i.e.*, $k = \frac{n(n+1)}{2}$, so that all candidate submanifolds are comparable to the submanifold endowed by SBM (denoted as M_{sbm}). Next, the rationale underlying the design of M_{sbm} can be illustrated using the general CIF.

Let the two-mixed-coordinates of $q(x)$ on S be $[\zeta]_2 = (\eta_i^1, \eta_{ij}^2, \theta_3^{i,j,k}, \ldots, \theta_n^{1,\ldots,n})$. Applying the general CIF on $[\zeta]_2$, our parametric reduction rule is to preserve the high confident part parameters $[\eta^{2-}]$ and replace low confident parameters $[\theta_{2+}]$ by a fixed neutral value of zero. Thus, we derive the two-tailored-mixed-coordinates: $[\zeta]_{2_t} = (\eta_i^1, \eta_{ij}^2, 0, \ldots, 0)$, as the optimal approximation of $q(x)$ by the k-dimensional submanifolds. On the other hand, given the two-mixed-coordinates of $q(x)$, the projection $p(x) \in M_{sbm}$ of $q(x)$ is proven to be $[\zeta]_p = (\eta_i^1, \eta_{ij}^2, 0, \ldots, 0)$ [8]. Thus, SBM defines a probabilistic parameter space that is derived from CIF.

4.3. The Learning Algorithms for SBM

Let $q(x)$ be the underlying probability distribution from which samples $D = \{d_1, d_2, \ldots, d_N\}$ are generated independently. Then, our goal is to train an SBM (with stationary probability $p(x)$) based on D that realizes $q(x)$ as faithfully as possible. Here, we briefly introduce two typical learning algorithms for SBM: maximum-likelihood and contrastive divergence [11,20,21].

Maximum-likelihood (ML) learning realizes a gradient ascent of log-likelihood of D:

$$\Delta U_{ij} = \varepsilon \frac{\partial l(\xi; D)}{\partial U_{ij}} = \varepsilon(E_q[x_i x_j] - E_p[x_i x_j]) \tag{10}$$

where ε is the learning rate and $l(\xi; D) = \frac{1}{N}\sum_{n=1}^{N} \log(d_n; \xi)$. $E_q[\cdot]$ and $E_p[\cdot]$ are expectations over $q(x)$ and $p(x)$, respectively. Actually, $E_q[x_i x_j]$ and $E_p[x_i x_j]$ are the coordinates η^2_{ij} of $q(x)$ and $p(x)$, respectively. $E_q[x_i x_j]$ could be unbiasedly estimated from the sample. Markov chain Monte Carlo [22] is often used to approximate $E_p[x_i x_j]$ with an average over samples from $p(x)$.

Contrastive divergence (CD) learning realizes the gradient descent of a different objective function to avoid the difficulty of computing the log-likelihood gradient, shown as follows:

$$\Delta U_{ij} = -\varepsilon \frac{\partial(KL(q_0||p) - KL(p_m||p))}{\partial U_{ij}} = \varepsilon(E_{q_0}[x_i x_j] - E_{p_m}[x_i x_j]) \tag{11}$$

where q_0 is the sample distribution, p_m is the distribution by starting the Markov chain with the data and running m steps and $KL(\cdot||\cdot)$ denotes the K-L divergence. Taking samples in D as initial states, we could generate a set of samples for $p_m(x)$. Those samples can be used to estimate $E_{p_m}[x_i x_j]$.

From the perspective of IG, we can see that ML/CD learning is to update parameters in SBM, so that its corresponding coordinates $[\eta^{2-}]$ are getting closer to the data (along with the decreasing gradient). This is consistent with our theoretical analysis in Section 3 and Section 4.2 that SBM uses the most confident information (*i.e.*, $[\eta^{2-}]$) for approximating an arbitrary distribution in an expected sense.

