Information Geometry on Hierarchy of Probability Distributions

Shun-ichi Amari, *Fellow, IEEE*

***Abstract*—An exponential family or mixture family of probability distributions has a natural hierarchical structure. This paper gives an "orthogonal" decomposition of such a system based on information geometry. A typical example is the decomposition of stochastic dependency among a number of random variables. In general, they have a complex structure of dependencies. Pairwise dependency is easily represented by correlation, but it is more difficult to measure effects of pure triplewise or higher order interactions (dependencies) among these variables. Stochastic dependency is decomposed quantitatively into an "orthogonal" sum of pairwise, triplewise, and further higher order dependencies. This gives a new invariant decomposition of joint entropy. This problem is important for extracting intrinsic interactions in firing patterns of an ensemble of neurons and for estimating its functional connections. The orthogonal decomposition is given in a wide class of hierarchical structures including both exponential and mixture families. As an example, we decompose the dependency in a higher order Markov chain into a sum of those in various lower order Markov chains.**

***Index Terms*—Decomposition of entropy, e- and m-projections, extended Pythagoras theorem, higher order interactions, higher order Markov chain, information geometry, Kullback divergence.**

I. Introduction

We study structures of hierarchical systems of probability distributions by information geometry. Examples of such systems are exponential families, mixture families, higher order Markov chains, autoregressive (AR) and moving average (MA) models, and others. Given a probability distribution, we decompose it into hierarchical components. Different from the Euclidean space, no orthogonal decomposition into components exists. However, when a system of probability distributions forms a dually flat Riemannian manifold, we can decompose the effects in various hierarchies in a quasi-orthogonal manner.

A typical example we study is interactions among a number of random variables $X_1, \ldots, X_n$. Interactions among them include not only pairwise correlations, but also triplewise and higher interactions, forming a hierarchical structure. This case has been studied extensively by the log-linear model [2] which gives a hierarchical structure, but the log-linear model itself does not give an orthogonal decomposition of interactions. Given a joint distribution of n random variables, it is important to search for an invariant "orthogonal" decomposition of their degrees or amounts of interactions into pairwise, triplewise, and higher order interactions. To this end, we study a family $\boldsymbol{E}_k$ $(k = 1, 2, \ldots, n)$ of joint probability distributions of n variables which have at most k-way interactions but no higher interactions. Two dual types of projections, namely, the e-projection and m-projection, to such subspaces play a fundamental role.

The present paper studies such a hierarchical structure and the related invariant "quasi-orthogonal" quantitative decomposition by using information geometry [3], [8], [12], [14], [28], [30]. Information geometry studies the intrinsic geometrical structure to be introduced in the manifold of a family of probability distributions. Its Riemannian structure was introduced by Rao [37]. Csiszár [21], [22], [23] studied the geometry of f-divergence in detail and applied it to information theory. It was Chentsov [19] who developed Rao's idea further and introduced new invariant affine connections in the manifolds of probability distributions. Nagaoka and Amari [31] developed a theory of dual structures and unified all of these theories in the dual differential-geometrical framework (see also [3], [14], [31]). Information geometry has been used so far not only for mathematical foundations of statistical inferences ([3], [12], [28] and many others) but also applied to information theory [5], [11], [25], [18], neural networks [6], [7], [9], [13], systems theory [4], [32], mathematical programming [33], statistical physics [10], [16], [38], and others. Mathematical foundations of information geometry in the function space were given by Pistone and his coworkers [35], [36] and are now developing further.

The present paper shows how information geometry gives an answer to the problem of invariant decomposition for hierarchical systems of probability distributions. This leads to a new invariant decomposition of entropy and information. It can be applied to the analysis of synchronous firing patterns of n neurons by decomposing their effects into hierarchies [34], [1]. Such a hierarchical structure also exists in the family of higher order Markov chains and also graphical conditional independence models [29].

The present paper is organized as follows. After the Introduction, Section II is devoted to simple introductory explanations of information geometry. We then study the e-flat and m-flat hierarchical structures, and give the quantitative "orthogonal decomposition" of higher order effects in these structures. We then show a simple example consisting of three binary variables and study how a joint distribution is quantitatively decomposed into pairwise and pure triplewise interactions. This gives a new decomposition of the joint entropy. Section IV is devoted to a general theory on decomposition of interactions among n variables. A new decomposition of entropy is also derived there. We

Manuscript received November 15, 1999; revised December 28, 2000.

The author is with the Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Saitama 351-0198, Japan (e-mail: amari@brain.riken.go.jp).

Communicated by I. Csiszár, Associate Editor for Shannon Theory.
Publisher Item Identifier S 0018-9448(01)04420-0.

touch upon the cases of multivalued random variables and continuous variables. We finally explain the hierarchical structures of higher order Markov chains.

II. Preliminaries from Information Geometry

A. Manifold, Curve, and Orthogonality

Let us consider a parameterized family of probability distributions $\boldsymbol{S} = \{p(x, \boldsymbol{\xi})\}$, where x is a random variable and $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)$ is a real vector parameter to specify a distribution. The family $\boldsymbol{S}$ is regarded as an n-dimensional manifold having $\boldsymbol{\xi}$ as a coordinate system. When the Fisher information matrix $G = (g_{ij})$

$$g_{ij}(\boldsymbol{\xi}) = E\left[\frac{\partial \log p(x, \boldsymbol{\xi})}{\partial \xi_i} \frac{\partial \log p(x, \boldsymbol{\xi})}{\partial \xi_j}\right] \tag{1}$$

where E denotes expectation with respect to $p(x, \boldsymbol{\xi})$, is nondegenerate, $\boldsymbol{S}$ is a Riemannian manifold, and $G(\boldsymbol{\xi})$ plays the role of a Riemannian metric tensor.

The squared distance ds^2 between two nearby distributions $p(x, \boldsymbol{\xi})$ and $p(x, \boldsymbol{\xi}, +d\boldsymbol{\xi})$ is given by the quadratic form of $d\boldsymbol{\xi}$

$$ds^2 = \sum g_{ij}(\boldsymbol{\xi})\, d\xi^i\, d\xi^j. \tag{2}$$

It is known that this is twice the Kullback–Leibler divergence

$$ds^2 = 2KL[p(x, \boldsymbol{\xi}) : p(x, \boldsymbol{\xi} + d\boldsymbol{\xi})] \tag{3}$$

where

$$KL[p : q] = \int p(x) \log \frac{p(x)}{q(x)}\, dx. \tag{4}$$

Let us consider a curve $\boldsymbol{\xi} = \boldsymbol{\xi}(t)$ parameterized by t in $\boldsymbol{S}$, that is, a one-parameter family of distributions $p(x, \boldsymbol{\xi}(t))$ in $\boldsymbol{S}$. It is convenient to represent the tangent vector $\dot{\boldsymbol{\xi}}(t) = (d/dt)\boldsymbol{\xi}(t)$ of the curve at t by the random variable called the score

$$\dot{\boldsymbol{\xi}}(t) = \frac{d}{dt} \log p(x, \boldsymbol{\xi}(t)) \tag{5}$$

which shows how the log probability changes as t increases. Given two curves $\boldsymbol{\xi}_1(t)$ and $\boldsymbol{\xi}_2(t)$ intersecting at t, the inner product of the two tangent vectors is given by

$$\begin{aligned}\left\langle \dot{\boldsymbol{\xi}}_1(t), \dot{\boldsymbol{\xi}}_2(t) \right\rangle &= E\left[\frac{d}{dt} \log p(x, \boldsymbol{\xi}_1(t)) \frac{d}{dt} \log p(x, \boldsymbol{\xi}_2(t))\right] \\ &= \sum g_{ij}(\boldsymbol{\xi}) \frac{d}{dt} \xi_{2j}(t) \frac{d}{dt} \xi_{1i}(t). \end{aligned} \tag{6}$$

The two curves intersect at $\boldsymbol{\xi}_1(t) = \boldsymbol{\xi}_2(t)$ orthogonally when

$$\left\langle \dot{\boldsymbol{\xi}}_1(t), \dot{\boldsymbol{\xi}}_2(t) \right\rangle = 0 \tag{7}$$

that is, when the two scores are noncorrelated.

B. Dually Flat Manifolds

A manifold $\boldsymbol{S}$ is said to be e-flat (exponential-flat), when there exists a coordinate system (parameterization) $\boldsymbol{\theta}$ such that, for all i, j, k

$$E\left[\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(x, \boldsymbol{\theta}) \frac{\partial}{\partial\theta_k} \log p(x, \boldsymbol{\theta})\right] = 0 \tag{8}$$

identically. Such $\boldsymbol{\theta}$ is called e-affine coordinates. When a curve $\boldsymbol{\theta}(t)$ is given by a linear function $\boldsymbol{\theta}(t) = t\boldsymbol{a} + \boldsymbol{b}$ in the $\boldsymbol{\theta}$-coordinates, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are constant vectors, it is called an e-geodesic. Any coordinate curve θ_i itself is an e-geodesic. (It is possible to define an e-geodesic in any manifold, but it is no more linear and we need the concept of the affine connection.)

