ARTICLE Communicated by Johnathan Victor

Information-Geometric Measure for Neural Spikes

Hiroyuki Nakahara
hiro@brain.riken.go.jp
Shun-ichi Amari
amari@brain.riken.go.jp
Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Wako, Saitama, 351-0198, Japan

This study introduces information-geometric measures to analyze neural firing patterns by taking not only the second-order but also higher-order interactions among neurons into account. Information geometry provides useful tools and concepts for this purpose, including the orthogonality of coordinate parameters and the Pythagoras relation in the Kullback-Leibler divergence. Based on this orthogonality, we show a novel method for analyzing spike firing patterns by decomposing the interactions of neurons of various orders. As a result, purely pairwise, triple-wise, and higher-order interactions are singled out. We also demonstrate the benefits of our proposal by using several examples.

1 Introduction

One of the central challenges in neuroscience is to understand what and how information is carried by a population of neural firing (Georgopoulos, Schwartz, & Kettner, 1986; Abeles, 1991; Aertsen & Arndt, 1993; Singer & Gray, 1995; Deadwyler & Hampson, 1997; Parker & Newsome, 1998). Many experimental studies have shown, as a first step toward this end, that the mean firing rate of each single neuron can be significantly modulated by experimental conditions and may thereby carry information about these experimental conditions, that is, sensory and motor signals. Information conveyed by a population of firing neurons, however, may be not only a sum of mean firing rates. Other statistical structures embedded in the neural firing may also carry behavioral information. In particular, growing attention has been paid to the possibility that coincident firing, correlated firing, synchronization, or specific firing patterns may alter conveyed information or carry significant behavioral information, whether such a possibility is supported or discarded (Gerstein, Bedenbaugh, & Aertsen, 1989; Engel, König, Kreiter, Schillen, & Singer, 1992; Wilson & McNaughton, 1993; Zohary, Shadlen, & Newsome, 1994; Vaadia et al., 1995; Nicolelis, Ghazanfar, Faggin, Votaw, & Oliveira, 1997; Riehle, Grün, Diesmann, & Aertsen, 1997; Lisman, 1997;

Zhang, Ginzburg, McNaughton, & Sejnowski, 1998; Maynard et al., 1999; Nadasdy, Hirase, Czurko, Csicsvari, & Buzsaki, 1999; Kudrimoti, Barnes, & McNaughton, 1999; Oram, Wiener, Lestienne, & Richmond, 1999; Nawrot, Aertsen, & Rotter, 1999; Baker & Lemon, 2000; Reinagel & Reid, 2000; Steinmetz et al., 2000; Laubach, Wessberg, & Nicolelis, 2000; Salinas & Sejnowski, 2001; Oram, Hatsopoulas, Richmond, & Donoghue, 2001). For this purpose, it is important to develop a sound statistical method for analyzing neural data. An obvious first step is to investigate a significant coincident firing between two neurons, that is, the pairwise correlation (Perkel, Gerstein, & Moore, 1967; Palm, 1981; Gerstein & Aertsen, 1985; Palm, Aertsen, & Gerstein, 1988; Aertsen, Gerstein, Habib, & Palm, 1989; Grün, 1996; Ito & Tsuji, 2000; Pauluis & Baker, 2000; Roy, Steinmetz, & Niebur, 2000; Grün, Diesmann, & Aertsen, 2002a, 2002b; Gütig, Aertsen, & Rotter, 2002).

In general, however, it is not sufficient to test a pairwise correlation of neural firing because there can be triplewise and higher correlations. For example, three variables (neurons) are not independent in general even when they are pairwise independent. We need to establish a systematic method of analysis that includes these higher-order correlations (Abeles & Gerstein, 1988; Abeles, Bergman, Margalit, & Vaadia, 1993; Martignon, Von Hasseln, Grün, Aertsen, & Palm, 1995; Grün, 1996; Tetko & Villa, 1992; Victor & Purpura, 1997; Prut et al., 1998; Del Prete & Martingon, 1998; MacLeod, Bäcker, & Laurent, 1998; Martignon et al., 2000; Bohte, Spekreijse, & Roelfsma, 2000; Roy et al., 2000). We are mostly interested in methods able to (1) analyze correlated firing of neurons, including higher-order interactions, and (2) connect such a technique with behavioral events, for which we use mutual information between firing and behavior (Tsukada, Ishii, & Sato, 1975; Optican & Richmond, 1987; Richmond, Optican, & Spitzer, 1990; McClurkin, Gawne, Optican, & Richmond, 1991; Bialek, Rieke, de Ruyter van Steveninck, & Warland, 1991; Gawne & Richmond, 1993; Tovee, Rolls, Treves, & Bellis, 1993; Gochin, Colombo, Dorfman, Gerstein, & Gross, 1994; Abbott, Rolls, & Tovee, 1996; Rolls, Treves, & Tovee, 1997; Richmond & Gawne, 1998; Kitazawa, Kimura, & Yin, 1998; Sugase, Yamane, Ueno, & Kawano, 1999; Panzeri, Schultz, Treves, & Rolls, 1999; Panzeri, Treves, Schultz, & Rolls, 1999; Brenner, Strong, Koberle, Bialek, & de Ruyter van Steveninck, 2000; Samengo, Montagnini, & Treves, 2000; Panzeri & Schultz, 2001).

To address these issues, this study uses the orthogonality of the natural and expectation parameters in the exponential family of distributions and proposes methods useful for analyzing a population of neural firing in a systematic manner, based on information geometry (Amari, 1985; Amari & Nagaoka, 2000) and the theory of hierarchical structure (Amari, 2001). By use of orthogonal coordinates, we will show that both hypothesis testing of neural interaction and calculation of mutual information can be drastically simplified. (For an extended abstract, see Nakahara & Amari, 2002.)

This article is organized as follows. In section 2, we briefly give our perspective on the merits of using an information-geometric measure. In

section 3, we begin with an introductory description of information geometry, using two random binary variables, and treat the application of this two variables' case to the analysis of two neurons' firing. Section 4 discusses the interaction of three binary variables and shows how to extract pure triple-wise correlation, which is different from pairwise correlation. Section 5 gives a general theory of decomposition of correlations among n variables and discusses some approaches to overcome practical difficulties that arise in this case. Section 6 provides illustrative examples. Section 7 contains the discussion.

2 Perspective

In this section, we state our perspective on the merits of using an information-geometric measure, briefly referring to a general case of n neurons. A detailed discussion in the general case is given in section 5.

We represent a neural firing pattern by a binary random vector variable so that the probability distribution of firing (of any number of neurons) can be exactly expanded by a log-linear model. Let $X = (X_1, \ldots, X_n)$ be n binary variables, and let $p = p(\boldsymbol{x}), \boldsymbol{x} = (x_1, \ldots, x_n), x_i = 0, 1$ be its probability, where we assume $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$. Each X_i indicates that the ith neuron is silent ($X_i(t_i) = 0$) or has a spike ($X_i(t_i) = 1$) in a short time bin, which is denoted by t_i. In general, t_i can be different for each neuron, but here, we assume $t_i = t$ for $i = 1, \ldots, n$ for simplicity and drop t in the following notation (see section 6).

Each $p(\boldsymbol{x})$ is given by 2^n probabilities,

$$p_{i_1 \cdots i_n} = \text{Prob}\{X_1 = i_1, \ldots, X_n = i_n\}, i_k = 0, 1,$$
$$\text{subject to} \sum_{i_1, \ldots, i_n} p_{i_1 \cdots i_n} = 1,$$

and hence, the set of all the probability distributions $\{p(\boldsymbol{x})\}$ forms a $(2^n - 1)$–dimensional manifold $\boldsymbol{S}_n$.

One coordinate system of $\boldsymbol{S}_n$ is given by the expectation parameters,

$$\eta_i = E[x_i] = \text{Prob}\{x_i = 1\}, \quad i = 1, \ldots, n$$
$$\eta_{ij} = E[x_i x_j] = \text{Prob}\{x_i = x_j = 1\}, \quad i < j$$
$$\eta_{12 \cdots n} = E[x_i, \ldots, x_n] = \text{Prob}\{x_1 = x_2 = \cdots = x_n = 1\},$$

which have 2^n–1 components. This coordinate system is called η-coordinates and, as in a more general term, defines m-flat structure in $\boldsymbol{S}_n$ (see section 5).

On the other hand, $p(\boldsymbol{x})$ can be exactly expanded by

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

where the indices of θ_{ijk}, etc. satisfy $i < j < k$, etc and ψ is a normalization term, corresponding to $-\log p(x_1 = x_2 = \cdots = x_n = 0)$. All $\theta_{ijk}, \ldots,$ together have $2^n - 1$ components and form another coordinate system, called θ-coordinates, corresponding to the e-flat structure in $\boldsymbol{S}_n$ (see section 5).

Findings in information geometry assure us that e-flat and m-flat manifolds are dually flat. The η-coordinates and θ-coordinates are dually orthogonal coordinates. The properties of the dual orthogonal coordinates remarkably simplify some apparently complicated issues. For example, the generalized Pythagoras theorem gives a decomposition of the Kullback-Leibler divergence by which we can inspect different contributions in the discrepancy of two probability distributions or contributions of different order interactions in neural firing. This is a global property of the dual orthogonal coordinates in the probability space. As a local property, the dual orthogonal coordinates give a simple form of the Fisher information metric, which is useful, for example, in hypothesis testing. This study exploits these properties. In the next section, we start with the case of two neurons.

3 Pairwise Interaction, Mutual Information, and Orthogonal Decomposition

3.1 Orthogonal Coordinates. Let us begin with two binary random variables X_1 and X_2 whose joint probability $p(\boldsymbol{x})$, $\boldsymbol{x} = (x_1, x_2)$, is given by

$$p_{ij} = \text{Prob}\left\{x_1 = i; x_2 = j\right\} > 0, \quad i, j = 0, 1.$$

Among four probabilities, $\{p_{00}, p_{01}, p_{10}, p_{11}\}$, only three are free, because of the constraint $p_{00}+p_{01}+p_{10}+p_{11} = 1$. Thus, the set of all such distributions of $\boldsymbol{x}$ forms a three-dimensional manifold $\boldsymbol{S}_2$, where the suffix 2 refers to the number of random variables in $\boldsymbol{x}$. Any three of p_{ij} can be used as a coordinate system of $\boldsymbol{S}_2$, which we call P-coordinates for later convenience. In the context of neural firing, random variables X_1 and X_2 stand for two neurons: neuron 1 and neuron 2. $X_i = 1$ and $X_i = 0$ indicate whether neuron i ($i = 1, 2$) has a spike in a short time bin.