5. Experimental Study: Incorporate Data into CIF

In the information-based CIF, the actual values of the data were not used to explicitly effect the output PDF (e.g., the derivation of SBM in Section 4). The data constrains the state of knowledge about the unknown pdf. In order to force the estimate of our probabilistic model to obey the data, we need to further reduce the difference between data information and physical information bound. How can this be done?

In this section, the CIF principle will also be used to modify existing SBM training algorithm (*i.e.*, CD-1) by incorporating data information. Given a particular dataset, the CIF can be used to further recognize less-confident parameters in SBM and to reduce them properly. Our solution here is to apply CIF to take effect on the learning trajectory with respect to specific samples and, hence, further confine the parameter space to the region indicated by the most confident information contained in the samples.

5.1. A Sample-Specific CIF-Based CD Learning for SBM

The main modification of our CIF-based CD algorithm (CD-CIF for short) is that we generate the samples for $p_m(x)$ based on those parameters with confident information, where the confident information carried by certain parameter is inherited from the sample and could be assessed using its Fisher information computed in terms of the sample.

For CD-1 (*i.e.*, m=1), the firing probability for the i-th neuron after a one-step transition from the initial state $x^{(0)} = \{x_1^{(0)}, x_2^{(0)}, \ldots, x_n^{(0)}\})$ is:

$$p(x_i^{(1)} = 1|x^{(0)}) = \frac{1}{1 + exp\{-\sum_{j \neq i} U_{ij} x_j^{(0)} - b_i\}} \tag{12}$$

For CD-CIF, the firing probability for the i-th neuron in Equation (12) is modified as follows:

$$p(x_i^{(1)}=1|x^{(0)})=\frac{1}{1+exp\{-\sum\limits_{(j\neq i)\&(F(U_{ij})>\tau)} U_{ij}x_j^{(0)}-b_i\}} \tag{13}$$

where τ is a pre-selected threshold, $F(U_{ij}) = E_{q_0}[x_i x_j] - E_{q_0}[x_i x_j]^2$ is the Fisher information of U_{ij} (see Equation (6)) and the expectations are estimated from the given sample D. We can see that those weights whose Fisher information are less than τ are considered to be unreliable w.r.t D. In practice, we could setup τ by the ratio r to specify the proportion of the total Fisher information T_{FI} of all parameters that we would like to remain, i.e., $\sum_{U_{ij}>\tau, i<j} F(U_{ij}) = r * T_{FI}$.

In summary, CD-CIF is realized in two phases. In the first phase, we initially “guess” whether certain parameter could be faithfully estimated based on the finite sample. In the second phase, we approximate the gradient using the CD scheme, except for when the CIF-based firing function in Equation (13) is used.