A typical example of an e-flat manifold is the well-known exponential family written as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\sum \theta_i k_i(x) - \psi(\boldsymbol{\theta})\right\} \tag{9}$$

where $k_i(x)$ are given functions and ψ is the normalizing factor. The e-affine coordinates are the canonical parameters $\boldsymbol{\theta} = (\theta_i)$, and (8) holds because

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \psi(\boldsymbol{\theta}) \tag{10}$$

does not depend on x and $E[\frac{\partial}{\partial\theta_i} \log p] = 0$.

Dually to the above, a manifold is said to be m-flat (mixture-flat), when there exists a coordinate system $\boldsymbol{\eta}$ such that

$$E\left[\frac{1}{p(x, \boldsymbol{\eta})} \frac{\partial^2}{\partial\eta_i \partial\eta_j} p(x, \boldsymbol{\eta}) \frac{\partial}{\partial\eta_k} \log p(x, \boldsymbol{\eta})\right] = 0 \tag{11}$$

identically. Here, $\boldsymbol{\eta}$ is called m-affine coordinates. A curve is called an m-geodesic when it is represented by a linear function $\boldsymbol{\eta}(t) = \boldsymbol{a}t + \boldsymbol{b}$ in the m-affine coordinates. Any coordinate curve η_i of $\boldsymbol{\eta}$ is an m-geodesic.

A typical example of an m-flat manifold is the mixture family

$$p(x, \boldsymbol{\eta}) = \sum \eta_i q_i(x) + \left(1 - \sum \eta_i\right) q_0(x) \tag{12}$$

where $q_i(x)$ are given probability distributions and $0 < \eta_i < 1$, $\sum \eta_i < 1$.

The following theorem is known in information geometry.

Theorem 1: A manifold $\boldsymbol{S}$ is e-flat when and only when it is m-flat and *vice versa*.

This shows that an exponential family is automatically m-flat although it is not necessarily a mixture family. A mixture family is e-flat, although it is not in general an exponential family. The m-affine coordinates ($\boldsymbol{\eta}$-coordinates) of an exponential family are given by

$$\eta_i = E[k_i(x)] = \frac{\partial}{\partial\theta_i} \psi(\boldsymbol{\theta}) \tag{13}$$

which is known as the expectation parameters. The coordinate transformation between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ is given by the Legendre transformation, and the inverse transformation is

$$\theta_i = \frac{\partial\varphi(\boldsymbol{\eta})}{\partial\eta_i} \tag{14}$$

where $\varphi(\boldsymbol{\eta})$ is the negative entropy

$$\varphi(\boldsymbol{\eta}) = E[\log p(x, \boldsymbol{\eta})] \tag{15}$$

and

$$\psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}) + \boldsymbol{\theta} \cdot \boldsymbol{\eta} = 0 \tag{16}$$

holds with $\boldsymbol{\theta} \cdot \boldsymbol{\eta} = \sum \theta_i \eta_i$. This was first remarked by Barndorff-Nielsen [15] in the case of exponential families.

Given a distribution in a flat manifold, $p(x, \boldsymbol{\theta})$ in the coordinates $\boldsymbol{\theta}$ and $p(x, \boldsymbol{\eta})$ in the coordinates $\boldsymbol{\eta}$ have different function forms, so that they should be denoted differently such as $p_\theta(x, \boldsymbol{\theta})$ and $p_\eta(x, \boldsymbol{\eta})$, respectively. However, we abuse notation such that the parameter $\boldsymbol{\theta}$ or $\boldsymbol{\eta}$ decides the function form of $p(x, \boldsymbol{\theta})$ or $p(x, \boldsymbol{\eta})$ automatically.

Dually to the above, a mixture family is e-flat, although it is not an exponential family in general. The e-affine coordinates $\boldsymbol{\theta}$ are derived from

$$\theta_i = \frac{\partial \varphi(\boldsymbol{\eta})}{\partial \eta_i} = \int \{q_i(x) - q_0(x)\} \log p(x, \boldsymbol{\eta})\, dx. \tag{17}$$

The ψ-function of (16) in a mixture family is

$$\psi(\boldsymbol{\theta}) = -\int q_0(x) p(x, \boldsymbol{\eta}(\boldsymbol{\theta}))\, dx. \tag{18}$$

When $\boldsymbol{S}$ is a dually flat manifold, the e-affine coordinates $\boldsymbol{\theta}$ and m-affine coordinates $\boldsymbol{\eta}$, connected by the Legendre transformations (13) and (14), satisfy the following dual relation.

Theorem 2: The tangent vectors (represented by random variables) of the coordinate curves θ_i

$$\boldsymbol{e}_i = \frac{\partial}{\partial \theta_i} \log p(x, \boldsymbol{\theta}) \tag{19}$$

and the tangent vectors of the coordinate curves η_j

$$\boldsymbol{e}_j^* = \frac{\partial}{\partial \eta_j} \log p(x, \boldsymbol{\eta}) \tag{20}$$

are orthonormal at all the points

$$\begin{aligned} \langle \boldsymbol{e}_i, \boldsymbol{e}_j^* \rangle &= E\left[\frac{\partial}{\partial \theta_i} \log p(x, \boldsymbol{\theta}) \frac{\partial}{\partial \eta_j} \log p(x, \boldsymbol{\eta})\right] \\ &= \delta_{ij} \end{aligned} \tag{21}$$

where δ_{ij} is the Kronecker delta.

C. Divergence and Generalized Pythagoras Theorem

Let $p = p(x, \boldsymbol{\theta})$ and $p' = p(x, \boldsymbol{\theta}')$ be two distributions in a dually flat manifold $\boldsymbol{S}$, and let $\boldsymbol{\eta}$ and $\boldsymbol{\eta}'$ be the corresponding m-affine coordinates. They have two convex potential functions $\psi(\boldsymbol{\theta})$ and $\varphi(\boldsymbol{\eta})$. In the case of exponential families, ψ is the cumulant generating function and φ is the negative entropy. For a mixture family, φ is also the negative entropy. By using the two functions, we can define a divergence from p to p' by

$$D[p : p'] = \psi(\boldsymbol{\theta}) + \varphi(\boldsymbol{\eta}') - \boldsymbol{\theta} \cdot \boldsymbol{\eta}'. \tag{22}$$

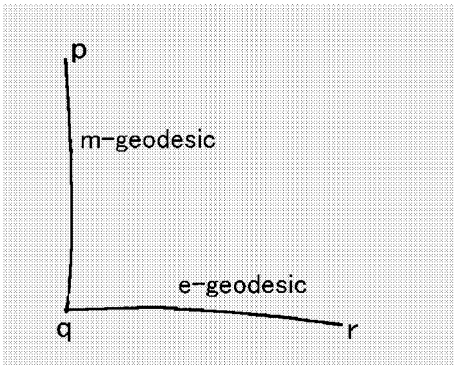

Fig. 1. Generalized Pythagoras theorem.

The divergence satisfies $D[p : p'] \geq 0$ with equality when, and only when, $p = p'$. In the cases of an exponential family and a mixture family, this is equal to the Kullback–Leibler divergence

$$D[p : p'] = E_{\boldsymbol{\theta}}\left[\log \frac{p(x, \boldsymbol{\theta})}{p(x, \boldsymbol{\theta}')}\right] \tag{23}$$

where $E_{\boldsymbol{\theta}}$ is the expectation with respect to $p(\boldsymbol{x}, \boldsymbol{\theta})$.

For a dually flat manifold $\boldsymbol{S}$, the following Pythagoras theorem plays a key role (Fig. 1).

Theorem 3: Let p, q, r be three distributions in $\boldsymbol{S}$. When the m-geodesic connecting p and q is orthogonal at q to the e-geodesic connecting q and r

$$D[p : q] + D[q : r] = D[p : r]. \tag{24}$$

The same theorem can be reformulated in a dual way.

Theorem 4: For $p, q, r \in \boldsymbol{S}$, when the e-geodesic connecting p and q is orthogonal at q to the m-geodesic connecting q and r

$$\overline{D}[p : q] + \overline{D}[q : r] = \overline{D}[p : r] \tag{25}$$

with

$$\overline{D}[p : q] = D[q : p]. \tag{26}$$

III. Flat Hierarchical Structures

We have summarized the geometrical features of dually flat families of probability distributions. We extend them to the geometry of flat hierarchical structures.

A. E-Flat Structures

Let $\boldsymbol{T} \subset \boldsymbol{S}$ be a submanifold of a dually flat manifold $\boldsymbol{S}$. It is called an e-flat submanifold, when $\boldsymbol{T}$ is written as a linear subspace in the e-affine coordinates $\boldsymbol{\theta}$ of $\boldsymbol{S}$. It is called an m-flat submanifold, when it is linear in the m-affine coordinates $\boldsymbol{\eta}$ of $\boldsymbol{S}$. An e-flat submanifold $\boldsymbol{T}$ is by itself an e-flat manifold, and hence is an m-flat manifold because of Theorem 1. However, it is not usually an m-flat submanifold of $\boldsymbol{S}$, because it is not linear in $\boldsymbol{\eta}$. (Mathematically speaking, an e-flat submanifold has vanishing embedding curvature in the sense of the e-affine connection, but its m-embedding curvature is nonvanishing, although both e- and m-Riemann–Christoffel curvatures vanish.) Dually, an m-flat submanifold is an e-flat manifold but is not an e-flat submanifold.