A distribution $p(\boldsymbol{x})$ can be decomposed into marginal and (pairwise) correlational components. The two quantities,

$$\eta_i = \text{Prob}\{x_i = 1\} = E[x_i], \quad i = 1, 2,$$

specify the marginal distributions of x_i, where E denotes the expectation. Obviously, we have $\eta_1 = p_{10} + p_{11}$, $\eta_2 = p_{01} + p_{11}$. Let us put

$$\eta_{12} = E[x_1 x_2] = p_{12}.$$

The three quantities,

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_{12}), \tag{3.1}$$

form another coordinate system of $\boldsymbol{S}_2$, called the η-coordinates. They are the coordinates of the expectation parameters in an exponential probability family in general (Cox & Hinkley, 1974; Barndorff-Nielsen, 1978; Lehmann, 1983). In the context of neural data, η_1 and η_2 are the mean firing rates of neurons 1 and 2, respectively, whereas η_{12} is the mean rate of their coincident firing.

The covariance,

$$\mathrm{Cov}\left[X_1, X_2\right] = E\left[\left(x_1 - \eta_1\right)\left(x_2 - \eta_2\right)\right] = \eta_{12} - \eta_1\eta_2,$$

may also be considered as a quantity representing the degree of correlation of X_1 and X_2. Therefore, $(\eta_1, \eta_2, \mathrm{Cov}[X_1, X_2])$ can be another coordinate system. The term $\mathrm{Cov}[X_1, X_2]$ becomes zero when the probability distribution is independent, because we have $\eta_{12} = \eta_1\eta_2$ in that case.

There are many candidates to specify the correlation component. The correlation coefficient

$$\rho = \frac{\eta_{12} - \eta_1\eta_2}{\sqrt{\eta_1\left(1 - \eta_1\right)\eta_2\left(1 - \eta_2\right)}}$$

is also such a quantity. The triplet (η_1, η_2, ρ) then forms another coordinate system of $\boldsymbol{S}_2$. The correlation coefficient is used to show the pairwise correlation of two neurons in N-JPSTH (Aertsen et al., 1989).

Which quantity is convenient for representing the pairwise correlational component? It is desirable to define the degree of pairwise interaction independent of the marginals η_1 and η_2. To this end, we use the orthogonal coordinates (η_1, η_2, θ) such that the coordinate curve of θ is always orthogonal to those of η_1 and η_2. This characteristic is particularly desirable in the context of neural data, as shown later.

Once such a θ is defined, we have a subset $\boldsymbol{E}(\theta)$ for each θ, a family of distributions having the same θ value (see Figure 1A). The $\boldsymbol{E}(\theta)$ is a two-dimensional submanifold on which (η_1, η_2) can vary freely but θ is fixed. We put the origin $\theta = 0$ when there is no correlation (i.e., $\eta_{12} = \eta_1\eta_2$) for convenience (see below), and then $\boldsymbol{E}(0)$ is the set of all the independent distributions. Similarly, we consider the set of all the probability distributions whose marginals are common, specified by (η_1, η_2), but only θ is free. This is denoted by $\boldsymbol{M}(\eta_1, \eta_2)$, forming a one-dimensional submanifold in $\boldsymbol{S}_2$. The tangential direction of $\boldsymbol{M}(\eta_1, \eta_2)$ represents the direction in which only the pure correlation changes, while the tangential directions of $\boldsymbol{E}(\theta)$ span the directions in which only η_1 and η_2 change but θ is fixed. We now require that $\boldsymbol{E}(\theta)$ and $\boldsymbol{M}(\eta_1, \eta_2)$ be orthogonal at any points, that is, the directions of changes in the correlation and marginals are to be mutually orthogonal.

The orthogonality of two directions in $\boldsymbol{S}_2$ is defined by using the Riemannian metric due to the Fisher information matrix (Rao, 1945; Barndorff-Nielsen, 1978; Amari, 1982; Nagaoka & Amari, 1982; Amari & Han, 1989; Amari & Nagaoka, 2000). Here, we define the orthogonality directly. Let us specify the probability distributions by $p(\boldsymbol{x}; \eta_1, \eta_2, \theta)$. The directions of

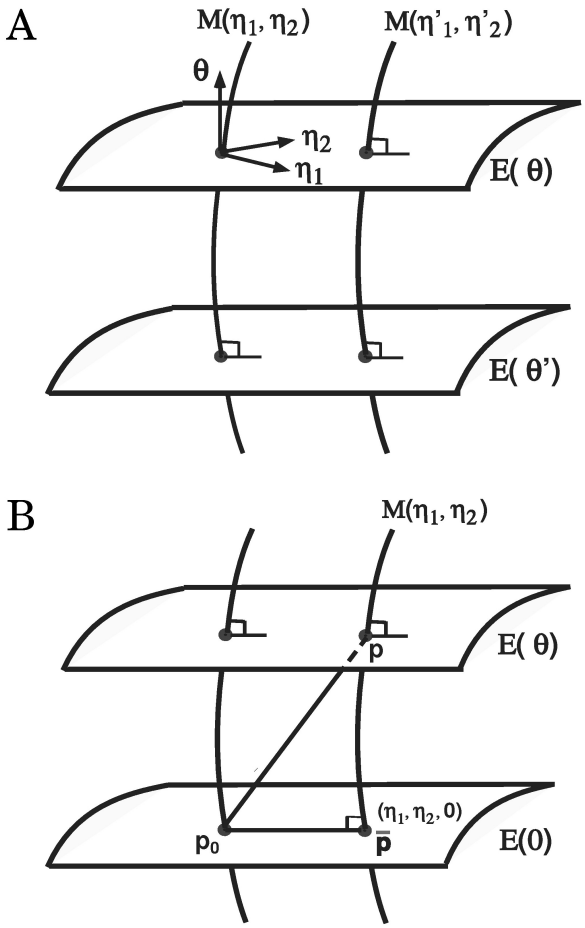


Figure 1: (A) Schematic diagram of $\boldsymbol{E}(\theta)$ and $\boldsymbol{M}(\eta_1, \eta_2)$. (B) A simple example of the generalized Pythagoras decomposition.

small changes in the coordinates η_i and θ are represented, respectively, by the score functions

$$\frac{\partial}{\partial \eta_i} l(\boldsymbol{x}; \eta_1, \eta_2, \theta), \quad (i = 1, 2) \quad \frac{\partial}{\partial \theta} l(\boldsymbol{x}; \eta_1, \eta_2, \theta),$$

where $l(\boldsymbol{x}; \eta_1, \eta_2, \theta) = \log p(\boldsymbol{x}; \eta_1, \eta_2, \theta)$.

They are random variables, denoting how the log probability changes by small changes in the parameters in the respective directions. These directions are said to be orthogonal when the corresponding random variables

are uncorrelated,

$$E\left[\frac{\partial}{\partial\theta}l\left(\boldsymbol{x};\eta_1,\eta_2,\theta\right)\frac{\partial}{\partial\eta_i}l\left(\boldsymbol{x};\eta_1,\eta_2,\theta\right)\right]=0, \tag{3.2}$$

where E denotes the expectation with respect to $p(\boldsymbol{x};\eta_1,\eta_2,\theta)$. This implies that the cross components of θ and η_i in the Fisher information matrix vanish. When the coordinate θ is defined to be orthogonal to the coordinates η_1 and η_2 of marginals, we say that θ represents the pure correlation independent of the marginals. Such θ is given by the following theorem.

Theorem 1. *The coordinate*

$$\theta=\log\frac{p_{11}p_{00}}{p_{01}p_{10}} \tag{3.3}$$

is orthogonal to the marginals η_1 and η_2.

The proof can be shown by direct calculations, which is omitted here. A more general result is shown later. We have another interpretation of θ. Let us expand $\log p(\boldsymbol{x})$ in the polynomial of $\boldsymbol{x}$,

$$\log p(\boldsymbol{x})=\sum_{i=1}^{2}\theta_i x_i+\theta_{12}x_1x_2-\psi. \tag{3.4}$$

Since x_i takes on the binary values $0,1$, this is an exact expansion. The coefficient θ_{12} is given by equation 3.3, while

$$\theta_1=\log\frac{p_{10}}{p_{00}},\quad \theta_2=\log\frac{p_{01}}{p_{00}},\quad \psi=-\log p_{00}. \tag{3.5}$$

We remark here that the above θ_{12} is well known, having frequently been used in the additive decomposition of log probabilities. It is 0 when and only when X_1 and X_2 are independent. The triple

$$\boldsymbol{\theta}=(\theta_1,\theta_2,\theta_{12})$$

forms another coordinate system of $\boldsymbol{S}_2$, called the θ-coordinates. They are the coordinates of the natural parameters in the exponential probability family in general (Cox & Hinkley, 1974; Barndorff-Nielsen, 1978; Lehmann, 1983). Furthermore, the triple

$$\boldsymbol{\zeta}\equiv(\eta_1,\eta_2,\theta_{12})$$

forms an orthogonal coordinate system of $\boldsymbol{S}_2$, called the mixed coordinates (Amari, 1985; Amari & Nagaoka, 2000).

3.2 KL-Divergence, Projections, and Pythagoras Relation. The Kullback-Leibler (KL) divergence between two probabilities $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is defined by

$$D\left[p:q\right]=\sum_{\boldsymbol{x}}p(\boldsymbol{x})\log\frac{p(\boldsymbol{x})}{q(\boldsymbol{x})}. \tag{3.6}$$

The KL divergence provides a quasi-distance between two probability distributions, $D[p:q]\geq 0$ with equality if and only if $p(\boldsymbol{x})=q(\boldsymbol{x})$, whereas the symmetrical relationship does not generally hold, that is, $D[p:q]\neq D[q:p]$.

Let $\bar{p}(\boldsymbol{x})$ be the independent distribution that is closest to a distribution $p(\boldsymbol{x})$,

$$\bar{p}(\boldsymbol{x})=\mathrm{argmin}_{q\in\boldsymbol{E}(0)}D\left[p:q\right],$$

where $\boldsymbol{E}(0)$ is the set of all the independent distributions. We call $\bar{p}(\boldsymbol{x})=\Pi_i p_i(x_i)$ the m-projection of p to $\boldsymbol{E}(0)$ (see Figure 1B). Let the mixed coordinates of p be (η_1,η_2,θ). The coordinates of $\bar{p}$ are given by $(\eta_1,\eta_2,0)$ because of the orthogonality, so that

$$\bar{p}(\boldsymbol{x})=\Pi_i p_i(x_i;\eta_i)=p_1(x_1;\eta_1)p_2(x_2;\eta_2),$$

where $p_i(x_i;\eta_i)$ is the marginal distribution of p. Interestingly, the minimized divergence is given by the mutual information

$$D\left[p:\bar{p}\right]=I(X_1;X_2)=\sum p(x_1,x_2)\log\frac{p(x_1,x_2)}{p_1(x_1)p_2(x_2)}.$$

We have another characterization of $\bar{p}$. Let p_0 be the uniform distribution whose mixed coordinates are $(0.5,0.5,0)$. Let $\boldsymbol{M}(\eta_1,\eta_2)$ be the subspace that includes p. Then,

$$\bar{p}=\mathrm{argmin}_{q\in\boldsymbol{M}(\eta_1,\eta_2)}D\left[q:p_0\right].$$

Such $\bar{p}$ is called the e-projection of p_0 to $\boldsymbol{M}(\eta_1,\eta_2)$, and it belongs to $\boldsymbol{E}(0)$. Since we easily have $D\left[q:p_0\right]=-H\left[q\right]+H_0$, where $H\left[q\right]$ is the entropy of q and $H_0=2\log 2$ is a constant, $\bar{p}$ has the maximal entropy among those belonging to $\boldsymbol{M}(\eta_1,\eta_2)$. This fact is called the maximum entropy principle (Jaynes, 1982).