5.2. Experimental Results

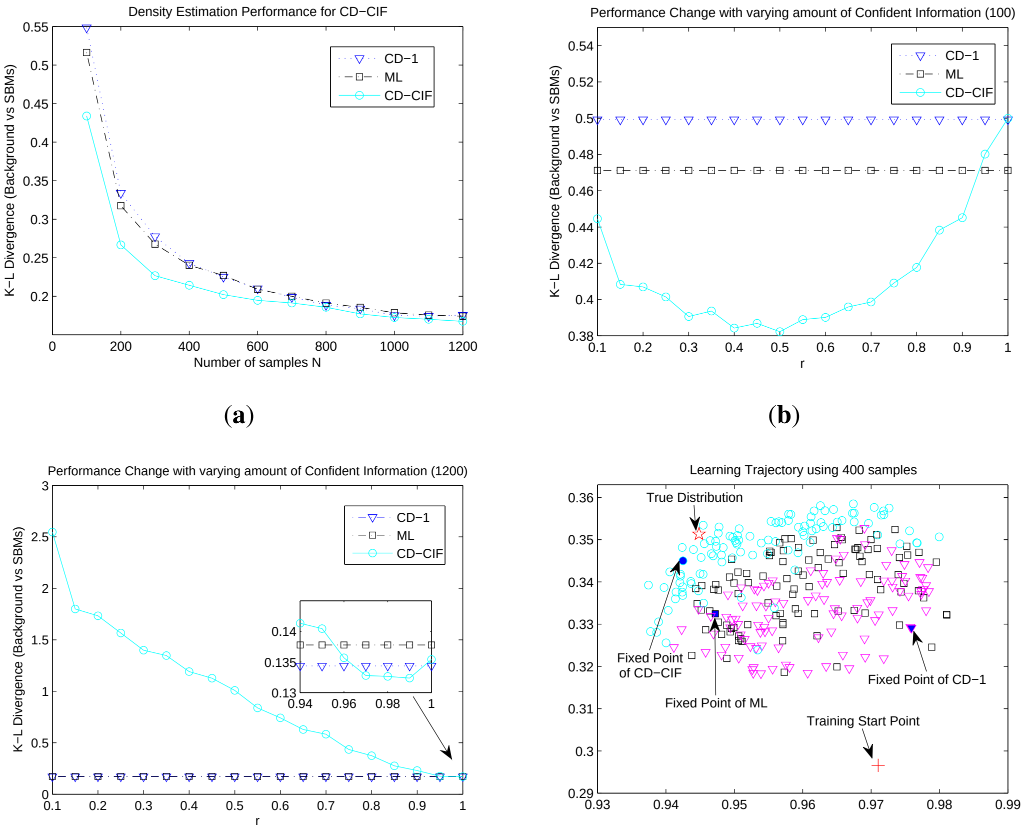
In this section, we empirically investigate our justifications for the CIF principle, especially how the sample-specific CIF-based CD learning (see Section 5) works in the context of density estimation.

Experimental Setup and Evaluation Metric: We utilize the random distribution uniformly generated from the open probability simplex over 10 variables as underlying distributions, whose samples size N may vary. Three learning algorithms are investigated: ML, CD-1 and our CD-CIF. K-L divergence is used to evaluate the goodness-of-fit of the SBM’s trained by various algorithms. For sample size N, we run 100 instances (20 (randomly generated distributions) $\times$ 5 (randomly running)) and report the averaged K-L divergences. Note that we focus on the case that the variable number is relatively small ($n = 10$) in order to analytically evaluate the K-L divergence and give a detailed study on algorithms. Changing the number of variables only offers a trivial influence for the experimental results, since we obtained qualitatively similar observations on various variable numbers (not reported here).

Automatically Adjusting r for Different Sample Sizes: The Fisher information is additive for i.i.d. sampling. When sample sizes change, it is natural to require that the total amount of Fisher information contained in all tailored parameters is steady. Hence, we have $\alpha = (1 - r)N$, where α indicates the amount of Fisher information and becomes a constant when the learning model and the underlying distribution family are given. It turns out that we can first identify α using the optimal r w.r.t several distributions generated from the underlying distribution family and then determine the optimal r’s for various sample sizes using $r = 1 - \alpha/N$. In our experiments, we set $\alpha = 35$.

Density Estimation Performance: The averaged K-L divergences between SBMs (learned by ML, CD-1 and CD-CIF with the r automatically determined) and the underlying distribution are shown in Figure 3a. In the case of relatively small samples ($N \leq 500$) in Figure 3a, our CD-CIF method shows significant improvements over ML (from 10.3% to 16.0%) and CD-1 (from 11.0% to 21.0%). This is because we could not expect to have reliable identifications for all model parameters from insufficient samples, and hence, CD-CIF gains its advantages by using parameters that could be confidently estimated. This result is consistent with our previous theoretical insight that Fisher information gives a reasonable guidance for parametric reduction via the confidence criterion. As the sample size increases ($N \geq 600$), CD-CIF, ML and CD-1 tend to have similar performances, since, with relatively large samples, most model parameters can be reasonably estimated, hence the effect of parameter reduction using CIF gradually becomes marginal. In Figure 3b and Figure 3c, we show how sample size affects the interval of r. For $N = 100$, CD-CIF achieves significantly better performances for a wide range of r. While, for $N = 1,200$, CD-CIF can only marginally outperform baselines for a narrow range of r.