Let us consider a nested series of e-flat submanifolds

$$\boldsymbol{E}_1 \subset \boldsymbol{E}_2 \subset \cdots \subset \boldsymbol{E}_n \tag{27}$$

where every $\boldsymbol{E}_k$ is an e-flat submanifold of $\boldsymbol{E}_{k+1}$. Each $\boldsymbol{E}_k$ is automatically dually flat, but is not an m-flat submanifold. We call such a nested series an e-flat hierarchical structure or, shortly, the e-structure. A typical example of the e-structure is the following exponential-type distributions:

$$p(x, \boldsymbol{\theta}) = \exp\left\{\sum_{\beta=1}^{n} \boldsymbol{\theta}_\beta \cdot \boldsymbol{g}_\beta(x) - \psi(\boldsymbol{\theta})\right\} \tag{28}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ is a partition of the entire parameter $\boldsymbol{\theta}$, $\boldsymbol{\theta}_\beta$ being the βth subvector, and $\boldsymbol{g}_\beta(x)$ is a random vector variable corresponding to it.

The expectation parameter $\boldsymbol{\eta}$ of (28) is partitioned correspondingly as

$$\boldsymbol{\eta}_\beta = E_{\boldsymbol{\theta}}[g_\beta(\boldsymbol{x})], \qquad \beta = 1, \ldots, n \tag{29}$$

where $\boldsymbol{E}_{\boldsymbol{\theta}}$ is expectation with respect to $p(x, \boldsymbol{\theta})$. Thus, $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$. Here, $\boldsymbol{\eta}_\beta(\boldsymbol{\theta})$ is a function of the entire $\boldsymbol{\theta}$, and not of $\boldsymbol{\theta}_\beta$ only.

B. Orthogonal Foliations

Let us consider a new coordinate system called the k-cut mixed ones

$$\boldsymbol{\xi}_k = (\boldsymbol{\eta}_{k-}; \boldsymbol{\theta}_{k+}) = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k; \boldsymbol{\theta}_{k+1}, \ldots, \boldsymbol{\theta}_n). \tag{30}$$

It consists of a pair of complementary parts of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$, namely,

$$\boldsymbol{\eta}_{k-} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_k) \tag{31}$$

$$\boldsymbol{\theta}_{k+} = (\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_{k+2}, \ldots, \boldsymbol{\theta}_n). \tag{32}$$

It is defined for any k.

We define a subset $\boldsymbol{E}_k(\boldsymbol{c}_{k+})$ for $\boldsymbol{c}_{k+} = (\boldsymbol{c}_{k+1}, \ldots, \boldsymbol{c}_n)$ of $\boldsymbol{S}$ consisting of all the distributions having the same $\boldsymbol{\theta}_{k+}$ coordinates, specified by $\boldsymbol{\theta}_{k+} = \boldsymbol{c}_{k+}$, but the other $\boldsymbol{\theta}$ coordinates are free. This is written as

$$\boldsymbol{E}_k(\boldsymbol{c}_{k+}) = \{p(\boldsymbol{x}, \boldsymbol{\theta}) \mid \boldsymbol{\theta}_{k+} = \boldsymbol{c}_{k+}\}. \tag{33}$$

They give a foliation of the entire manifold

$$\bigcup_{\boldsymbol{c}_{k+}} \boldsymbol{E}_k(\boldsymbol{c}_{k+}) = \boldsymbol{S}. \tag{34}$$

The hierarchical e-structure is introduced in $\boldsymbol{S}$ by putting $\boldsymbol{c}_{k+} = 0$

$$\boldsymbol{E}_1(0) \subset \boldsymbol{E}_2(0) \subset \cdots \subset \boldsymbol{E}_n(0) = \boldsymbol{S}. \tag{35}$$

Dually to the above, let

$$\boldsymbol{M}_k(\boldsymbol{d}_{k-}) = \{p(\boldsymbol{x}, \boldsymbol{\theta}) \mid \boldsymbol{\eta}_{k-} = \boldsymbol{d}_{k-}\} \tag{36}$$

be the subset in $\boldsymbol{S}$ in which the $\boldsymbol{\eta}_{k-}$-part of the distributions have a common fixed value $\boldsymbol{d}_{k-}$. This is an m-flat submanifold. They define m-flat foliations. Because of (21), the two foliations are orthogonal in the sense that submanifolds $\boldsymbol{M}_k$ and $\boldsymbol{E}_k$ are complementary and orthogonal at any point.

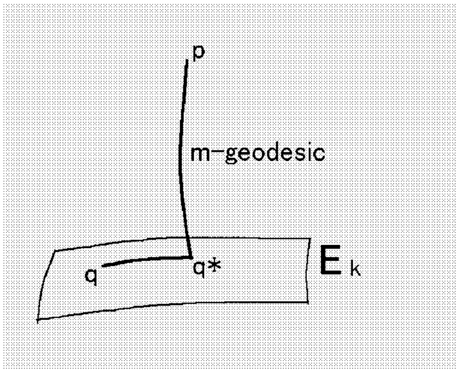

Fig. 2. Information projection (m-projection).

C. Projections and Pythagoras Theorem

Given a distribution $p(\boldsymbol{x}, \boldsymbol{\theta})$ with partition $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, it belongs to $\boldsymbol{E}_k = \boldsymbol{E}_k(0)$ when $\boldsymbol{\theta}_{k+} = (\boldsymbol{\theta}_{k+1}, \ldots, \boldsymbol{\theta}_n) = 0$. In general, $\boldsymbol{\theta}_k$ is considered to represent the effect emerging from $\boldsymbol{E}_k$ but not from $\boldsymbol{E}_{k-1}$. We call it the "kth effect" of $p(x, \boldsymbol{\theta})$. In later examples, it represents the kth-order effect of mutual interactions of random variables. The submanifold $\boldsymbol{E}_k$ includes only the probability distributions that do not have effects higher than k. Consider the problem of evaluating the effects higher than k. To this end, we define the information projection of p to $\boldsymbol{E}_k$ by

$$p^{(k)}(x) = \arg\min_{q \in \boldsymbol{E}_k} D[p(x) : q(x)]. \tag{37}$$

This $p^{(k)}$ is the point in $\boldsymbol{E}_k$ that is closest to p in the sense of divergence.

Theorem 5: Let q^* be the point in $\boldsymbol{E}_k$ such that the m-geodesic connecting p and q^* is orthogonal to $\boldsymbol{E}_k$. Then, q^* is unique and $p^{(k)}$ is given by q^*.

Proof: For any point $q \in \boldsymbol{E}_k$, the e-geodesic connecting q and q^* is included in $\boldsymbol{E}_k$, and hence is orthogonal to the m-geodesic connecting p and q^* (Fig. 2). Hence, the Pythagoras theorem

$$D[p : q] = D[p : q^*] + D[q^* : q] \tag{38}$$

holds. This proves $p^{(k)} = q^*$ and q^* is unique. □

We call $p^{(k)}$ the m-projection of p to $\boldsymbol{E}_k$, and write

$$p^{(k)} = \prod^{k} p. \tag{39}$$

Let $p_0 \in \boldsymbol{E}_0$ be a fixed distribution. We then have

$$D[p : p_0] = D\left[p : p^{(k)}\right] + D\left[p^{(k)} : p^0\right]. \tag{40}$$

Hence, $D[p : p^{(k)}]$ is regarded as the amount representing effects of p higher than k, whereas $D[p^{(k)} : p_0]$ is that not higher than k.

The coordinates of $p^{(k)}$ are not simply

$$(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k, 0, \ldots, 0)$$

since the manifold is not Euclidean but is Riemannian. In order to obtain $p^{(k)}$, the k-cut mixed coordinates $\boldsymbol{\xi}_k$ are convenient.

Theorem 6: Let the e- and m-coordinates of p be $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$. Let the e- and m-coordinates of the projection $p^{(k)} = \prod^k p$ be $\boldsymbol{\theta}^{(k)}$ and $\boldsymbol{\eta}^{(k)}$, respectively. Then, the k-cut mixed coordinates of $p^{(k)}$ is

$$\boldsymbol{\xi}_k = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k; 0, \ldots, 0) \tag{41}$$

that is, $\boldsymbol{\eta}_{k-}^{(k)} = \boldsymbol{\eta}_{k-}$ and $\boldsymbol{\theta}_{k+}^{(k)} = 0$.