It is well known that we have the decomposition

$$D\left[p:p_0\right]=D\left[p:\bar{p}\right]+D\left[\bar{p}:p_0\right].$$

Now let us generalize the above observation and let $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ be two probability distributions whose mixed coordinates are $\boldsymbol{\zeta}^p=(\eta_1^p,\eta_2^p,\theta_3^p)$ and $\boldsymbol{\zeta}^q=(\eta_1^q,\eta_2^q,\theta_3^q)$, respectively. Let $r^*(\boldsymbol{x})$ be the m-projection of $p(\boldsymbol{x})$ to $\boldsymbol{E}(\theta^q)$ and $r^{**}(\boldsymbol{x})$ be the e-projection of $p(\boldsymbol{x})$ to $\boldsymbol{M}(\eta_1^q,\eta_2^q)$:

$$r^*(\boldsymbol{x})=\mathrm{argmin}_{r\in\boldsymbol{E}(\theta^q)}D\left[p:r\right],\quad r^{**}(\boldsymbol{x})=\mathrm{argmin}_{r\in\boldsymbol{M}(\eta_1^q,\eta_2^q)}D\left[r:p\right].$$

The mixed coordinates of r^* and r^{**} are explicitly given by $(\eta_1^p, \eta_2^p, \theta_3^q)$ and $(\eta_1^q, \eta_2^q, \theta_3^p)$, respectively. Hence, the following Pythagoras relation holds (see Figure 1B).

Theorem 2.

$$D\left[p : q\right] = D\left[p : r^*\right] + D\left[r^* : q\right] \tag{3.7}$$

$$D\left[q : p\right] = D\left[q : r^{**}\right] + D\left[r^{**} : p\right]. \tag{3.8}$$

Theorem 2 shows that the divergence $D\left[p : q\right]$ from p to q is decomposed into two terms, $D\left[p : r^*\right]$ and $D\left[r^* : q\right]$, where the former represents the degree of difference in their correlation and the latter the difference in their marginals.

3.3 Local Orthogonality and Fisher Information. For any parameterization $p(\boldsymbol{x}; \boldsymbol{\xi})$, the Fisher information matrix $G = (g_{ij})$ in terms of the coordinates $\boldsymbol{\xi}$ is given by

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

This $G(\boldsymbol{\xi})$ plays the role of a Riemannian metric tensor.

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

$$ds^2 = \sum_{i,j\in(1,2,3)} g_{ij}(\boldsymbol{\xi}) d\xi_i d\xi_j.$$

It is known that this is approximately twice the KL divergence,

$$ds^2 \approx 2D\left[p(\boldsymbol{x}; \boldsymbol{\xi}) : p(\boldsymbol{x}; \boldsymbol{\xi} + d\boldsymbol{\xi})\right].$$

When we use the mixed coordinates $\boldsymbol{\zeta}$, the Fisher information is of the form

$$G_\zeta = (g_{ij}^\zeta) = \begin{bmatrix} g_{11}^\zeta & g_{12}^\zeta & 0 \\ g_{12}^\zeta & g_{22}^\zeta & 0 \\ 0 & 0 & g_{33}^\zeta \end{bmatrix},$$

as is seen from equation 3.2. This is the local property induced by the orthogonality of θ and η_i. In this case, by putting

$$ds_1^2 = g_{33}^\zeta (d\xi_3)^2, \quad ds_2^2 = \sum_{i,j\in(1,2)} g_{ij}^\zeta d\eta_i d\eta_j,$$

we have the orthogonal decomposition

$$ds^2 = ds_1^2 + ds_2^2, \tag{3.9}$$

corresponding to equation 3.2.

We show the merits of the orthogonal coordinates for statistical inference. Let us estimate the parameter $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and θ from N observed data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. The maximum likelihood estimator is asymptotically unbiased and efficient, where the covariance of the estimation error, $\Delta\boldsymbol{\eta}$ and $\Delta\theta$, is given asymptotically by

$$\mathrm{Cov}\begin{bmatrix}\Delta\boldsymbol{\eta}\\ \Delta\theta\end{bmatrix} = \frac{1}{N}G_{\zeta}^{-1}.$$

Since the cross terms of G or G^{-1} vanish for the orthogonal coordinates, we have

$$\mathrm{Cov}\,[\Delta\boldsymbol{\eta}, \Delta\theta] = 0, \tag{3.10}$$

implying that the estimation error $\Delta\boldsymbol{\eta}$ of marginals and that of interaction are mutually independent. Such a property does not hold for other nonorthogonal parameterizations such as the correlation coefficient ρ and the covariance. This property greatly simplifies procedures of hypothesis testing, as shown below.

3.4 Hypothesis Testing. Let us consider the estimation of θ and $\boldsymbol{\eta}$ more directly. A natural estimate for the η-coordinates is

$$\hat{\eta}_i = \frac{1}{N}\#\{x_i = 1\} \quad (i = 1, 2), \quad \hat{\eta}_{12} = \frac{1}{N}\#\{x_1 x_2 = 1\}. \tag{3.11}$$

This is the maximum likelihood estimator. The estimator $\hat{\theta}$ is obtained by the coordinate transformation from η- to θ-coordinates,

$$\hat{\theta} = \log\frac{\hat{\eta}_{12}(1 - \hat{\eta}_1 - \hat{\eta}_2 + \hat{\eta}_{12})}{(\hat{\eta}_1 - \hat{\eta}_{12})(\hat{\eta}_2 - \hat{\eta}_{12})}.$$

Notably, the estimation of θ can be performed independently from the estimator of $\boldsymbol{\eta}$ in the sense of equation 3.10. This brings a simple procedure of hypothesis testing concerning the null hypothesis,

$$H_0 : \theta = \theta_0,$$

against

$$H_1 : \theta \neq \theta_0.$$

In previous studies, under different frameworks (e.g., using N-JPSTH), the null hypothesis of independent firing is often examined. This corresponds to the null hypothesis of $\theta_0 = 0$ in the current framework.

Let the log-likelihood of the models $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$, respectively, be

$$l_0 = \max_{\boldsymbol{\eta}} \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\eta}, \theta_0), \quad l_1 = \max_{\boldsymbol{\eta}, \theta} \log p(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\eta}, \theta),$$

where N is the number of observations.

The likelihood ratio test uses the test statistics

$$\lambda = 2 \log \frac{l_0}{l_1}, \tag{3.12}$$

which is subject to the χ^2 distribution. With the orthogonal coordinates, the likelihood maximization with respect to $\boldsymbol{\eta} = (\eta_1, \eta_2)$ and θ can be performed independently, so that we have

$$l_0 = \log p(\bar{\boldsymbol{x}}; \hat{\boldsymbol{\eta}}, \theta_0), \quad l_1 = \log p(\bar{\boldsymbol{x}}; \hat{\boldsymbol{\eta}}, \hat{\theta}),$$

where $\hat{\boldsymbol{\eta}}$ denotes the same marginals in both models. If nonorthogonal parameterization is used, this property does not hold. A similar situation holds in the case of testing $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ against $\boldsymbol{\eta} \neq \boldsymbol{\eta}_0$ for unknown θ.

Now let us calculate the test statistics λ in more detail. Under the hypothesis H_0, λ is approximated for a large N as

$$\begin{aligned}
\lambda &= 2 \sum_{i=1}^{N} \log \frac{p(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}, \theta_0)}{p(\boldsymbol{x}_i; \hat{\boldsymbol{\eta}}, \hat{\theta})} \\
&\approx 2N\tilde{E}\left[\log \frac{p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \theta_0)}{p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \hat{\theta})} \right] \\
&\approx 2ND\left[p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \theta_0) : p(\boldsymbol{x}; \hat{\boldsymbol{\eta}}, \hat{\theta}) \right] \\
&\approx N g_{33}^{\zeta} (\hat{\theta} - \theta_0)^2,
\end{aligned} \tag{3.13}$$

where $\tilde{E}$ is the expectation over the empirical distribution and the approximation in the third line comes from our assumption of the null hypothesis H_0. g_{33}^{ζ} is the Fisher information of the mixed coordinates $\boldsymbol{\zeta}$ in the θ direction at $\boldsymbol{\zeta}^0 = (\hat{\boldsymbol{\eta}}; \theta_0)$, which is easily calculated as

$$g_{33}^{\zeta} = g_{33}(\boldsymbol{\zeta}^0) = \frac{\hat{\eta}_3(\hat{\eta}_1 - \hat{\eta}_3)(\hat{\eta}_2 - \hat{\eta}_3)(\hat{\eta}_1 + \hat{\eta}_2 - \hat{\eta}_3 - 1)}{\hat{\eta}_1 \hat{\eta}_2 (\hat{\eta}_1 + \hat{\eta}_2 - 1 - 2\hat{\eta}_3) + \hat{\eta}_3^2}.$$

Asymptotically, we have $\sqrt{N}\sqrt{g_{33}^{\zeta}}(\hat{\theta}_3 - \theta_3) \sim \mathcal{N}(0, 1)$ and hence,

$$\lambda \sim \chi^2(1),$$

where m in $\chi^2(m)$ indicates the degrees of freedom m in the χ^2 distribution, that is, in our case, the degree of freedom is 1.

We must note that the above approach is valid regardless of $\theta_3 = 0$ or $\neq 0$. In contrast, the decomposition as shown in equation 3.9 cannot exist, for example, for the coordinate system (η_1, η_2, ρ), where ρ is the correlation coefficient. The plane $\theta_3 = 0$, or $\boldsymbol{E}(0)$, coincides with the plane $\rho = 0$, which is $\eta_3 = \eta_1\eta_2$. However, $\boldsymbol{E}(c)$ ($c = const \neq 0$) cannot be equal to any plane defined by $\rho = c'$ where $c' = const$. Only in the case of $\rho = 0$ it is possible to formulate testing for ρ similarly to the above discussion, which is testing against the hypothesis of independent firing.