Figure 3. **(a)**: the performance of CD-CIF on different sample sizes; **(b)** and **(c)**: The performances of CD-CIF with various values of r on two typical sample sizes, *i.e.*, 100 and 1200; **(d)** illustrates one learning trajectory of the last 100 steps for ML (squares), CD-1 (triangles) and CD-CIF (circles).



(a)

(b)

(c)

(d)

Effects on Learning Trajectory: We use the 2D visualizing technology SNE [20] to investigate learning trajectories and dynamical behaviors of three comparative algorithms. We start three methods with the same parameter initialization. Then, each intermediate state is represented by a 55-dimensional vector formed by its current parameter values. From Figure 3d, we can see that: (1) In the final 100 steps, the three methods seem to end up staying in different regions of the parameter space, and CD-CIF confines the parameter in a relatively thinner region compared to ML and CD-1; (2) The true distribution is usually located on the side of CD-CIF, indicating its potential for converging to the optimal solution. Note that the above claims are based on general observations, and Figure 3d is shown as an illustration. Hence, we may conclude that CD-CIF regularizes the learning trajectories in a desired region of the parameter space using the sample-specific CIF.

6. Conclusions

Different from the traditional EPI, the CIF principle proposed in this paper aims at finding a way to derive computational models for universal cognitive systems by a dimensionality reduction approach in parameter spaces: specifically, by preserving the confident parameters and reducing the less confident parameters. In principle, the confidence of parameters should be assessed according to their contributions to the expected information distance between the ideal distribution and its fluctuated observations. This is called the distance-based CIF. For some coordinated systems, e.g., the mixed-coordinate system $[\zeta]_l$, the confidence of a parameter can also be directly evaluated by its Fisher information, which establishes a connection with the inverse variance of any unbiased estimate for the parameter via the Cramér–Rao bound. This is called the information-based CIF. The criterion of information-based CIF (*i.e.*, Fisher information) can be seen as an approximation to distance-based CIF, since it neglects the influence of parameter scaling to the expected information distance. However, considering the standard mixed-coordinates $[\zeta]_l$ for the manifold of multivariate binary distributions, it turns out that both distance-based CIF and information-based CIF entail the same optimal submanifold M.

The CIF provides a strategy for the derivation of probabilistic models. The SBM is a specific example in this regard. It has been theoretically shown that the SBM can achieve a reliable representation in parameter spaces by using the CIF principle.

The CIF principle can also be used to modify existing SBM training algorithms by incorporating data information, such as CD-CIF. One interesting result shown in our experiments is that: although CD-CIF is a biased algorithm, it could significantly outperform ML when the sample is insufficient. This suggests that CIF gives us a reasonable criterion for utilizing confident information from the underlying data, while ML lacks a mechanism to do so.

In the future, we will further develop the formal justification of CIF w.r.t various contexts (e.g., distribution families or models).

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments. We also thank Mengjiao Xie and Shuai Mi for their helpful discussions. This work is partially supported by the Chinese National Program on Key Basic Research Project (973 Program, Grant No. 2013CB329304 and 2014CB744604), the Natural Science Foundation of China (Grant Nos. 61272265, 61070044, 61272291, 61111130190 and 61105072).

Appendix

A. Proof of Proposition 1

Proof. By definition, we have:

$$g_{IJ} = \frac{\partial^2 \psi(\theta)}{\partial \theta^I \partial \theta^J}$$

where $\psi(\theta)$ is defined by Equation (4). Hence, we have:

$$g_{IJ} = \frac{\partial^2 (\sum_I \theta^I \eta_I - \phi(\eta))}{\partial \theta^I \partial \theta^J} = \frac{\partial \eta_I}{\partial \theta^J}$$

By differentiating η_I, defined by Equation (1), with respect to θ^J, we have:

$$\begin{aligned} g_{IJ} &= \frac{\partial \eta_I}{\partial \theta^J} = \frac{\partial \sum_x X_I(x)(exp\{\sum_I \theta^I X_I(x) - \psi(\theta)\})}{\partial \theta^J} \\ &= \sum_x X_I(x)[X_J(x) - \eta_J]p(x;\theta) = \eta_{I\bigcup J} - \eta_I \eta_J \end{aligned}$$