Proof: The point given by (41) is included in $\boldsymbol{E}_k(0)$. Since the m-coordinates of p and $p^{(k)}$ differ only in $\boldsymbol{\eta}_{k+1}, \ldots \boldsymbol{\eta}_n$, the m-geodesic connecting $p(x, \boldsymbol{\eta})$ and $p^{(k)}(x, \boldsymbol{\eta}^{(k)})$ is included in $\boldsymbol{M}_k(\boldsymbol{\eta}_{k-})$. Since $\boldsymbol{M}_k$ is orthogonal to $\boldsymbol{E}_k$, this $p^{(k)}$ is the m-projection of p to $\boldsymbol{E}_k$. □

In order to obtain the full $\boldsymbol{\theta}$- or $\boldsymbol{\eta}$-coordinates of $p^{(k)}$, $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\theta}_{k-}^{(k)}, 0)$ and $\boldsymbol{\eta}^{(k)} = (\boldsymbol{\eta}_{k-}, \boldsymbol{\eta}_{k+}^{(k)})$, we need to solve the set of equations

$$\boldsymbol{\theta}_{k-}^{(k)} = \frac{\partial}{\partial \boldsymbol{\eta}_{k-}} \varphi\left(\boldsymbol{\eta}_{k-}, \boldsymbol{\eta}_{k+}^{(k)}\right) \tag{42}$$

$$0 = \frac{\partial}{\partial \boldsymbol{\eta}_{k+}} \varphi\left(\boldsymbol{\eta}_{k-}, \boldsymbol{\eta}_{k+}^{(k)}\right) \tag{43}$$

$$\boldsymbol{\eta}_{k-} = \frac{\partial}{\partial \boldsymbol{\theta}_{k-}} \psi\left(\boldsymbol{\theta}_{k-}^{(k)}, 0\right) \tag{44}$$

$$\boldsymbol{\eta}_{k+}^{(k)} = \frac{\partial}{\partial \boldsymbol{\theta}_{k+}} \psi\left(\boldsymbol{\theta}_{k-}^{(k)}, 0\right). \tag{45}$$

Note that, $\boldsymbol{\eta}_{k-}^{(k)} = \boldsymbol{\eta}_{k-}$, $\boldsymbol{\theta}_{k-}^{(k)} \neq \boldsymbol{\theta}_{k-}$. Hence, $\boldsymbol{\theta}_{k-}$ changes by the m-projection. This implies that $(\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ does not give any orthogonal decomposition of the effects of various orders and that $\boldsymbol{\theta}_k$ does not give the pure order k effect except for the case of $\boldsymbol{\theta}_{k+} = 0$.

D. Maximal Principle

The projection $p^{(k)}$ is closely related with the maximal entropy principle [27]. The projection $p^{(k)}$ belongs to $\boldsymbol{M}_k(p)$ which consists of all the probability distributions having the same k-marginal distributions as p, that is, the same $\boldsymbol{\eta}_{k-}$ coordinates. For any $q \in \boldsymbol{M}_k$, $p^{(k)}$ is its projection to $\boldsymbol{E}_k$. Hence, because of the Pythagoras theorem

$$D[q : p_0] = D\left[q : p^{(k)}\right] + D\left[p^{(k)} : p_0\right] \tag{46}$$

the minimizer of $D[q : p_0]$ for $q \in \boldsymbol{M}_k$ is $p^{(k)}$.

We have

$$\begin{aligned} D[q : p_0] &= \sum q(\boldsymbol{x}) \log q(\boldsymbol{x}) - \sum q(\boldsymbol{x}) \log p_0(\boldsymbol{x}) \\ &= -H[q] - \text{const} \end{aligned} \tag{47}$$

because $\sum q(\boldsymbol{x}) \log p_0(\boldsymbol{x})$ depends only on the marginal distributions of q and is constant for all $q \in \boldsymbol{M}_k$. Hence, we have the geometric form of the maximum principle [27].

Theorem 7: The projection $p^{(k)}$ of p to $\boldsymbol{E}_k$ is the maximizer of entropy among $q \in \boldsymbol{M}_k$ having the same k-marginals as p

$$p^{(k)} = \underset{q \in \boldsymbol{M}_k}{\arg\max}\ H[q]. \tag{48}$$

This relation is useful for calculating $p^{(k)}$.

E. Orthogonal Decomposition

The next problem is to single out the amount of the kth-order effects in $p(\boldsymbol{x}, \boldsymbol{\theta})$, by separating it from the others. This is not given by $\boldsymbol{\theta}_k$. The amount of the effect of order k is given by the divergence

$$D_k = D\left[p^{(k)} : p^{(k-1)}\right] \tag{49}$$

which is certified by the following orthogonal decomposition.

Theorem 8:

$$D[p : p_0] = \sum_{k=1}^{n} D\left[p^{(k)} : p^{(k-1)}\right]. \tag{50}$$

Theorem 8 shows that the kth-order effects are "orthogonally" decomposed in (50). The theorem might be thought as a trivial result from the Pythagoras theorem. This is so when $\boldsymbol{S}$ is a Euclidean space. However, this is not trivial in the case of a dually flat manifold. To show this, let us consider the "theorem of three squares" in a Euclidean space: Let p, q, r, s be four corners in a rectangular parallelpiped. Then

$$(\overline{ps})^2 = (\overline{pq})^2 + (\overline{qr})^2 + (\overline{rs})^2. \tag{51}$$

In a dually flat manifold, the Pythagoras theorem holds when one edge is an m-geodesic and the other is an e-geodesic. Hence, (51) cannot trivially be extended to this case, because $\overline{qr}$ cannot be an e-geodesic and an m-geodesic at the same time.

The theorem is proved by the properties of the orthogonal foliation. We show the simplest case. Let us consider

$$\{p_0\} = \boldsymbol{E}_0 \subset \boldsymbol{E}_1 \subset \boldsymbol{E}_2 \subset \boldsymbol{E}_3 = \boldsymbol{S}$$

and let $p^{(1)}$ and $p^{(2)}$ be the m-projections of p. We write their coordinates as

$$p\colon \boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3), \qquad \boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3) \tag{52}$$

$$p^{(2)}\colon \boldsymbol{\theta} = \left(\bar{\boldsymbol{\theta}}_1, \bar{\boldsymbol{\theta}}_2, 0\right), \qquad \boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \bar{\boldsymbol{\eta}}_3) \tag{53}$$

$$p^{(1)}\colon \boldsymbol{\theta} = \left(\bar{\bar{\boldsymbol{\theta}}}_1, 0, 0\right), \qquad \boldsymbol{\eta} = \left(\boldsymbol{\eta}_1, \bar{\bar{\boldsymbol{\eta}}}_2, \bar{\bar{\boldsymbol{\eta}}}_3\right). \tag{54}$$

Then, we see that the m-projection of $p^{(2)}$ to $\boldsymbol{E}_1$ is $p^{(1)}$, although the m-geodesic connecting $p^{(2)}$ and $p^{(1)}$ is not usually included in $\boldsymbol{E}_2$. This proves

$$D[p : p_0] = D\left[p : p^{(2)}\right] + D\left[p^{(2)} : p^{(1)}\right] + D\left[p^{(1)} : p^{(0)}\right]. \tag{55}$$

The general case is similar.

F. M-Flat Structure

Dually to the e-structure, we can study the m-hierarchical structure

$$\boldsymbol{M}_0 \subset \boldsymbol{M}_1 \subset \boldsymbol{M}_2 \subset \cdots \subset \boldsymbol{M}_n = \boldsymbol{S} \tag{56}$$

where $\boldsymbol{M}_{k-1}$ is an m-flat submanifold of $\boldsymbol{M}_k$. A typical example is the mixture family

$$\begin{aligned}\boldsymbol{S} = \Big\{ p(\boldsymbol{x}, \boldsymbol{\eta}) \mid p(\boldsymbol{x}, \boldsymbol{\eta}) \\ = \sum \eta_i \{q_i(x) - q_0(x)\} + \Big(1 - \sum \eta_i\Big) q_0(x) \Big\}\end{aligned} \tag{57}$$

where

$$\boldsymbol{M}_k = \{p(\boldsymbol{x}, \boldsymbol{\eta}) \mid \eta_{k+1} = \cdots = \eta_n = 0\}. \tag{58}$$

The e-projection of $p \in \boldsymbol{S}$ to $\boldsymbol{M}_k$ is defined by

$$\overline{p}^{(k)}(\boldsymbol{x}) = \underset{q \in \boldsymbol{M}_k}{\arg\min}\, \overline{D}[p, q] = \underset{q \in \boldsymbol{M}_k}{\arg\min}\, D[q : p]. \tag{59}$$

We can also show the orthogonal decomposition theorem

$$\overline{D}[p : p_0] = \sum_{k=1}^{n} \overline{D}\left[\overline{p}^{(k)} : \overline{p}^{(k-1)}\right] \tag{60}$$

where $\overline{p}^{(n)} = p$ and $\overline{p}^{(0)} = p_0 = q_0$.

The quantity

$$\overline{D}_k = \overline{D}[\overline{p}^{(k)} : \overline{p}^{(k-1)}]$$

is regarded as the kth-order effect of p in the mixture family or in a more general m-structure. A hierarchical MA model in time series analysis is another example of the m-structure [4], where the minimal principle holds instead of the maximal principle [4].