3.5 Application to Firing of Two Neurons. Here, we discuss the application of the theoretical results already noted to the firing of two neurons and relate different choices of the null hypothesis with corresponding hypothesis testings. Given N trials of experiments, the probability distribution of X in a time bin $[t, t+\delta t]$ can be estimated, denoted by $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}) = p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}(t, t+\Delta t))$, where $\hat{\boldsymbol{\xi}}$ can be any coordinate system. If stationarity is assumed in a certain time interval, we obtain the probability distribution in the interval by averaging the estimated probabilities of many bins of the interval.

The maximum likelihood estimate (MLE) of the P coordinates is given by

$$\hat{p}_{ij} = \frac{N_{ij}}{N},$$

where N_{ij} $(i, j = 0, 1)$ indicates the number of trials in which the event $(X_1 = i, X_2 = j)$ occurs. The MLE is retained by any coordinate transformation. Any coordinate transformation is easy in the case of two neurons, so we freely change the coordinate systems in this section.

Let us denote our estimated probability distribution by the mixed coordinates $\hat{\boldsymbol{\zeta}}$. We also denote by $\boldsymbol{\zeta}^0$ the probability distribution according to our null hypothesis. Then, we have

$$D\left[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}\right] = D\left[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}'\right] + D\left[\hat{\boldsymbol{\zeta}}' : \hat{\boldsymbol{\zeta}}\right] = D_1 + D_2, \tag{3.14}$$

where $D_1 = D[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}']$, $D_2 = D[\hat{\boldsymbol{\zeta}}' : \hat{\boldsymbol{\zeta}}]$, and $\hat{\boldsymbol{\zeta}}' = (\zeta_1^0, \zeta_2^0, \hat{\zeta}_3)$. In equation 3.14, we use abbreviation such that $D[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}'] = D[p(\boldsymbol{x}; \boldsymbol{\zeta}^0) : p(\boldsymbol{x}; \hat{\boldsymbol{\zeta}}')]$.

Here, D_1 and D_2 are the quantities representing the discrepancies of $p(\hat{\boldsymbol{\zeta}})$ from $p(\boldsymbol{\zeta}^0)$ with respect to the coincident firing and the marginals, respectively. We have

$$\lambda_1 = 2ND_1 \approx Ng_{33}(\boldsymbol{\zeta}^0)(\zeta_3^0 - \hat{\zeta}_3')^2 \sim \chi^2(1)$$

$$\lambda_2 = 2ND_2 \approx N\sum_{i,j=1}^{2} g_{ij}(\boldsymbol{\zeta}^0)(\zeta_i^0 - \hat{\zeta}_j')^2 \sim \chi^2(2).$$

Here, λ_1 is to test whether the estimated coincident firing significantly differs from that of the null hypothesis, while λ_2 is to test whether the estimated marginals differ significantly from the hypothesized marginals.

In particular, a test of whether the estimated coincident firing $\hat{\zeta}_3$ is significantly different from zero is given by $\boldsymbol{\zeta}^0 = (\hat{\zeta}_1, \hat{\zeta}_2, 0)$. This $p(\boldsymbol{x}; \boldsymbol{\zeta}^0)$ is the probability distribution that gives the same marginals as those of $p(\boldsymbol{x}; \hat{\boldsymbol{\zeta}})$ but with independent firings. In this case, $\lambda_1 = 2nD_1 = 2nD[\hat{\boldsymbol{\zeta}} : \boldsymbol{\zeta}^0]$ gives a test statistic against $\hat{\theta}_3 = 0$, while $D_2 = 0$.

Let us consider another typical situation, where we need to compare two estimated probability distributions. This case is very important but somewhat ignored in the testing of coincident firings. Many previous studies often assumed independent firing as the null hypothesis. However, for example, to say a single neuron firing as task related, for example, in a memory-guided saccade task (Hikosaka & Wurtz, 1983), the existence of firing in task period alone does not guarantee that the firing is task related. It is normal to examine the firing in the task period against that in the control period. The firing in the control period serves as the resting level activity or as the null hypothesis. We hence propose that a procedure for testing coincident firing should be performed in a similar manner: we should test if two neurons have any significant pairwise interaction in one period in comparison to the other (control) period.

Investigation of coincident firing in the task period against the null hypothesis of independent firing may lead to a wrong interpretation of its significance when there is already a weak correlation in the control period (see the examples in section 6). The similar arguments can be applied to different tasks. One example would be a rat's maze task. A rat is in the left room in one period, while in the other period, it is in the right room. We may want to test if coincident firing of the two neurons, say, in the hippocampus, is significantly larger or smaller in one room than in the other room. The null hypothesis of independent firing is not plausible in this case.

Let us denote the estimated probability distribution in two periods by $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}^1)$ and $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}^2)$. Using the mixed coordinates, by theorem 2, we have

$$D\left[\hat{\boldsymbol{\zeta}}^1 : \hat{\boldsymbol{\zeta}}^2\right] = D\left[\hat{\boldsymbol{\zeta}}^1 : \hat{\boldsymbol{\zeta}}^3\right] + D\left[\hat{\boldsymbol{\zeta}}^3 : \hat{\boldsymbol{\zeta}}^2\right],$$

where $\hat{\boldsymbol{\zeta}}^3 = (\hat{\zeta}_1^1, \hat{\zeta}_2^1, \hat{\zeta}_3^2) = (\hat{\eta}_1^1, \hat{\eta}_2^1, \hat{\theta}_3^2)$.

Here, $\hat{\boldsymbol{\zeta}}^1$ is an estimated probability distribution. If we can guarantee that $\hat{\boldsymbol{\zeta}}^1$ is a true underlying distribution, denoted by $\boldsymbol{\zeta}^1$, we can have

$$\lambda = 2ND\left[\boldsymbol{\zeta}^1 : \hat{\boldsymbol{\zeta}}^3\right] \approx Ng_{33}(\boldsymbol{\zeta}^1)(\hat{\theta}_3^2 - \theta_3^1)^2 \sim \chi^2(1). \tag{3.15}$$

This χ^2 test is, precisely speaking, to examine if $\hat{\theta}_3^2$ is significantly different from θ_3^1 when $\hat{\boldsymbol{\zeta}}^1$ is a true distribution.

In general, when $\hat{\zeta}^1$ is an estimated distribution, we should test whether $\hat{\theta}_3^1$ and $\hat{\theta}_3^2$ are from the same interaction component, which we denote by θ_3. In this case, the MLEs, denoted by $\hat{\zeta}^{10}$ and $\hat{\zeta}^{20}$, are given by

$$(\hat{\zeta}^{10}, \hat{\zeta}^{20}) = \text{argmax} \sum_j^N \log p(X^j : \zeta^1) p(X^j : \zeta^2)$$
$$\text{subject to } \zeta_3^1 = \zeta_3^2 = \theta_3.$$

Then our likelihood ratio test against this null hypothesis yields

$$\lambda' = 2ND\left[\hat{\zeta}^{10} : \hat{\zeta}^1\right] + 2ND\left[\hat{\zeta}^{20} : \hat{\zeta}^2\right]$$
$$\approx Ng_{33}(\hat{\zeta}^{10})(\hat{\theta}_3^1 - \hat{\theta}_3^{10})^2 + Ng_{33}(\hat{\zeta}^{20})(\hat{\theta}_3^2 - \hat{\theta}_3^{20})^2, \tag{3.16}$$

where $\hat{\theta}_3 = \hat{\theta}_3^{10} = \hat{\theta}_3^{20}$. In equation 3.15, we can convert λ into a χ^2 test, because g_{33} is the true value by our assumption. In equation 3.16, however, rigorously speaking, we cannot convert this λ into a χ^2 test because both g_{33} are estimates, determined at each estimated point $\hat{\zeta}$, that is, depending on $\hat{\theta}_3^{10}$ and $\hat{\theta}_3^{20}$, respectively. This issue is analogous to the famous Fisher-Beherens problem in the context of the *t*-test (Stuart, Ord, & Arnold, 1999). Yet since all of the terms in equation 3.16 asymptotically converge to their true values, we suggest using

$$\lambda' \approx Ng_{33}(\hat{\zeta}^{10})(\hat{\theta}_3^1 - \hat{\theta}_3^{10})^2 + Ng_{33}(\hat{\zeta}^{20})(\hat{\theta}_3^2 - \hat{\theta}_3^{20})^2 \sim \chi^2(2).$$

This $\chi^2(2)$ formulation gives a more appropriate test under the null hypothesis against the average activity in control period. At the same time, to compare significant events between the two null hypotheses—against independent firing and against the average activity in the control period—we still suggest using the $\chi^2(1)$ formulation for the latter hypothesis.

3.6 Relationship Between Neural Firing and Behavior. The orthogonality between θ and η parameters played a fundamental role in the above results so that pairwise coincident firing, characterized by θ_3, can be examined by a simple hypothesis testing procedure. In the analysis of neural data, it is also important to investigate whether any coincident firing has any behavioral significance. For this purpose, we use the mutual information to relate neural firing with behavioral events. The above orthogonality can again play an important role.

Let us denote by Y a discrete random variable representing behavioral choices, for example, making saccade right or left, or presented stimuli, for

example, red dots, blue rectangles, or green triangles. The mutual information between $X = (X_1, X_2)$ and Y is defined by

$$I(X, Y) = E_{p(X,Y)}\left[\log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)}\right],$$

which is equivalent to

$$I(X, Y) = E_{p(Y)}\left[D\left[p(X \mid y) : p(X)\right]\right] = E_{p(X)}\left[D\left[p(Y \mid x) : p(Y)\right]\right].$$

We can apply the Pythagoras decomposition to the above equation. We use the mixed coordinates for $p(X \mid y)$ and $p(X)$, denoted by $\boldsymbol{\zeta}(X \mid y)$ and $\boldsymbol{\zeta}(X)$, respectively. Then we have

$$\begin{aligned} D\left[p(X \mid y) : p(X)\right] &= D\left[\boldsymbol{\zeta}(X \mid y) : \boldsymbol{\zeta}(X)\right] \\ &= D\left[\boldsymbol{\zeta}(X \mid y) : \boldsymbol{\zeta}'\right] + D\left[\boldsymbol{\zeta}' : \boldsymbol{\zeta}(X)\right], \end{aligned}$$

where

$$\boldsymbol{\zeta}' = \boldsymbol{\zeta}'(X, y) = (\zeta_1(X \mid y), \zeta_2(X \mid y), \zeta_3(X)) = (\eta_1(X \mid y), \eta_2(X \mid y), \theta_3(X)).$$

Thus, $\boldsymbol{\zeta}'$ has the first two components (i.e., η_1, η_2) as the same as those of $\boldsymbol{\zeta}(X \mid y)$ and the third term (i.e., θ_3) as the same as that of $\boldsymbol{\zeta}(X)$. Using this relationship, the mutual information between X and Y is decomposed.