This completes the proof. □

B. Proof of Proposition 2

Proof. By definition, we have:

$$g^{IJ} = \frac{\partial^2 \phi(\eta)}{\partial \eta_I \partial \eta_J}$$

where $\phi(\eta)$ is defined by Equation (4). Hence, we have:

$$g^{IJ} = \frac{\partial^2 (\sum_J \theta^J \eta_J - \psi(\theta))}{\partial \eta_I \partial \eta_J} = \frac{\partial \theta^I}{\partial \eta_J}$$

Based on Equations (2) and (1), the θ^I and p_K could be calculated by solving a linear equation of $[p]$ and $[\eta]$, respectively. Hence, we have:

$$\theta^I = \sum_{K \subseteq I} (-1)^{|I-K|} log(p_K); \;\; p_K = \sum_{K \subseteq J} (-1)^{|J-K|} \eta_J$$

Therefore, the partial derivation of θ^I with respect to η_J is:

$$g^{IJ} = \frac{\partial \theta^I}{\partial \eta_J} = \sum_K \frac{\partial \theta^I}{\partial p_K} \cdot \frac{\partial p_K}{\partial \eta_J} = \sum_{K \subseteq I \cap J} (-1)^{|I-K|+|J-K|} \cdot \frac{1}{p_K}$$

This completes the proof. □

C. Proof of Proposition 3

Proof. The Fisher information matrix of $[\zeta]$ could be partitioned into four parts: $G_\zeta = \begin{pmatrix} A & C \\ D & B \end{pmatrix}$. It can be verified that in the mixed coordinate, the θ-coordinate of order k is orthogonal to any η-coordinate less than k-order, implying the corresponding element of the Fisher information matrix is zero ($C = D = 0$) [23]. Hence, G_ζ is a block diagonal matrix.

According to the Cramér–Rao bound [3], a parameter (or a pair of parameters) has a unique asymptotically tight lower bound of the variance (or covariance) of the unbiased estimate, which is given by the corresponding element of the inverse of the Fisher information matrix involving this parameter (or this pair of parameters). Recall that I_η is the index set of the parameters shared by $[\eta]$ and $[\zeta]_l$ and that J_θ is the index set of the parameters shared by $[\theta]$ and $[\zeta]_l$; we have $(G_\zeta^{-1})_{I_\zeta} = (G_\eta^{-1})_{I_\eta}$ and $(G_\zeta^{-1})_{J_\zeta} = (G_\theta^{-1})_{J_\theta}$, *i.e.*, $G_\zeta^{-1} = \begin{pmatrix} (G_\eta^{-1})_{I_\eta} & 0 \\ 0 & (G_\theta^{-1})_{J_\theta} \end{pmatrix}$. Since G_ζ is a block tridiagonal matrix, the proposition follows. □

D. Proof of Proposition 4

Proof. Assume the Fisher information matrix of $[\theta]$ to be: $G_\theta = \begin{pmatrix} U & X \\ X^T & V \end{pmatrix}$, which is partitioned based on I_η and J_θ. Based on Proposition 3, we have $A = U^{-1}$. Obviously, the diagonal elements of U are all smaller than one. According to the succeeding Lemma 6, we can see that the diagonal elements of A (*i.e.*, U^{-1}) are greater than one.