IV. Simple Example: Triplewise Interactions

The general results are explained by a simple example of the set of joint probability distributions $\boldsymbol{S}_3 = \{p(\boldsymbol{x})\}$, $\boldsymbol{x} = (x_1, x_2, x_3)$, of binary random variables X_1, X_2 and X_3, where $x_i = 0$ or 1. We can expand $\log p(\boldsymbol{x})$

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum \theta_{ij} x_i x_j + \theta_{123} x_1 x_2 x_3 - \psi \tag{61}$$

obtaining the log-linear model [2]. This shows that $\boldsymbol{S}_3$ is an exponential family. The canonical or e-affine coordinates are $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3; \theta_{12}, \theta_{23}, \theta_{31}; \theta_{123})$ which are partitioned as

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3) \tag{62}$$

$$\boldsymbol{\theta}_1 = (\theta_1, \theta_2, \theta_3), \qquad \boldsymbol{\theta}_2 = (\theta_{12}, \theta_{23}, \theta_{31}) \tag{63}$$

$$\boldsymbol{\theta}_3 = (\theta_{123}). \tag{64}$$

This defines a hierarchical e-structure in $\boldsymbol{S}_3$, where $\boldsymbol{\theta}_2$ represents pairwise interactions and $\boldsymbol{\theta}_3$ represents the triple interaction, although they are not orthogonal. The corresponding m-affine coordinates are partitioned as

$$\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3) \tag{65}$$

where $\boldsymbol{\eta}_1 = (\eta_1, \eta_2, \eta_3)$ $\boldsymbol{\eta}_2 = (\eta_{12}, \eta_{23}, \eta_{13})$, and $\boldsymbol{\eta}_3 = (\eta_{123})$, with

$$\eta_i = E[x_i] = \text{Prob}\,\{x_i = 1\} \tag{66}$$

$$\eta_{ij} = E[x_i x_j] = \text{Prob}\,\{x_i = x_j = 1\} \tag{67}$$

$$\eta_{123} = E[x_1 x_2 x_3] = \text{Prob}\,\{x_1 = x_2 = x_3 = 1\}. \tag{68}$$

We have a hierarchical e-structure

$$\boldsymbol{E}_0 \subset \boldsymbol{E}_1 \subset \boldsymbol{E}_2 \subset \boldsymbol{E}_3 = \boldsymbol{S}_3 \tag{69}$$

where $\boldsymbol{E}_0$ is a singleton $p_0(\boldsymbol{x}) = 1/8$ with $\boldsymbol{\theta} = 0$, $\boldsymbol{E}_1$ is defined by $\boldsymbol{\theta}_{1+} = (\boldsymbol{\theta}_2, \boldsymbol{\theta}_3) = 0$, $\boldsymbol{E}_2$ is defined by $\boldsymbol{\theta}_{2+} = \boldsymbol{\theta}_3 = 0$. On the other hand, $\boldsymbol{\eta}_{1-} = \boldsymbol{\eta}_1 = (\eta_1, \eta_2, \eta_3)$ gives the marginal distributions of X_1, X_2 and X_3, and $\boldsymbol{\eta}_{2-} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$ gives the all the pairwise marginal distributions.

Consider the two mixed cut coordinates: $\boldsymbol{\xi}_1 = (\boldsymbol{\eta}_{1-}; \boldsymbol{\theta}_{1+})$ and $\boldsymbol{\xi}_2 = (\boldsymbol{\eta}_{2-}; \boldsymbol{\theta}_{2+})$. Since $\boldsymbol{\theta}_{1+}$ are orthogonal to the coordinates that specify the marginal distributions $\boldsymbol{\eta}_{1-}$, $\boldsymbol{\theta}_{1+} = (\boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ represents the effect of mutual interactions of X_1, X_2, and X_3, independently of their marginals. Similarly, $\boldsymbol{E}_2$ defined by $\boldsymbol{\theta}_{123} = 0$ is composed of all the distributions which have no intrinsic triplewise interactions but pairwise correlations. The two distributions given by $\boldsymbol{\xi}_2 = (\boldsymbol{\eta}_{2-}, \boldsymbol{\theta}_3)$ and $\boldsymbol{\xi}_2' = (\boldsymbol{\eta}_{2-}, \boldsymbol{\theta}_3')$ have the same pairwise marginal distributions but differ only in the pure triplewise interactions. Since θ_{123} is orthogonal to $\boldsymbol{\eta}_{2-} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2)$, it represents purely triplewise interactions, as is well known in the log-linear model [2], [18].

The partitioned coordinates $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3)$ are not orthogonal, so that we cannot say that $\boldsymbol{\theta}_2$ summarizes all the pure pairwise correlations, except for the special case of $\boldsymbol{\theta}_3 = 0$. Given $p(\boldsymbol{x}, \boldsymbol{\theta})$, we need to separate pairwise correlations and triplewise interactions invariantly and obtain the "orthogonal" quantitative decomposition of these effects.

We project p to $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$, giving $p^{(1)}$ and $p^{(2)}$, respectively. Then, $p^{(1)}$ is the independent distribution having the same marginals as p, and $p^{(2)}$ is the distribution having the same pairwise marginals as p but no triplewise interaction. By putting

$$D_2 = D\left[p : p^{(2)}\right] \tag{70}$$

$$D_1 = D\left[p^{(2)} : p^{(1)}\right] \tag{71}$$

$$D_0 = D\left[p^{(1)} : p_0\right] \tag{72}$$

we have the decomposition

$$D[p : p_0] = D_2 + D_1 + D_0. \tag{73}$$

Here, D_2 represents the amount of purely triplewise interaction, D_1 the amount of purely pairwise interaction, and D_0 the degree of deviations of the marginals of p from the uniform distribution p_0. We have thus the "orthogonal" quantitative decomposition of interactions.

It is interesting to show a new type of decomposition of entropy or information from this result. We have

$$\begin{aligned} D[p : p_0] &= \sum p(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{p_0(\boldsymbol{x})} \\ &= -H(X_1, X_2, X_3) + 3 \end{aligned} \tag{74}$$

$$D\left[p^{(1)} : p_0\right] = -\sum_{i=1}^{3} H(X_i) + 3 \tag{75}$$

where $H(X_1, X_2, X_3)$ and $H(X_i)$ are the total and marginal entropies of $\boldsymbol{X} = (X_1, X_2, X_3)$. Hence,

$$H[X_1, X_2, X_3] = \sum_{i=1}^{3} H(X_i) - D\left[p : p^{(2)}\right] - D\left[p^{(2)} : p^{(1)}\right]. \quad (76)$$

One can define mutual information among X_1, X_2, and X_3 by

$$I[X_1, X_2, X_3] = \sum p(x_1, x_2, x_3) \log \frac{p(x_1, x_2, x_3)}{p(x_1)p(x_2)p(x_3)} = \sum H(X_i) - H[X_1, X_2, X_3]. \quad (77)$$

Then, (73) gives a new invariant positive decomposition of $I[X_1, X_2, X_3]$

$$I[X_1, X_2, X_3] = D\left[p : p^{(2)}\right] + D\left[p^{(2)} : p^{(1)}\right]. \quad (78)$$

In information theory, the mutual information among three variables is sometimes defined by

$$I_3(X_1 : X_2 : X_3) = I(X_1 : X_2) - I(X_1 : X_2 \mid X_3). \quad (79)$$

Unfortunately, this quantity is not necessarily nonnegative ([20], see also [24]). Hence, our decomposition is new and is completely different from the conventional one [20]

$$I(X_1, X_2, X_3) = I(X_1 : X_2) + I(X_2 : X_3) + I(X_1 : X_3) - I(X_1 : X_2 : X_3). \quad (80)$$

It is easy to obtain $p^{(1)}$ from given p. The coordinates of $p^{(2)}$ are obtained by solving (42)–(45). In the present case, the mixed cut of $p^{(2)}$ is $\boldsymbol{\xi}_2 = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2; 0)$. Hence, the $\boldsymbol{\eta}$-coordinates of $p^{(2)}$ are $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, , \overline{\eta}_{123})$, where the marginals η_i and η_{ij} are the same as p and $\overline{\eta}_{123}$ is determined such that θ_{123} becomes 0. Since we have

$$\theta_{123} = \log \frac{p_{111}p_{100}p_{010}p_{001}}{p_{110}p_{101}p_{011}p_{000}} \quad (81)$$

by putting $\theta_{123} = 0$, the $\overline{\eta}_{123}$ of $p^{(2)}$ is given by solving (82) shown at the bottom of the page.

V. Higher Order Interactions of Random Variables

A. Coordinate Systems of $\boldsymbol{S}_n$

Let $X_1, \ldots, X_n$ be n binary variables and let $p = p(\boldsymbol{x})$ be its probability, $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_i = 0, 1$. We assume that $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$. The set of all such probability distributions is a $(2^n - 1)$-dimensional manifold $\boldsymbol{S}_n$, where 2^n probabilities

$$p_{i_1 \cdots i_n} = \mathrm{Prob}\{X_1 = i_1, \ldots, X_n = i_n\}, \qquad i_k = 0, 1 \quad (83)$$

constitute a coordinate system in which one of them is determined from

$$\sum_{i_1, \ldots, i_n} p_{i_1 \cdots i_n} = 1. \quad (84)$$

The manifold $\boldsymbol{S}_n$ is an exponential family and is also a mixture family at the same time. Let us expand $\log p(\boldsymbol{x})$ to obtain the log-linear model [2], [17]

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \sum_{i<j<k} \theta_{ijk} x_i x_j x_k \cdots + \theta_{1 \cdots n} x_1 \cdots x_n - \psi \quad (85)$$

where indexes of θ_{ijk}, etc., satisfy $i < j < k$, etc. Then

$$\boldsymbol{\theta} = (\theta_i, \theta_{ij}, \ldots, \theta_{1 \cdots n}) \quad (86)$$

has $2^n - 1$ components. They define the e-flat coordinate system of $\boldsymbol{S}_n$.