Theorem 3.

$$I(X, Y) = I_1(X, Y) + I_2(X, Y), \tag{3.17}$$

where $I_1(X, Y)$, $I_2(X, Y)$ are given by

$$\begin{aligned} I_1(X, Y) &= E_{p(Y)}\left[D\left[\boldsymbol{\zeta}(X \mid y) : \boldsymbol{\zeta}'(X, y)\right]\right], \\ I_2(X, Y) &= E_{p(Y)}\left[D\left[\boldsymbol{\zeta}'(X, y) : \boldsymbol{\zeta}(X)\right]\right]. \end{aligned}$$

A similar result holds with respect to conditional distribution $p(Y \mid X)$. The above decomposition states that the mutual information $I(X, Y)$ is the sum of the two terms: $I_1(X, Y)$ is the mutual information by modulation of the correlation components of X, while $I_2(X, Y)$ is the mutual information by modulation of the marginal means of X. This observation helps us investigate the behavioral significance for each modulation of the coincident firing and the mean firing rate.

4 Triple Interactions Among Three Variables

The previous section discussed pairwise interaction between two variables. Given more than two variables, we need to look into not only pairwise interaction but also higher-order interactions. It is useful to study triplewise interactions before stating the general case.

4.1 Orthogonal Coordinates and Pure Triple Interaction. Let us consider three binary random variables X_1, X_2, and X_3, and let $p(\boldsymbol{x}) > 0$, $\boldsymbol{x} = (x_1, x_2, x_3)$ be their joint probability distribution. We put $p_{ijk} = \text{Prob}\{x_1 = i, x_2 = j, x_3 = k\} > 0$, $i, j, k = 0, 1$. The set of all such distributions forms a seven-dimensional manifold $\boldsymbol{S}_3$, because $\sum p_{ijk} = 1$ among the eight p_{ijk}'s.

The marginal and pairwise marginal distributions of X_i are defined by

$$\eta_i = E[x_i] = \text{Prob}\{x_i = 1\} \quad (i = 1, 2, 3),$$
$$\eta_{ij} = E\left[x_i x_j\right] = \text{Prob}\left\{x_i = x_j = 1\right\}, \quad (i, j = 1, 2, 3).$$

The three quantities η_i, η_j, and η_{ij} together determine the joint marginal distribution of any two random variables X_i and X_j. Let us further put

$$\eta_{123} = E[x_1 x_2 x_3] = \text{Prob}\{x_i = x_j = x_k = 1\}.$$

All of these have seven degrees of freedom,

$$\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_7) = (\eta_1, \eta_2, \eta_3; \eta_{12}, \eta_{23}, \eta_{13}; \eta_{123}), \tag{4.1}$$

which specify any distribution $p(\boldsymbol{x})$ in $\boldsymbol{S}_3$. Hence, this $\boldsymbol{\eta}$ is a coordinate system of $\boldsymbol{S}_3$ called the m- or η-coordinates.

The pairwise correlation between any two of X_1, X_2, and X_3 is determined from the marginal distributions of X_i and X_j or η_i, η_j, and η_{ij}. However, even when all the pairwise correlations vanish, this does not imply that X_1, X_2, and X_3 are independent. Therefore, one should define intrinsic triplewise interaction independent of pairwise correlations. The coordinate η_{123} itself does not directly give the degree of pure triplewise interaction.

In order to define the degree of pure triplewise interaction, the orthogonality plays a fundamental role. Let us fix the three pairwise marginal distributions, specified by the six coordinates,

$$\boldsymbol{\eta}_2 = (\eta_1, \eta_2, \eta_3; \eta_{12}, \eta_{23}, \eta_{13}).$$

There are many distributions with the same $\boldsymbol{\eta}_2$. Let us consider the set $\boldsymbol{M}_2(\boldsymbol{\eta}_2)$ of all the distributions in which we have the same single and pairwise marginals $\boldsymbol{\eta}_2$, but η_{123} may take any value. This is a one-dimensional submanifold specified by $\boldsymbol{\eta}_2$. Let us introduce a coordinate θ in $\boldsymbol{M}_2(\boldsymbol{\eta}_2)$. $(\boldsymbol{\eta}_2, \theta)$ is a coordinate system of $\boldsymbol{S}_3$. When the coordinate θ is orthogonal to $\boldsymbol{\eta}_2$, that is, a change in the log-likelihood along θ is not correlated with that in any of the components of $\boldsymbol{\eta}_2$, we may say that θ represents the degree of pure triple interaction regardless of pairwise marginals $\boldsymbol{\eta}_2$ and require that θ has this property.

The tangent direction of $\boldsymbol{M}_2$, that is, the direction in which only θ changes but the second-order marginals $\boldsymbol{\eta}_2$ are fixed, represents a change in the pure triple interaction among X_1, X_2, and X_3. To show this geometrically, let us consider a family of submanifolds $\boldsymbol{E}_{2^*}(\theta)$ in which all the distributions have

the same θ but the single and pairwise marginals $\boldsymbol{\eta}_2$ are free. A $\boldsymbol{E}_{2^*}(\theta)$ is a six-dimensional submanifold transversal to $\boldsymbol{M}_2(\boldsymbol{\eta}_2)$. Tangent directions of $\boldsymbol{E}_{2^*}(\theta)$ represent changes in marginals $\boldsymbol{\eta}_2$, keeping θ fixed, and $\boldsymbol{E}_{2^*}(\theta)$ and $\boldsymbol{M}_2(\boldsymbol{\eta}_2)$ are orthogonal at any θ and $\boldsymbol{\eta}_2$.

In order to obtain such a θ, let us expand $\log p(\boldsymbol{x})$ in the polynomial of $\boldsymbol{x}$,

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum \theta_{ij} x_i x_j + \theta_{123} x_1 x_2 x_3 - \psi. \tag{4.2}$$

This is an exact formula, since x_i $(i = 1, 2, 3)$ is binary. One can check that the coefficient $\theta = \theta_{123}$ is given by

$$\theta_{123} = \log \frac{p_{111} p_{100} p_{010} p_{001}}{p_{110} p_{101} p_{011} p_{000}}. \tag{4.3}$$

The other coefficients are:

$$\theta_1 = \log \frac{p_{100}}{p_{000}}, \quad \theta_2 = \log \frac{p_{010}}{p_{000}}, \quad \theta_3 = \log \frac{p_{001}}{p_{000}}, \tag{4.4}$$

$$\theta_{12} = \log \frac{p_{110} p_{000}}{p_{100} p_{010}}, \quad \theta_{23} = \log \frac{p_{011} p_{000}}{p_{010} p_{001}}, \quad \theta_{13} = \log \frac{p_{101} p_{000}}{p_{100} p_{001}}, \tag{4.5}$$

$$\psi = -\log p_{000}. \tag{4.6}$$

Information geometry gives the following theorem.

Theorem 4. *The quantity θ_{123} represents the pure triplewise interaction in the sense that it is orthogonal to any changes in the single and pairwise marginals.*

We can prove this directly by calculating the derivatives of the log likelihood. Equation 4.2 shows that $\boldsymbol{S}_3$ is an exponential family with the canonical parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3; \theta_{12}, \theta_{23}, \theta_{13}; \theta_{123})$. The corresponding expectation parameters are $\boldsymbol{\eta} = (\boldsymbol{\eta}_2; \eta_{123})$, so that they are orthogonal. We can compose the mixed orthogonal coordinates, denoted by $\boldsymbol{\zeta}_2$, as

$$\boldsymbol{\zeta}_2 = (\boldsymbol{\eta}_2; \theta_{123}) = (\eta_1, \eta_2, \eta_3; \eta_{12}, \eta_{23}, \eta_{13}; \theta_{123}). \tag{4.7}$$

In this coordinate system, $\boldsymbol{\eta}_2$ and $\theta = \theta_{123}$ are orthogonal. Note that θ_{123} is not orthogonal to $\theta_{12}, \theta_{23}, \theta_{13}$. Hence, except when there is no triplewise interaction $(\theta_{123} = 0)$, the quantities θ_{12}, θ_{23}, and θ_{13} in equation 4.5 do not directly represent the degrees of pairwise correlations of the respective two random variables.

Notably, the submanifold $\boldsymbol{E}_{2^*}(0)$ consists of all the distributions having no triple interactions but pairwise interactions. The log probability $\log p(\boldsymbol{x})$ is quadratic and given by $\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum \theta_{ij} x_i x_j - \psi$. A stable distribution of a Boltzmann machine in neural networks belongs to this class,

because there are no triple interactions among neurons (Amari, Kurata, & Nagaoka, 1992). The submanifold $\boldsymbol{E}_{2^*}(0)$ is characterized by $\theta_{123} = 0$, or in terms of $\boldsymbol{\eta}$ (see equation 4.3) by

$$\eta_{123} = \frac{\begin{array}{c}(\eta_{12} - \eta_{123})(\eta_{13} - \eta_{123})(\eta_{23} - \eta_{123}) \\ \times(1 - \eta_1 - \eta_2 - \eta_3 + \eta_{12} + \eta_{23} + \eta_{13} - \eta_{123})\end{array}}{(\eta_1 - \eta_{12} - \eta_{13} + \eta_{123})(\eta_2 - \eta_{23} - \eta_{12} + \eta_{123})(\eta_3 - \eta_{13} - \eta_{23} + \eta_{123})}.$$

4.2 Another Orthogonal Coordinate System. In the above, we extracted the pure triple interaction by using the coordinate θ_{123}, such that $\boldsymbol{\eta}_2$ and θ_{123} are orthogonal. If we are interested in separating simple marginals from various kinds of interactions, we can use another decomposition. Let us summarize the three simple marginals in $\boldsymbol{\eta}_1 = (\eta_1, \eta_2, \eta_3)$ and then summarize all of the interaction terms in $\boldsymbol{\theta}_{1^*} = (\theta_{12}, \theta_{23}, \theta_{13}, \theta_{123})$. Here, $\boldsymbol{\theta}_{1^*}$ denotes the coordinates complementary to $\boldsymbol{\eta}_1$. Using this pair, we have another mixed coordinate system, denoted by $\boldsymbol{\zeta}_1$, as

$$\boldsymbol{\zeta}_1 = (\zeta_{11}, \zeta_{12}, \ldots, \zeta_{17}) = \left(\boldsymbol{\eta}_1, \boldsymbol{\theta}_{1^*}\right). \tag{4.8}$$

Here, $\boldsymbol{\eta}_1$ and $\boldsymbol{\theta}_{1^*}$ are orthogonal. Geometrically, let $\boldsymbol{M}_1(\boldsymbol{\eta}_1)$, specified by $\boldsymbol{\eta}_1 = (\eta_1, \eta_2, \eta_3)$, be the set of all the distributions having the same simple marginals $\boldsymbol{\eta}_1 = (\eta_1, \eta_2, \eta_3)$ but having any pairwise and triplewise correlations. $\boldsymbol{M}_1(\boldsymbol{\eta}_1)$ is a four-dimensional submanifold in which $\boldsymbol{\theta}_{1^*}$ takes arbitrary values. On the other hand, let $\boldsymbol{E}_{1^*}(\boldsymbol{\theta}_{1^*})$ be a three-dimensional submanifold in which all of the distributions have the same $\boldsymbol{\theta}_{1^*} = (\theta_{12}, \theta_{23}, \theta_{31}, \theta_{123})$ but different marginals $\boldsymbol{\eta}_1$. We have the following theorem.