Next, we need to show that the diagonal elements of B are smaller than 1. Using the Schur complement of G_θ, the bottom-right block of G_θ^{-1}, *i.e.*, $(G_\theta^{-1})_{J_\theta}$, equals to $(V - X^T U^{-1} X)^{-1}$. Thus, the diagonal elements of B: $B_{jj} = (V - X^T U^{-1} X)_{jj} < V_{jj} < 1$. Hence, we complete the proof. □

Lemma 6. *With a $l \times l$ positive definite matrix H, if $H_{ii} < 1$, then $(H^{-1})_{ii} > 1, \forall i \in \{1, 2, \ldots, l\}$.*

Proof. Since H is positive definite, it is a Gramian matrix of l linearly independent vectors $v_1, v_2, \ldots, v_l$, *i.e.*, $H_{ij} = \langle v_i, v_j \rangle$ ($\langle \cdot, \cdot \rangle$ denotes the inner product). Similarly, H^{-1} is the Gramian matrix of l linearly independent vectors $w_1, w_2, \ldots, w_l$ and $(H^{-1})_{ij} = \langle w_i, w_j \rangle$. It is easy to verify that $\langle w_i, v_i \rangle = 1, \forall i \in \{1, 2, \ldots, l\}$. If $H_{ii} < 1$, we can see that the norm $\|v_i\| = \sqrt{H_{ii}} < 1$. Since $\|w_i\| \times \|v_i\| \geq \langle w_i, v_i \rangle = 1$, we have $\|w_i\| > 1$. Hence, $(H^{-1})_{ii} = \langle w_i, w_i \rangle = \|w_i\|^2 > 1$. □

E. Proof of Proposition 5

Proof. Let B_q be a ε-ball 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 Kullback–Leibler divergence and ε is small. ζ_q is the coordinates of $q(x)$. Let $q(x) + dq$ be a neighbor of $q(x)$ uniformly sampled on B_q and $\zeta_{q(x)+dq}$ be its corresponding coordinates. For a small ε, we can calculate the expected information distance between $q(x)$ and $q(x) + dq$ as follows:

$$E_{B_q} = \int [(\zeta_{q(x)+dq} - \zeta_q)^T G_\zeta (\zeta_{q(x)+dq} - \zeta_q)]^{\frac{1}{2}} dB_q \tag{A1}$$

where G_ζ is the Fisher information matrix at $q(x)$.

Since Fisher information matrix G_ζ is both positive definite and symmetric, there exists a singular value decomposition $G_\zeta = U^T \Lambda U$ where U is an orthogonal matrix and Λ is a diagonal matrix with diagonal entries equal to the eigenvalues of G_ζ (all ≥ 0).

Applying the singular value decomposition into Equation (A1), the distance becomes:

$$E_{B_q} = \int [(\zeta_{q(x)+dq} - \zeta_q)^T U^T \Lambda U (\zeta_{q(x)+dq} - \zeta_q)]^{\frac{1}{2}} dB_q \tag{A2}$$

Note that U is an orthogonal matrix, and the transformation $U(\zeta_{q(x)+dq} - \zeta_q)$ is a norm-preserving rotation.

Now, we need to show that among all tailored k-dimensional submanifolds of S, $[\zeta]_{l_t}$ is the one that preserves maximum information distance. 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 $[\zeta]$. According to the fundamental analytical properties of the surface of the hyper-ellipsoid and the orthogonality of the mixed-coordinates, there exists a strict positive monotonicity between the expected information distance E_{B_q} for T and the sum of eigenvalues of the sub-matrix $(G_\zeta)_{I_T}$, where the sum equals to the trace of $(G_\zeta)_{I_T}$. That is, the greater the trace of $(G_\zeta)_{I_T}$, the greater the expected information distance E_{B_q} for T.

Next, we show that the sub-matrix of G_ζ specified by $[\zeta]_{l_t}$ gives a maximum trace. Based on Proposition 4, the elements on the main diagonal of the sub-matrix A are lower bounded by one and those of B upper bounded by one. Therefore, $[\zeta]_{l_t}$ gives the maximum trace among all sub-matrices of G_ζ. This completes the proof. □