In order to simplify index notation, we introduce the following two hyper-index notations. Indexes I, J, K, etc., run over n-tuples of binary numbers

$$I = (i_1, i_2, \ldots, i_n), \qquad i_1, \ldots, i_n = 0, 1 \quad (87)$$

except for $(0, \ldots, 0)$. Hence, the cardinality $|I|$ of the index set is $2^n - 1$. Indexes A, B, C, etc., run over the set consisting of a single index i, a pair of indexes (i, j) where $i < j$, a triple of indexes (i, j, k) where $i < j < k, \ldots$, and n-tuple $(1, 2, \ldots, n)$, that is, A stands for any element in the set $\{i, \ ij, \ ijk, \ldots, 1 \cdots n\}$.

In terms of these indexes, the two coordinate systems given by (83) and (86) are written as

$$\boldsymbol{p} = (p_I) \quad (88)$$

$$\boldsymbol{\theta} = (\theta_A) \quad (89)$$

respectively. We now study the coordinate transformations among them.

Let us consider $2^n - 1$ functions of $\boldsymbol{x} = (x_1, \ldots, x_n)$ defined by

$$\delta_I(\boldsymbol{x}) = \begin{cases} 1, & \boldsymbol{x} = I \\ 0, & \text{otherwise.} \end{cases} \quad (90)$$

Here, $\boldsymbol{x} = I$ implies that, for $I = (i_1, \ldots, i_n)$, $x_1 = i_1, \ldots, x_n = i_n$. Each $\delta_I(\boldsymbol{x})$ is a polynomial of degree n, written as

$$\delta_I(\boldsymbol{x}) = x_1^{i_1} \cdots x_n^{i_n} \quad (91)$$

where we put

$$x^i = \begin{cases} x, & i = 1 \\ 1 - x, & i = 0. \end{cases} \quad (92)$$

The polynomials can be expanded as

$$\delta_I(\boldsymbol{x}) = \sum b_i^I x_i + \sum b_{ij}^I x_i x_j + \cdots + \sum b_{1 \cdots n}^I x_1 \cdots x_n \quad (93)$$

or, shortly, as

$$\delta_I(\boldsymbol{x}) = \sum_A B_A^I X_A \quad (94)$$

$$\overline{\eta}_{123} = \frac{(\eta_{12} - \overline{\eta}_{123})(\eta_{13} - \overline{\eta}_{123})(\eta_{23} - \overline{\eta}_{123})(1 - \eta_1 - \eta_2 - \eta_3 + \eta_{12} + \eta_{23} + \eta_{13} - \overline{\eta}_{123})}{(\eta_1 - \eta_{12} - \eta_{13} + \overline{\eta}_{123})(\eta_2 - \eta_{23} - \eta_{12} + \overline{\eta}_{123})(\eta_3 - \eta_{13} - \eta_{23} + \overline{\eta}_{123})}. \quad (82)$$

where we set

$$X_A = x_{i_1} \cdots x_{ik} \tag{95}$$

when A is $i_1 \cdots i_k$. The matrix $\boldsymbol{B} = (B_A^I)$ is a $(2^n - 1) \times (2^n - 1)$ nonsingular matrix. We have

$$X_A = \sum_I \left(\boldsymbol{B}^{-1}\right)_A^I \delta_I(\boldsymbol{x}) \tag{96}$$

where $(\boldsymbol{B}^{-1})_A^I$ are the elements of the inverse of $\boldsymbol{B}$. It is not difficult to write down these elements explicitly.

The probability distributions of $\boldsymbol{S}_n$ is rewritten as

$$p(\boldsymbol{x}, \boldsymbol{p}_I) = \sum p_I \, \delta_I(\boldsymbol{x}) + \left(1 - \sum p_I\right) \delta_0(\boldsymbol{x}) \tag{97}$$

where $\delta_0(\boldsymbol{x}) = 1$ when $\boldsymbol{x} = (0, \ldots, 0)$ and 0 otherwise. This shows that $\boldsymbol{S}_n$ is a mixture family. The expansion (85) is rewritten as

$$p(\boldsymbol{x}, \boldsymbol{\theta}) = \exp\left\{\sum \theta_A X_A - \psi(\boldsymbol{\theta})\right\}. \tag{98}$$

This shows that $\boldsymbol{S}_n$ is an exponential family.

The corresponding m-affine coordinates are given by the expectation parameters $\boldsymbol{\eta} = (\eta_A)$

$$\eta_A = E[X_A]. \tag{99}$$

For $A = i_1, \ldots i_k$

$$\eta_A = E\left[x_{i_1} \cdots x_{i_k}\right] = \text{Prob}\,\{x_{i_1} = 1, \ldots, x_{i_k} = 1\}. \tag{100}$$

The relations among the coordinate systems p_I, θ_A, and η_A are given by the following theorem.

Theorem 9:

$$\log(p_I / p_{0\cdots0}) = \sum_A \left(\boldsymbol{B}^{-1}\right)_A^I \theta_A \tag{101}$$

$$\theta_A = \sum \boldsymbol{B}_A^I \log(p_I / p_{0\cdots0}) \tag{102}$$

$$p_I = \sum_A \boldsymbol{B}_A^I \eta_A, \qquad \eta_A = \sum \left(\boldsymbol{B}^{-1}\right)_A^I p_I \tag{103}$$

$$\theta_A = \sum_I \boldsymbol{B}_A^I \log\left\{\left(\sum_B \boldsymbol{B}_B^I \eta_B\right) \Big/ p_{0\cdots0}\right\} \tag{104}$$

$$\eta_A = p_{00\cdots0} \sum \left(\boldsymbol{B}^{-1}\right)_A^I \exp\left\{\sum \left(\boldsymbol{B}^{-1}\right)_B^I \theta_B\right\}. \tag{105}$$

Proof: Given a probability distribution $p(\boldsymbol{x})$, we have

$$p(\boldsymbol{x}) = \sum p_I \, \delta_I(\boldsymbol{x}), \qquad p_{0\cdots0} = 1 - \sum_I p_I \tag{106}$$

$$\log p(\boldsymbol{x}) = \sum (\log p_I) \, \delta_I(\boldsymbol{x}) + (\log p_{0\cdots0}) \left[1 - \sum_I \delta_I(\boldsymbol{x})\right]. \tag{107}$$

On the other hand, from (85) we have

$$\log p(\boldsymbol{x}) = \sum_A \theta_A X_A = \sum \theta_A \left(\boldsymbol{B}^{-1}\right)_A^I \delta_I(\boldsymbol{x}) \tag{108}$$

which proves (103). We have

$$\eta_A = E[X_A] = \sum_I \left(\boldsymbol{B}^{-1}\right)_A^I E[\delta_I(\boldsymbol{x})]. \tag{109}$$

□

B. E-Hierarchical Structure of $\boldsymbol{S}_n$

Let us naturally decompose the $\boldsymbol{\theta}$ coordinates into

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_n) \tag{110}$$

where $\boldsymbol{\theta}_k$ summarizes θ_A with $|A| = k$, that is, those consisting of k indexes $i_1 \cdots i_k$. Hence $\boldsymbol{\theta}_k$ has ${}_nC_k$ components. Let $\boldsymbol{E}_k$ be the subspace defined by $\boldsymbol{\theta}_{k+} = 0$. Then, $\boldsymbol{E}_k$'s form an e-hierarchical structure.

We consider the corresponding partition of $\boldsymbol{\eta}$

$$\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n). \tag{111}$$

Then, the coordinates $\boldsymbol{\eta}_{k-} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k)$ consist of all

$$\eta_{i_1 \cdots i_s} = \text{Prob}\,\{x_{i_1} = \cdots = x_{i_s} = 1\} \tag{112}$$

for $1 \le s \le k$. All the marginal distributions of k random variables $x_{i_1}, \ldots, x_{i_k}$ are completely determined by $\boldsymbol{\eta}_{k-}$. For any k, we can define the k-cut coordinates

$$\boldsymbol{\xi}_k = (\boldsymbol{\eta}_{k-}; \boldsymbol{\theta}_{k+}). \tag{113}$$

C. Orthogonal Decomposition of Interactions and Entropy

Given $p(\boldsymbol{x})$, $p^{(k)}(\boldsymbol{x}) = \prod^{(k)} p$ is the point closest to p among those that do not have intrinsic interactions more than k variables. The amount of interactions higher than k is defined by $D[p : p^{(k)}]$. Moreover, $D_k(p) = D[p^{(k)} : p^{(k-1)}]$ may be interpreted as the degree of purely order k interactions among n variables. We then have the following decomposition.

From the orthogonal decomposition, we have a new invariant decomposition of entropy and information, which is different from those studied by many researchers as the decomposition of entropy of a number of random variables.

Theorem 10:

$$D[p : p_0] = \sum_{k=1}^{n} D_k \tag{114}$$

$$H(X_1, \ldots, X_n) = \sum_{i=1}^{n} H(X_i) - \sum_{k=2}^{n} D_k \tag{115}$$

$$I(X_1, \ldots, X_n) = \sum D_k \tag{116}$$

where

$$I(X_1, \ldots, X_n) = \sum H(X_i) - H(X_1, \ldots, X_n).$$

D. Extension to Finite Alphabet

We have so far treated binary random variables. The hierarchical structure of interactions and the decomposition of divergence or entropy holds in the general case. We consider the case where $X_1, \ldots, X_n$ are random variables taking on a common finite alphabet set $\mathcal{A} = \{0, 1, \ldots, m\}$. The argument is extended easily to the case where X_i take on different alphabet sets $\mathcal{A}_i$. Let us define the indicator functions for X_i by

$$\delta_i^a(x_i) = \begin{cases} 1, & x_i = a \\ 0, & x_i \neq a \end{cases} \tag{117}$$

where a denotes any element in $\mathcal{A}$.