Theorem 5. *The coordinates $\boldsymbol{\eta}_1$ and $\boldsymbol{\theta}_{1^*}$ are orthogonal, that is, $\boldsymbol{E}_{1^*}(\boldsymbol{\theta}_{1^*})$ is orthogonal to $\boldsymbol{M}_1(\boldsymbol{\eta}_1)$.*

Here, $\boldsymbol{\theta}_{1^*}$ represents degrees of pure correlations independent of marginals $\boldsymbol{\eta}_1$ and includes correlations resulting from the triplewise interaction in addition to the pairwise interactions. Because of the non-Euclidean character of $\boldsymbol{S}_3$ (Amari & Nagaoka, 2000; Amari, 2001), we cannot have a coordinate system, $(\eta_1, \eta_2, \eta_3; \theta'_{12}, \theta'_{23}, \theta'_{13}; \theta_{123})$, with $\{\eta_i\}$, $\{\theta'_{ij}\}$, and θ_{123} being mutually orthogonal. The submanifold $\boldsymbol{E}_{1^*}(0)$ has zero pairwise and triplewise correlations and, hence, consists entirely of independent distributions in which $\eta_{ij} = \eta_i\eta_j$ and $\eta_{123} = \eta_1\eta_2\eta_3$ hold. The function $\log p(\boldsymbol{x})$ is linear in $\boldsymbol{x}$ because $\boldsymbol{\theta}_{1^*} = 0$ (see equation 4.2).

4.3 Projections and Decompositions of Divergence. Using the above two mixed coordinates, we decompose a probability distribution in the following two ways. Let us consider two probability distributions, $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$, where any coordinate system $\boldsymbol{\xi}$ is denoted by $\boldsymbol{\xi}^p$ and $\boldsymbol{\xi}^q$, respectively.

First, let us consider the case where q is the independent uniform distribution. By using the mixed orthogonal coordinate system $\boldsymbol{\zeta}_2$, we now seek to extract a pure triplewise interaction θ_{123}. For q, we have

$$\theta^q_{123} = 0, \quad \eta^q_1 = \eta^q_2 = \eta^q_3 = \frac{1}{2}, \quad \eta^q_{12} = \eta^q_{23} = \eta^q_{13} = \frac{1}{4}.$$

Furthermore, we note that $q \in \boldsymbol{E}_{2^*}(0)$ and also $q \in \boldsymbol{E}_{1^*}(0)$.

Let us m-project p to $\boldsymbol{E}_{2^*}(0)$ by

$$\bar{p}(\boldsymbol{x}) = \underset{r \in \boldsymbol{E}_{2^*}(0)}{\arg\min} D\left[p(\boldsymbol{x}) : r(\boldsymbol{x})\right].$$

This $\bar{p}$ has the same pairwise marginals as p but does not include any triplewise interaction, and its mixed coordinates are given by $\boldsymbol{\zeta}_2^{\bar{p}} = (\boldsymbol{\eta}_2^p; \theta^q_{123}) = (\boldsymbol{\eta}_2^p; 0)$. The Pythagorean theorem gives us

$$D\left[p : q\right] = D\left[p : \bar{p}\right] + D\left[\bar{p} : q\right],$$

where $D[p : \bar{p}]$ represents the degree of pure triplewise interaction, while $D[\bar{p} : q]$ represents how p differs from q in simple marginals and pairwise correlations.

Let us next extract the pairwise interactions in $p(\boldsymbol{x})$ by using another mixed coordinate $\boldsymbol{\zeta}_1$. To this end, let us project p to $\boldsymbol{E}_{1^*}(0)$, which is composed of independent distributions,

$$\widetilde{p}(\boldsymbol{x}) = \underset{s \in \boldsymbol{E}_{1^*}(0)}{\arg\min} D\left[p(\boldsymbol{x}) : s(\boldsymbol{x})\right].$$

More explicitly, we have $\boldsymbol{\zeta}_1^{\tilde{p}} = (\boldsymbol{\eta}_1^{\tilde{p}}; \boldsymbol{\theta}^q_{1^*}) = (\boldsymbol{\eta}_1^p; \mathbf{0})$ and

$$D\left[p : \widetilde{p}\right] = D\left[p : \bar{p}\right] + D\left[\bar{p} : \widetilde{p}\right].$$

Here, $D[p : \bar{p}]$ summarizes the effect of all the pairwise and triplewise interactions, while $D[\bar{p} : q]$ represents the difference of the simple marginals from the uniformity.

By taking the two decompositions together, we have

$$D\left[p : q\right] = D\left[p : \bar{p}\right] + D\left[\bar{p} : \widetilde{p}\right] + D\left[\widetilde{p} : q\right]. \tag{4.9}$$

Here $D[p : \bar{p}]$ represents the degree of pure triplewise interaction in the probability distribution p, $D[\bar{p} : \widetilde{p}]$ of pairwise interactions, and $D[\widetilde{p} : q]$ of the nonuniformity of firing rate.

Let us generalize equation 4.9 by dropping our assumption on q as the independent uniform distribution. We then redefine $\bar{p}$ and $\widetilde{p}$ as

$$\bar{p}(\boldsymbol{x}) = \underset{r \in \boldsymbol{E}_{2^*}(\theta^q_{123})}{\arg\min} D\left[p(\boldsymbol{x}) : r(\boldsymbol{x})\right], \quad \widetilde{p}(\boldsymbol{x}) = \underset{s \in \boldsymbol{E}_{1^*}(\boldsymbol{\theta}^q_{1^*})}{\arg\min} D\left[p(\boldsymbol{x}) : s(\boldsymbol{x})\right].$$

We now have theorem 6:

Theorem 6.

$$D\left[p:q\right] = D\left[p:\bar{p}\right] + D\left[\bar{p}:q\right] \tag{4.10}$$

$$= D\left[p:\widetilde{p}\right] + D\left[\widetilde{p}:q\right] \tag{4.11}$$

$$= D\left[p:\bar{p}\right] + D\left[\bar{p}:\widetilde{p}\right] + D\left[\widetilde{p}:q\right]. \tag{4.12}$$

The decompositions in the first and second lines are particularly interesting for neural data analysis purpose, as shown in the next section.

Any coordinate transformation can be done freely in this three-variable case in a numerical sense. In general, however, coordinate transformations between $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are not easy when the dimensions are high. Later, we discuss several practical approaches in n neuron case for use in neural data analysis.

4.4 Applications to Firing of Three Neurons. Here, we briefly discuss our application of the above results to firing of three neurons. The discussion in section 3.5 can be naturally extended. We consider three binary random variables, $X = (X_1, X_2, X_3)$, and denote our estimated probability distribution and the distribution of our null hypothesis by $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}})$ and $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$, respectively, where $\boldsymbol{\xi}$ is now a seven-dimensional coordinate system. We use the following decompositions,

$$D\left[\boldsymbol{\xi}^0 : \hat{\boldsymbol{\xi}}\right] = D\left[\boldsymbol{\zeta}_2^0 : \hat{\boldsymbol{\zeta}}_2''\right] + D\left[\hat{\boldsymbol{\zeta}}_2'' : \hat{\boldsymbol{\zeta}}\right] = D\left[\boldsymbol{\zeta}_1^0 : \hat{\boldsymbol{\zeta}}_1'\right] + D\left[\hat{\boldsymbol{\zeta}}_1' : \hat{\boldsymbol{\zeta}}\right],$$

where $\hat{\boldsymbol{\zeta}}_1' = (\boldsymbol{\eta}_{1^*}^0; \hat{\boldsymbol{\theta}}_{1^*})$ and $\hat{\boldsymbol{\zeta}}_2'' = (\boldsymbol{\eta}_2^0; \hat{\theta}_{123})$.

In the first decomposition, $D[\boldsymbol{\zeta}_2^0 : \hat{\boldsymbol{\zeta}}_2'']$ represents the discrepancy in the triplewise interaction of $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}})$ from $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$, fixing the pairwise interaction and marginals as specified by $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$. $D[\hat{\boldsymbol{\zeta}}_2'' : \hat{\boldsymbol{\zeta}}]$ then collects all the residual discrepancy and, more precisely, represents the discrepancy of the distribution $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}})$ from $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}_2'')$, which has the same simple and second-order marginals as those of $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$ (i.e., $\boldsymbol{\eta}_2^0$) and the same triplewise interaction $\hat{\theta}_{123}$ as that of $p(\boldsymbol{x}; \hat{\boldsymbol{\zeta}})$. Therefore, $D[\boldsymbol{\zeta}_2^0 : \hat{\boldsymbol{\zeta}}_2'']$ is particularly useful for investigating if there is any significant triplewise interaction in data, that is, $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}})$, in comparison with our null hypothesis $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$. A significant triplewise interaction, for example, may be considered indicative of three neurons functioning together. As for hypothesis testing, we can use

$$\lambda_2 = 2ND\left[\boldsymbol{\zeta}_2^0 : \hat{\boldsymbol{\zeta}}_2''\right] \approx N g_{77}^{\zeta}(\boldsymbol{\zeta}_2^0)(\theta_{123}^0 - \hat{\theta}_{123})^2 \sim \chi^2(1), \tag{4.13}$$

where N is the number of trials and the indices are $\boldsymbol{\zeta}_1 = (\zeta_1, \ldots, \zeta_6; \zeta_7) = (\boldsymbol{\eta}_2; \theta_{123})$.