Author Contributions

Theoretical study and proof: Yuexian Hou and Xiaozhao Zhao. Conceived and designed the experiments: Xiaozhao Zhao, Yuexian Hou, Dawei Song and Wenjie Li. Performed the experiments: Xiaozhao Zhao. Analyzed the data: Xiaozhao Zhao, Yuexian Hou. Wrote the manuscript: Xiaozhao Zhao, Dawei Song, Wenjie Li and Yuexian Hou. All authors have read and approved the final manuscript.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Wheeler, J.A. *Time Today*; Cambridge University Press: Cambridge, UK, 1994; pp. 1–29.
2. Frieden, B.R. *Science from Fisher Information: A Unification*; Cambridge University Press: Cambridge, UK, 2004.
3. Rao, C.R. Information and the accuracy attainable in the estimation of statistical parameters. *Bull. Calcutta Math. Soc.* **1945**, *37*, 81–91.
4. Frieden, B.R.; Gatenby, R.A. Principle of maximum Fisher information from Hardy's axioms applied to statistical systems. *Phys. Rev. E* **2013**, *88*, 042144.
5. Burnham, K.P.; Anderson, D.R. *Model Selection and Multimodel Inference: A Practical Information—Theoretic Approach*; Springer: Berlin/Heidelberg, Germany, 2002.

6. Vstovsky, G.V. Interpretation of the extreme physical information principle in terms of shift information. *Phys. Rev. E* **1995**, *51*, 975–979.
7. Amari, S.; Nagaoka, H. *Methods of Information Geometry*; Translations of Mathematical Monographs; Oxford University Press: Oxford, UK, 1993.
8. Amari, S.; Kurata, K.; Nagaoka, H. Information geometry of Boltzmann machines. *IEEE Trans. Neural Netw.* **1992**, *3*, 260–271.
9. Hou, Y.; Zhao, X.; Song, D.; Li, W. Mining pure high-order word associations via information geometry for information retrieval. *ACM Trans. Inf. Syst.* **2013**, *31*, 12:1–12:32.
10. Kass, R.E. The geometry of asymptotic inference. *Stat. Sci.* **1989**, *4*, 188–219.
11. Ackley, D.H.; Hinton, G.E.; Sejnowski, T.J. A learning algorithm for Boltzmann machines. *Cogn. Sci.* **1985**, *9*, 147–169.
12. Hinton, G.E.; Salakhutdinov, R.R. Reducing the Dimensionality of Data with Neural Networks. *Science* **2006**, *313*, 504–507.
13. Rasch, G. *Probabilistic Models for Some Intelligence and Attainment Tests*; Danish Institute for Educational Research: Copenhagen, Denmark, 1960.
14. Bond, T.; Fox, C. *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*; Psychology Press: London, UK, 2013.
15. Gibilisco, P. *Algebraic and Geometric Methods in Statistics*; Cambridge University Press: Cambridge, UK, 2010.
16. Čencov, N.N. *Statistical Decision Rules and Optimal Inference*; American Mathematical Society: Washington, D.C., USA, 1982.
17. Kullback, S.; Leibler, R.A. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 79–86.
18. Buhlmann, P.; van de Geer, S. *Statistics for High-Dimensional Data: Methods, Theory And Applications*; Springer: Berlin/Heidelberg, Germany, 2011.
19. Bobrovsky, B.; Mayer-Wolf, E.; Zakai, M. Some classes of global Cramér-Rao bounds. *Ann. Stat.* **1987**, *15*, 1421–1438.
20. Hinton, G.E. Training products of experts by minimizing contrastive divergence. *Neural Comput.* **2002**, *14*, 1771–1800.
21. Carreira-Perpinan, M.A.; Hinton, G.E. On contrastive divergence learning. In Proceedings of the International Workshop on Artificial Intelligence and Statistics, Key West, FL, USA, 6–8 January 2005; pp. 33–40.
22. Gilks, W.R.; Richardson, S.; Spiegelhalter, D. Introducing markov chain monte carlo. In *Markov Chain Monte Carlo in Practice*; Chapman and Hall/CRC: London, UK, 1996; pp. 1–19.
23. Nakahara, H.; Amari, S. Information geometric measure for neural spikes. *Neural Comput.* **2002**, *14*, 2269–2316.