Let the joint probability of $\boldsymbol{x} = (a_1, \ldots, a_n)$ be $p_{1\cdots n}^{a_1\;a_n}$. The joint probability function $p(\boldsymbol{x})$ is now written as

$$p(\boldsymbol{x}) = \sum_{a_1, \ldots, a_n} p_{1\cdots n}^{a_1\cdots a_n}\,\delta_1^{a_1}(x_1)\,\delta_2^{a_2}(x_2)\cdots\delta_n^{a_n}(x_n). \tag{118}$$

We now expand $\log p(\boldsymbol{x})$ as

$$\log p(\boldsymbol{x}) = \sum_{k=1}^{n}\sum\sum\theta_{i_1\cdots i_k}^{a_1\cdots a_k}\,\delta_{i_1}^{a_1}(x_1)\cdots\delta_{i_k}^{a_k}(x_k) - \psi \tag{119}$$

where a_i stands for nonzero elements of $\mathcal{A}$ and $i_1 < \cdots < i_k$. The coefficients θ's form the θ-coordinate system. Here, the term representing interactions of k variables $X_{i_1}, \ldots, X_{i_k}$ have a number of components as $\theta_{i_1\cdots i_k}^{a_1\cdots a_k}$.

The corresponding η coordinates consist of

$$\begin{aligned} \eta_{i_1\cdots i_k}^{a_1\cdots a_k} &= E\left[\delta_{i_1}^{a_1}(x_{i_1})\cdots\delta_{i_k}^{a_k}(x_{i_k})\right] \\ &= \mathrm{Prob}\left\{x_{i_1} = a_1, \ldots, x_{i_k} = a_k\right\} \end{aligned} \tag{120}$$

which represent the marginal distributions of k variables $X_{i_1}, \ldots, X_{i_k}$.

We can then define the k-cut, $(\boldsymbol{\eta}_{k-}, \boldsymbol{\theta}_{k+})$ in which $\boldsymbol{\eta}_{k-}$ and $\boldsymbol{\theta}_{k+}$ are orthogonal. Given $p(\boldsymbol{x})$, we can define

$$p^{(k)}(\boldsymbol{x}) = \prod^{(k)} p(\boldsymbol{x})$$

which has the same marginals as $p(\boldsymbol{x})$ up to k variables, but has no intrinsic interactions more than k random variables. The orthogonal decompositions of D and mutual information I in terms of D_k's hold as before.

E. Continuous Distributions

It was difficult to give a rigorous mathematical foundation to the function space $\boldsymbol{S} = \{p(\boldsymbol{x})\}$ of all the density functions $p(\boldsymbol{x}) > 0$ with respect to the Lebesgue measure on the real $\boldsymbol{R}^n$. Recently, Pistone *et al.* [35], [36] have succeeded in constructing information geometry in the infinite-dimensional case, although we do not enter in its mathematical details.

Roughly speaking, the η-coordinates of $\boldsymbol{S}$ is the density function itself, $\eta(\boldsymbol{x}) = p(\boldsymbol{x})$, and the θ-coordinates are the function

$$\theta(\boldsymbol{x}) = \log p(\boldsymbol{x}). \tag{121}$$

They are dually coupled.

We define higher order interactions among n variables $x_1, \ldots, x_n$. In order to avoid notational complications, we show only the case of $n = 3$. The three marginals and three joint marginals of two variables are given by $\eta_i(x_i)$ and $\eta_{ij}(x_i, x_j)$, respectively. They are obtained by integrating $\eta_{123}(x_1, x_2, x_3)$.

The corresponding θ-coordinates are given by

$$\theta_i(x_i) = \log\frac{p(0, x_i, 0)}{p(0, 0, 0)} \tag{122}$$

$$\theta_{ij}(x_i, x_j) = \log\frac{p(x_i, x_j, 0)p(0, 0, 0)}{p(x_i, 0, 0)p(0, x_j, 0)} \tag{123}$$

$$\begin{aligned} &\theta_{123}(x_1, x_2, x_3) \\ &\quad = \log\frac{p(x_1, x_2, x_3)p(x_1, x_2, 0)p(x_1, 0, x_3)p(0, x_2, x_3)}{p(x_1, 0, 0)p(0, x_2, 0)p(0, 0, x_3)p(0, 0, 0)}. \end{aligned} \tag{124}$$

When $\theta_{123}(x_1, x_2, x_3) = 0$ holds identically, there are no intrinsic interactions among the three, so that all interactions are pairwise in this case. When $\theta_{ij} = 0$, $\theta_{123} = 0$, there are no interactions so that the three random variables are independent.

Given $p(x_1, x_2, x_3)$, the independent distribution $p^{(1)} \in \boldsymbol{E}_1$ is easily obtained by

$$p^{(1)}(x_1, x_2, x_3) = p(x_1)p(x_2)p(x_3). \tag{125}$$

However, it is difficult to obtain the analytical expression of $p^{(2)}$. In general, $\overline{p} \in \boldsymbol{E}_2$ which consists of distributions having pairwise correlations but no intrinsic triplewise interactions, is written in the form

$$\log\overline{p}(x_1, x_2, x_3) = \sum_{i=1}^{3}\theta_i(x_i) + \sum\theta_{ij}(x_i, x_j) - \psi. \tag{126}$$

The $p^{(2)}$ is the one that satisfies

$$\overline{p}(x_i) = p(x_i) \tag{127}$$

$$\overline{p}(x_i, x_j) = p(x_i, x_j). \tag{128}$$

VI. Higher Order Markov Processes

As another example of the e-structure, we briefly touch upon the higher order Markov processes of the binary alphabet. The kth-order Markov process is specified by the conditional probability of the next output x

$$p\left(x \mid x^k\right) \tag{129}$$

where $x^k = x_1x_2\cdots x_k$ is the past sequence of k letters. We define functions α: $\{0, 1\}^k \to \boldsymbol{R}$ and $\overline{\alpha}$

$$\alpha\left(x^k\right) = \log p\left(1 \mid x^k\right) \tag{130}$$

$$\overline{\alpha}\left(x^k\right) = \log p\left(0 \mid x^k\right) = \log\left\{1 - p\left(1 \mid x^k\right)\right\}. \tag{131}$$

Then, the Markov chain is specified by 2^n parameters $\alpha(x^k)$, x^k taking on any binary sequences of length k.

For an observed long data sequence $x^T = x_1\cdots x_T$, let $f(x^k)$ be the relative number of subsequence x^k appearing in x^T. The Markov chain is an exponential family when T is sufficiently large, and $f(x^k)$'s are sufficient statistics. The probability of x^T is written in terms of the relative frequencies of various $k+1$ letter sequences $x^k1 = x_1\cdots x_k1$ and $x^k0 = x_1\cdots x_k0$

$$p(x^T, \alpha) = \exp\left\{\sum\alpha(x^k)\,f(x^k1) + \sum\overline{\alpha}(x^k)\,f(x^k0)\right\}. \tag{132}$$

Hence, the kth-order Markov chain forms an e-flat manifold $\boldsymbol{E}_k$ of 2^k dimensions. The set of Markov chains of various orders naturally has the hierarchical structure

$$\boldsymbol{E}_0 \subset \boldsymbol{E}_1 \subset \boldsymbol{E}_2 \subset \cdots \tag{133}$$

where $\boldsymbol{E}_0$ is the Bernoulli process (independent and identically distributed process) and $\boldsymbol{E}_1$ is the first-order Markov chain.

In order to decompose the degree of dependency of a Markov chain into various orders, we use the following frequencies observed from the sequence x^T. In order to avoid complicated notations, we use as an exemple the second-order Markov chain, but generalization is easy. We define random variables related to a sequence by

$$f_{1\cdot\cdot} = \text{ relative frequency of 1} \tag{134}$$

$$f_{11\cdot} = \text{ relative frequency of 11} \tag{135}$$

$$f_{1\cdot 1} = \text{ relative frequency of } 1x1 \tag{136}$$

$$\text{where } x \text{ is arbitrary}$$

$$f_{111} = \text{ relative frequency of 111.} \tag{137}$$

We summarize their expectations as

$$\begin{aligned} \eta_1 &= E[f_{1\cdot\cdot}], \qquad \eta_{11} = E[f_{11\cdot}] \\ \eta_{1\cdot 1} &= E[f_{1\cdot 1}], \qquad \eta_{111} = E[f_{111}]. \end{aligned} \tag{138}$$

They form a mixture affine coordinate system to specify the second-order Markov chain, and are functions of $\alpha(00)$, $\alpha(01)$, $\alpha(10)$, and $\alpha(11)$. Higher order generalization is easy. In order to obtain the third-order Markov chain, for example, we obtain eight f's by adding $\cdot$ and 1 in the suffix of each f, for example, $f_{1\cdots}$ and $f_{1\cdot\cdot 1}$ emerge from $f_{1\cdots}$. We then have eight quantities whose expectations form the η-coordinates.