In the second decomposition, $D[\zeta_1^0 : \hat{\zeta}_1']$ represents the discrepancy in both the triplewise and pairwise interactions of $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}})$ from $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$, fixing the marginals as specified by $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$, while $D[\hat{\zeta}_1' : \hat{\zeta}]$ collects all the residual discrepancy. $D[\zeta_1^0 : \hat{\zeta}_1']$ is useful to investigate if there is a significant coincident firing, taking the pairwise and triplewise interactions together, compared with the null hypothesis. We now have

$$\lambda_1 = 2ND\left[\zeta_1^0 : \hat{\zeta}_1'\right] \approx N \sum_{i,j=4}^{7} g_{ij}^{\zeta}(\zeta_1^0)(\zeta_i^0 - \hat{\zeta}_i)(\zeta_j^0 - \hat{\zeta}_j) \sim \chi^2(4), \tag{4.14}$$

where the indices are given by $\zeta = (\zeta_1, \ldots, \zeta_7) = (\boldsymbol{\eta}_{1^*}; \boldsymbol{\theta}_{1^*})$.

We can also compare two probability distributions estimated under different experimental conditions. Let us denote two estimated distributions by $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}^1)$ and $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}^2)$. We first detect the triplewise interaction. The MLE, denoted by $\hat{\zeta}_2^{10}$ and $\hat{\zeta}_2^{20}$, of our null hypothesis, that is, $\hat{\theta}_{123}^1 = \hat{\theta}_{123}^2$, is given by

$$\begin{aligned}(\hat{\zeta}_2^{10}, \hat{\zeta}_2^{20}) = \operatorname{argmax} \sum_{j}^{N} \log p(X^j : \zeta_2^1) p(X^j : \zeta_2^2) \\ \text{subject to } \theta_{123}^1 = \theta_{123}^2.\end{aligned} \tag{4.15}$$

Then we have

$$\begin{aligned}\lambda_2' &= 2ND\left[\hat{\zeta}_2^{10} : \hat{\zeta}_2^1\right] + 2ND\left[\hat{\zeta}_2^{20} : \hat{\zeta}_2^2\right] \\ &\approx N g_{77}^{\zeta}(\hat{\zeta}_2^{10})(\theta_{123}^{10} - \hat{\theta}_{123}^1)^2 + N g_{77}^{\zeta}(\hat{\zeta}_2^{20})(\theta_{123}^{20} - \hat{\theta}_{123}^2)^2 \\ &\sim \chi^2(2).\end{aligned} \tag{4.16}$$

When we investigate the coincident firing, taking the pairwise and triplewise interactions together, we use the second decomposition above. The MLE of our null hypothesis in this case is given by

$$\begin{aligned}(\hat{\zeta}_1^{10}, \hat{\zeta}_1^{20}) = \operatorname{argmax} \sum_{j}^{N} \log p(X^j : \zeta_1^1) p(X^j : \zeta_1^2) \\ \text{subject to } \boldsymbol{\theta}_{1^*}^1 = \boldsymbol{\theta}_{1^*}^2.\end{aligned} \tag{4.17}$$

For hypothesis testing, we can use

$$\begin{aligned}\lambda_1' &= 2ND\left[\hat{\zeta}_1^{10} : \hat{\zeta}_1^1\right] + 2ND\left[\hat{\zeta}_1^{20} : \hat{\zeta}_1^2\right] \\ &\approx N \sum_{i,j=4}^{7} g_{ij}^{\zeta}(\hat{\zeta}_1^{10})(\zeta_i^{10} - \hat{\zeta}_i)(\zeta_j^{10} - \hat{\zeta}_j)\end{aligned}$$

$$+ N \sum_{i,j=4}^{7} g_{ij}^{\zeta}(\hat{\zeta}_1^{20})(\zeta_i^{10} - \hat{\zeta}_i)(\zeta_j^{10} - \hat{\zeta}_j)$$

$$\sim \chi^2(8). \tag{4.18}$$

The decompositions in the KL divergence also allow us to decompose mutual information between the firing pattern of three neurons $X = (X_1, X_2, X_3)$ and the behavior Y in a similar manner to section 3.6.

Theorem 7.

$$I(X, Y) = E_{p(X,Y)}\left[\log \frac{p(\boldsymbol{x}, y)}{p(\boldsymbol{x})p(y)}\right]$$

$$= I_1(X, Y) + I_2(X, Y) \tag{4.19}$$

$$= I_3(X, Y) + I_4(X, Y) \tag{4.20}$$

where

$$I_1(X, Y) = E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_1(X \mid y) : \boldsymbol{\zeta}_1(X, y)\right]\right],$$
$$I_2(X, Y) = E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_1(X, y) : \boldsymbol{\zeta}_1(X)\right]\right]$$

and we define $\boldsymbol{\zeta}_1(X, y) = (\boldsymbol{\eta}_1(X \mid y); \boldsymbol{\theta}_{1^*}(X))$. *Similarly,*

$$I_3(X, Y) = E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_2(X \mid y) : \boldsymbol{\zeta}_2(X, y)\right]\right],$$
$$I_4(X, Y) = E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_2(X, y) : \boldsymbol{\zeta}_2(X)\right]\right]$$

and we define $\boldsymbol{\zeta}_2(X, y) = (\boldsymbol{\eta}_2(X \mid y); \boldsymbol{\theta}_{2^*}(X))$.

In equation 4.19, the mutual information $I(X, Y)$ is decomposed into two parts: I_1, the mutual information conveyed by the pairwise and triplewise interactions of the firing, and I_2, the mutual information conveyed by the mean firing-rate modulation. In equation 4.20, $I(X, Y)$ is decomposed differently: I_3, conveyed by the triplewise interaction, and I_4, conveyed by the other terms, that is, the pairwise and mean firing-rate modulations.

5 General Case: Joint Distributions of $X_1, \ldots, X_n$

Here we study a general case of n neurons. Let $X = (X_1, \ldots, X_n)$ be n binary variables, and let $p = p(\boldsymbol{x}), \boldsymbol{x} = (x_1, \ldots, x_n), x_i = 0, 1$, be its probability, where we assume $p(\boldsymbol{x}) > 0$ for all $\boldsymbol{x}$. We begin with briefly recapitulating Amari (2001) for the theoretical framework and then move to its applications.

5.1 Coordinate Systems of S_n. As mentioned in section 2, the set of all probability distributions $\{p(\boldsymbol{x})\}$ forms a (2^n-1)–dimensional manifold $\boldsymbol{S}_n$. Any $p(\boldsymbol{x})$ in $\boldsymbol{S}_n$ can be represented by the P-coordinate system, θ-coordinate system, or η-coordinate system. The P-coordinate system is defined by

$$p_{i_1\cdots i_n} = \text{Prob}\{X_1 = i_1, \ldots, X_n = i_n\}, \quad i_k = 0, 1,$$
$$\text{subject to} \sum_{i_1,\ldots,i_n} p_{i_1\cdots i_n} = 1.$$

The θ-coordinate system is defined by the expansion of $\log p(\boldsymbol{x})$ as

$$\log p(\boldsymbol{x}) = \sum \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \sum_{i<j<k} \theta_{ijk} x_i x_j x_k \cdots + \theta_{1\cdots n} x_1 \cdots x_n - \psi, \tag{5.1}$$

where the indices of $\theta_{ijk}, \ldots,$ satisfy $i < j < k$ and then

$$\boldsymbol{\theta} = (\theta_i, \theta_{ij}, \theta_{ijk}, \ldots, \theta_{12,\ldots,n}) \tag{5.2}$$

has $2^n - 1$ components and forms the θ-coordinate system. It is easy to compute any components of $\boldsymbol{\theta}$; for example, we can get $\theta_1 = \log \frac{p_{10,\ldots,0}}{p_{0,\ldots,0}}$. For later convenience, we use the notation $\boldsymbol{\theta}_1 = (\theta_i)$, $\boldsymbol{\theta}_2 = (\theta_{ij})$, $\boldsymbol{\theta}_3 = (\theta_{ijk}), \ldots, \boldsymbol{\theta}_n = \theta_{12,\ldots,n}$, where l in $\boldsymbol{\theta}_l$ runs over l-tuple among n binary numbers, yielding ${}_nC_l$ components (${}_nC_l$ is the binomial coefficient). Then we can write

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

On the other hand, the η-coordinate system is defined by using

$$\eta_i = E[x_i] \ (i = 1, \ldots, n),$$
$$\eta_{ij} = E\left[x_i x_j\right] \quad (i < j), \ldots, \eta_{12,\ldots,n} = E[x_i, \ldots, x_n],$$

which has $2^n - 1$ components (see section 2)—in other words,

$$\boldsymbol{\eta} = \left(\eta_i, \eta_{ij}, \ldots, \eta_{1,\ldots,n}\right)$$

forms the η-coordinate system in $\boldsymbol{S}_n$. We also write $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_n)$, which is linearly related to $\{p_{i_1,\ldots,i_n}\}$.

In the rest of this section, we provide some general concepts in information geometry. Readers who are interested in more detail can refer to Amari and Nagaoka (2000). When a submanifold of $\boldsymbol{S}_n$, denoted by $\boldsymbol{E}$, is represented by linear constraints among the θ coordinates, $\boldsymbol{E}$ is called exponentially flat or e flat. When a submanifold of $\boldsymbol{S}_n$, denoted by $\boldsymbol{M}$, is represented by linear constraints among the η coordinates, $\boldsymbol{M}$ is called mixture flat or m flat.

The Fisher information matrices in the respective coordinate systems play the role of Riemannian metric tensors. The two coordinate systems $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are dually coupled in the following sense. Let $A, B, C, \ldots$ denote ordered subsets of indices, which stand for components of $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$, that is, $\boldsymbol{\theta} = (\theta_A)$, $\boldsymbol{\eta} = (\eta_B)$.

Theorem 8. *The two metric tensors $G(\boldsymbol{\theta})$ and $\bar{G}(\boldsymbol{\eta})$ are mutually inverse,*

$$\bar{G}(\boldsymbol{\theta}) = G(\boldsymbol{\eta})^{-1}, \tag{5.3}$$

where $G(\boldsymbol{\eta}) = (g_{AB}(\boldsymbol{\eta}))$ and $\bar{G}(\boldsymbol{\theta}) = (\bar{g}_{AB}(\boldsymbol{\theta}))$ are defined by

$$g_{AB}(\boldsymbol{\theta}) = E\left[\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_A}\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_B}\right],$$
$$\bar{g}_{AB}(\boldsymbol{\eta}) = E\left[\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\eta})}{\partial \eta_A}\frac{\partial \log p(\boldsymbol{x};\boldsymbol{\eta})}{\partial \eta_B}\right].$$

The following generalized Pythagoras theorem has been known in $\boldsymbol{S}_n$ (Csiszár, 1967, 1975; Amari et al., 1992; Amari & Han, 1989). It holds in more general cases, playing an important role in information geometry (Amari, 1987; Amari & Nagaoka, 2000).