The coordinate η_1 is responsible only for the Bernoulli structure. Therefore, if the process is Bernoulli, η_1 determines its probability distribution. The coordinates η_1 and η_{11} together are responsible for the first-order Markov structure and all of η_1, η_{11}, $\eta_{1\cdot 1}$, η_{111} are necessary to determine the second-order Markov structure. Therefore, we have the following decomposition of the η-coordinates:

$$\boldsymbol{\eta} = (\boldsymbol{\eta}_0, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots) \tag{139}$$

where $\boldsymbol{\eta}_0 = \eta_1, \boldsymbol{\eta}_1 = \eta_{11}, \boldsymbol{\eta}_2 = (\eta_{1\cdot 1}, \eta_{111}), \ldots$. The parameters $\boldsymbol{\eta} = (\boldsymbol{\eta}_0, \boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k)$ define the kth-order Markov structure.

We have the corresponding $\boldsymbol{\theta}$-coordinates, because $\boldsymbol{f}$'s are linear combinations of $f(x^k 1)$'s. They are given in the partitioned form in this case as

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_0, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots) \tag{140}$$

$$\boldsymbol{\theta}_0 = \theta_1 = \alpha(00) + \overline{\alpha}(10) + \overline{\alpha}(01) - 3\overline{\alpha}(00) \tag{141}$$

$$\boldsymbol{\theta}_1 = \theta_{11} = \alpha(01) + \overline{\alpha}(11) + 2\overline{\alpha}(00) - \alpha(00) \tag{142}$$

$$- \overline{\alpha}(10) - 2\overline{\alpha}(01) \tag{143}$$

$$\boldsymbol{\theta}_2 = (\theta_{1\cdot 1}, \theta_{111}) \tag{144}$$

$$\theta_{1\cdot 1} = \alpha(10) + \overline{\alpha}(00) - \alpha(00) - \overline{\alpha}(10) \tag{145}$$

$$\theta_{111} = \alpha(11) + \alpha(00) + \overline{\alpha}(01) + \overline{\alpha}(10) \tag{146}$$

$$- \alpha(10) - \alpha(01) - \overline{\alpha}(00) - \overline{\alpha}(11). \tag{147}$$

We can see here that $\theta_{1\cdot 1}$ and θ_{111} together show how the Markov structure is apart from the first-order one, because those coordinates are orthogonal to $\boldsymbol{M}_1$.

The projections to lower order Markov chains are defined in the same way as before, and we have the following decomposition:

$$D[p : p_0] = \sum_{i=0}^{k} D\left[p^{(i)} : p^{(i-1)}\right] \tag{148}$$

where $p^{(i)}$ is the projection to the ith-order Markov chain and $p^{(-1)}$ is a fair Bernoulli process (that is, $\eta_1 = 1/2$).

VII. Conclusion

The present paper used information geometry to elucidate the hierarchical structures of random variables and their quasi-orthogonal decomposition. A typical example is the decomposition of interactions among n binary random variables into a quasi-orthogonal sum of interactions among exactly k random variables. The Kullback–Leibler divergence, entropy, and information are decomposed into a sum of nonnegative quantities representing the order k interactions. This is a new result in information theory. This problem is important for analyzing joint firing patterns of an ensemble of neurons. The present paper gives a method of extracting various degrees of interactions among these neurons.

The present theories of hierarchical structures can be applicable to more general cases including mixture families. Applications to higher order Markov chains are briefly described.

References

[1] A. M. H. J. Aertsen, G. L. Gerstein, M. K. Habib, and G. Palm, "Dynamics of neuronal firing correlation: Modulation of 'effective connectivity'," *J. Neurophysiol.*, vol. 61, pp. 900–918, 1989.

[2] A. Agresti, *Categorical Data Analysis*, New York: Wiley, 1990.

[3] S. Amari, *Differential-Geometrical Methods of Statistics (Lecture Notes in Statistics)*. Berlin, Germany: Springer, 1985, vol. 25.

[4] ——, "Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections and divergence," *Math. Syst. Theory*, vol. 20, pp. 53–82, 1987.

[5] ——, "Fisher information under restriction of Shannon information," *Ann. Inst. Statist. Math.*, vol. 41, pp. 623–648, 1989.

[6] ——, "Dualistic geometry of the manifold of higher-order neurons," *Neural Networks*, vol. 4, pp. 443–451, 1991.

[7] ——, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, vol. 8, no. 9, pp. 1379–1408, 1995.

[8] ——, "Information geometry," *Contemp. Math.*, vol. 203, pp. 81–95, 1997.

[9] ——, "Natural gradient works efficiently in learning," *Neural Comput.*, vol. 10, pp. 251–276, 1998.

[10] S. Amari, S. Ikeda, and H. Shimokawa, "Information geometry of α-projection in mean-field approximation," in *Recent Developments of Mean Field Approximation*, M. Opper and D. Saad, Eds. Cambridge, MA: MIT Press, 2000.

[11] S. Amari and T. S. Han, "Statistical inference under multi-terminal rate restrictions—A differential geometrical approach," *IEEE Trans. Inform. Theory*, vol. 35, pp. 217–227, Jan. 1989.

[12] S. Amari and M. Kawanabe, "Information geometry of estimating functions in semi parametric statistical models," *Bernoulli*, vol. 3, pp. 29–54, 1997.

[13] S. Amari, K. Kurata, and H. Nagaoka, "Information geometry of Boltzmann machines," *IEEE Trans. Neural Networks*, vol. 3, pp. 260–271, Mar. 1992.

[14] S. Amari and H. Nagaoka, *Methods of Information Geometry*. New York: AMS and Oxford Univ. Press, 2000.

[15] O. E. Barndorff-Nielsen, *Information and Exponential Families in Statistical Theory*, New York: Wiley, 1978.

[16] C. Bhattacharyya and S. S. Keerthi, "Mean-field methods for stochastic connections networks," *Phys. Rev.*, to be published.

[17] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press, 1975.

[18] L. L. Campbell, "The relation between information theory and the differential geometric approach to statistics," *Inform. Sci.*, vol. 35, pp. 199–210, 1985.

[19] N. N. Chentsov, *Statistical Decision Rules and Optimal Inference* (in Russian). Moscow, U.S.S.R.: Nauka, 1972. English translation: Providence, RI: AMS, 1982.

[20] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[21] I. Csiszár, "On topological properties of f-divergence," *Studia Sci. Math. Hungar.*, vol. 2, pp. 329–339, 1967.

[22] ——, "I-divergence geometry of probability distributions and minimization problems," *Ann. Probab.*, vol. 3, pp. 146–158, 1975.

[23] I. Csiszár, T. M. Cover, and B. S. Choi, "Conditional limit theorems under Markov conditioning," *IEEE Trans. Inform. Theory*, vol. IT-33, pp. 788–801, Nov. 1987.

[24] T. S. Han, "Nonnegative entropy measures of multivariate symmetric correlations," *Inform. Contr.*, vol. 36, pp. 133–156, 1978.

[25] T. S. Han and S. Amari, "Statistical inference under multiterminal data compression," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2300–2324, Oct. 1998.

[26] H. Ito and S. Tsuji, "Model dependence in quantification of spike interdependence by joint peri-stimulus time histogram," *Neural Comput.*, vol. 12, pp. 195–217, 2000.

[27] E. T. Jaynes, "On the rationale of maximum entropy methods," *Proc. IEEE*, vol. 70, pp. 939–952, 1982.

[28] R. E. Kass and P. W. Vos, *Geometrical Foundations of Asymptotic Inference*, New York: Wiley, 1997.

[29] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Science, 1996.

[30] M. K. Murray and J. W. Rice, *Differential Geometry and Statistics*. London, U.K.: Chapman & Hall, 1993.

[31] H. Nagaoka and S. Amari, "Differential geometry of smooth families of probability distributions," Univ. Tokyo, Tokyo, Japan, METR, 82-7, 1982.

[32] A. Ohara, N. Suda, and S. Amari, "Dualistic differential geometry of positive definite matrices and its applications to related problems," *Linear Alg. its Applic.*, vol. 247, pp. 31–53, 1996.

[33] A. Ohara, "Information geometric analysis of an interior point method for semidefinite programming," in *Geometry in Present Day Science*, O. E. Barndorff-Nielsen and E. B. V. Jensen, Eds. Singapore: World Scientific, 1999, pp. 49–74.

[34] G. Palm, A. M. H. J. Aertsen, and G. L. Gerstein, "On the significance of correlations among neuronal spike trains," *Biol. Cybern.*, vol. 59, pp. 1–11, 1988.

[35] G. Pistone and C. Sempi, "An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one," *Ann. Statist.*, vol. 23, pp. 1543–1561, 1995.

[36] G. Pistone and M.-P. Rogantin, "The exponential statistical manifold: Mean parameters, orthogonality, and space transformation," *Bernoulli*, vol. 5, pp. 721–760, 1999.

[37] C. R. Rao, "Information and accuracy attainable in the estimation of statistical parameters," *Bull. Calcutta. Math. Soc.*, vol. 37, pp. 81–91, 1945.

[38] T. Tanaka, "Information geometry of mean field approximation," *Neural Comput.*, vol. 12, pp. 1951–1968, 2000.