Theorem 9. *Let $p(\boldsymbol{x})$, $q(\boldsymbol{x})$ and $r(\boldsymbol{x})$ be three distributions where the m geodesic connecting $p(\boldsymbol{x})$ and $q(\boldsymbol{x})$ is orthogonal to the e geodesic connecting $q(\boldsymbol{x})$ and $r(\boldsymbol{x})$. Then,*

$$D\left[p:q\right] + D\left[q:r\right] = D\left[p:r\right]. \tag{5.4}$$

5.2 Higher-Order Interactions. This section aims at defining the higher-order interactions using the k-cut mixed coordinate system. Section 5.1 introduced $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$ and $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$, each of which spans $\boldsymbol{S}_n$. Let us define their partitions, called a k-cut, as follows,

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_{k-}; \boldsymbol{\theta}_{k+}\right), \quad \boldsymbol{\eta} = \left(\boldsymbol{\eta}_{k-}; \boldsymbol{\theta}_{k+}\right), \tag{5.5}$$

where $\boldsymbol{\theta}_{k-}$ and $\boldsymbol{\eta}_{k-}$ consist of coordinates whose subindices have no more than k indices, that is, $\boldsymbol{\theta}_{k-} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \ldots, \boldsymbol{\theta}_k)$, $\boldsymbol{\eta}_{k-} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \ldots, \boldsymbol{\eta}_k)$, and $\boldsymbol{\theta}_{k+}$ and $\boldsymbol{\eta}_{k+}$ consist of the coordinates whose subindices have more than k indices, that is, $\boldsymbol{\theta}_{k+} = (\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_{k+2}, \ldots, \boldsymbol{\theta}_n)$, $\boldsymbol{\eta}_{k+} = (\boldsymbol{\eta}_{k+1}, \boldsymbol{\eta}_{k+2}, \ldots, \boldsymbol{\eta}_n)$.

First, note that $\boldsymbol{\eta}_{k-}$ specifies the marginal distributions of any k (or less than k) random variables among $X_1, \ldots, X_n$. Let us consider a family of m-flat submanifold in $\boldsymbol{S}_n$,

$$\boldsymbol{M}_k(\boldsymbol{m}_k) = \left\{\boldsymbol{\eta} \mid \boldsymbol{\eta}_{k-} = \boldsymbol{m}_k\right\}.$$

It consists of all the distributions having the same k-marginals specified by a fixed $\boldsymbol{\eta}_{k-} = \boldsymbol{m}_k$. They differ from one another only by higher-order interactions of more than k variables.

Second, all coordinate curves represented by $\boldsymbol{\theta}_{k+}$ are orthogonal to $\boldsymbol{\eta}_{k-}$ or any components of $\boldsymbol{\eta}_{k-}$. Hence, $\boldsymbol{\theta}_{k+}$ represents interactions among more than k variables independent of the k marginals, $\boldsymbol{\eta}_{k-}$. Then, for a constant vector $\boldsymbol{c}_k$, let us compose a family of e-flat submanifolds,

$$\boldsymbol{E}_{k+}(\boldsymbol{c}_k) = \left\{\boldsymbol{\theta} \mid \boldsymbol{\theta}_{k+} = \boldsymbol{c}_k\right\}.$$

Third, $\boldsymbol{E}_{k+}(\boldsymbol{c}_k)$ and $\boldsymbol{M}_k(\boldsymbol{m}_k)$ are mutually orthogonal and introduce a new coordinate system, called the k-cut mixed coordinate system, defined by

$$\boldsymbol{\zeta}_k = \left(\boldsymbol{\eta}_{k-}; \boldsymbol{\theta}_{k+}\right).$$

Any k-cut mixed coordinate system forms the coordinate system of $\boldsymbol{S}_n$. A change in the $\boldsymbol{\theta}_{k+}$ part preserves the k-marginals of $p(\boldsymbol{x})$ (i.e., $\boldsymbol{\eta}_{k-}$), while a change in the $\boldsymbol{\eta}_{k-}$ part preserves the interactions among more than k variables. These changes are mutually orthogonal. Thus, $\boldsymbol{E}_{k+}(\boldsymbol{\theta}_{k+})$ is regarded as the submanifold consisting of distributions having the same degree of higher-order interactions. When $\boldsymbol{\theta}_{k+} = 0$, $\boldsymbol{E}_{k+}(0)$ denotes the set of all the distributions having no intrinsic interactions of more than k variables.

5.3 Projections and Decompositions of Higher-Order Interactions. Given $p(\boldsymbol{x})$, we define $p^{(k)}(\boldsymbol{x}) = \prod^{(k)} p$ by

$$p^{(k)}(\boldsymbol{x}) = \prod^{(k)} p = \underset{q \in \boldsymbol{E}_{k+}(0)}{\arg\min} D\left[p : q\right].$$

This is the point closest to p among those that do not have intrinsic interactions of more than k variables. We note that another characterization of $p^{(k)}$ is given by

$$p^{(k)}(\boldsymbol{x}) = \underset{q \in \boldsymbol{M}_k(\boldsymbol{\eta}^p_{k-})}{\arg\min} D\left[q : p^{(0)}\right],$$

where it should be easy to see $p^{(0)}$ a uniform distribution by definition of $p^{(0)}$. The e geodesic connecting $p^{(k)}$ and $p^{(0)}$ is orthogonal to $\boldsymbol{M}_k(\boldsymbol{\eta}^p_{k-})$ to which the original p belongs.

The k-cut mixed coordinates of $p^{(k)}$ are given by $\boldsymbol{\zeta}_k(p^{(k)}) = (\boldsymbol{\eta}_{k-}, \boldsymbol{\theta}_{k+} = 0)$. The degree of interactions higher than k is hence defined by $D[p : p^{(k)}]$. Since the m geodesic connecting p and $p^{(k)}$ is orthogonal to $\boldsymbol{E}_{k+}(0)$, the Pythagoras theorem guarantees the following decomposition:

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

Let us put

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

Then $D_k(p)$ is interpreted as the degree of interaction purely among the k variables. We then have the following decomposition in which $D_k(p)$ denotes the degree of interaction among k variables.

Theorem 10.

$$D\left[p : p^{(0)}\right] = \sum_{k=1}^{n} D_k(p). \tag{5.6}$$

It is straightforward to generalize the above results when we are given two distributions $p(\boldsymbol{x})$, $q(\boldsymbol{x})$. Let us define

$$\boldsymbol{\zeta}_k^{(k')} = \boldsymbol{\zeta}_k(p^{(k')}) = (\boldsymbol{\eta}_{k-}(p); \boldsymbol{\theta}_{k+}(q)).$$

Then we have

$$D[p : q] = D[\boldsymbol{\zeta}_k^p : \boldsymbol{\zeta}_k^q] = D[\boldsymbol{\zeta}_k^p : \boldsymbol{\zeta}_k^{(k')}] + D[\boldsymbol{\zeta}_k^{(k')} : \boldsymbol{\zeta}_k^q],$$

which is induced from theorem 9. By defining,

$$D_{k'}(p) = D\left[p^{(k')} : p^{((k-1)')}\right],$$

we obtain

$$D\left[p : q\right] = \sum_{k=1}^{n} D_{k'}(p). \tag{5.7}$$

The decompositions shown in equations 5.6 and 5.7 are obviously similar to each other. A critical difference, however, exists in the interpretation of the two decompositions. Each term in equation 5.6, $D_k(p)$, represents the degree of purely kth order interaction, whereas $D_{k'}(p)$ in equation 5.7 does not necessarily do so. This is because $\boldsymbol{\zeta}_k(p^{(k)})$ always has $\boldsymbol{\theta}_{k+} = 0$, that is, the higher-order coordinates than the kth order. On the other hand, $\boldsymbol{\zeta}_k(p^{(k')})$ does not necessarily have zero in the corresponding part. In other words, $\boldsymbol{\theta}_k$ represents the pure kth order interaction only if $\boldsymbol{\theta}_{k+} = 0$.

5.4 Application to Neural Firing. To apply the results in the above sections to neural firing data, the discussion for the case of the three neurons can be directly applied. Hence, we mainly provide some remarks in this section.

First, suppose that we have an estimated probability distribution of n neurons, denoted by $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}})$, and a probability distribution of our null hypothesis, $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$. Then, using the k-cut mixed coordinates, we obtain the decomposition,

$$D[\boldsymbol{\xi}^0 : \hat{\boldsymbol{\xi}}] = D[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}_k'] + D[\hat{\boldsymbol{\zeta}}_k' : \hat{\boldsymbol{\zeta}}],$$

where we define $\hat{\boldsymbol{\zeta}}_k' = (\boldsymbol{\eta}_{k-}^0; \hat{\boldsymbol{\theta}}_{k+})$. In this decomposition, $D[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}_k']$ represents the discrepancy between $\boldsymbol{\xi}^0$ and $\hat{\boldsymbol{\xi}}$ in the interactions higher than the kth order and $D[\hat{\boldsymbol{\zeta}}_k' : \hat{\boldsymbol{\zeta}}]$ equal to and lower than the kth order. We can also convert these divergences to a χ^2 test. For example, we have

$$\lambda_{k+} = 2ND[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}_k'] \approx N \sum_{A,B>k} g_{AB}^{\zeta_k}(\boldsymbol{\zeta}^0)(\zeta_A^0 - \hat{\zeta}_A)(\zeta_B^0 - \hat{\zeta}_B) \sim \chi^2(m), \tag{5.8}$$

where N is the number of trials and A, B runs over all of the indices higher than the k-tuple, included in $\boldsymbol{\theta}_{k+}$, and the degree of freedom m in the χ^2 test is given by $m = \sum_{l=k+1}^{n} {}_nC_l$. By using λ_{k+}, we can test whether there is any significant contribution of a higher-order interaction by summing all interactions higher than the kth order.

In equation 5.8, $g_{AB}^{\zeta_k}$ corresponds to the part, denoted by D_{ζ_k}, of the Fisher information matrix of the k-cut mixed coordinates, G_{ζ_k}. Let us write the Fisher information matrix of different coordinate systems in block form as follows:

$$G_{\zeta_k} = \begin{bmatrix} A_{\zeta_k} & O \\ O & D_{\zeta_k} \end{bmatrix}, \qquad G_\eta = \begin{bmatrix} A_\eta & B_\eta \\ B_\eta^T & D_\eta, \end{bmatrix} \qquad G_\theta = \begin{bmatrix} A_\theta & B_\theta \\ B_\theta^T & D_\theta \end{bmatrix}.$$

The following theorem gives an explicit form of G_{ζ_k}.

Theorem 11. *The Fisher information matrix of the k-cut mixed coordinates, G_{ζ_k}, is given by*

$$A_{\zeta_k} = A_\theta^{-1}, \quad D_{\zeta_k} = D_\eta^{-1}. \tag{5.9}$$

Note that G_θ is easy to obtain from the experimental data because $g_{AB}^\theta = E_\theta[X_A X_B] - \eta_A \eta_B$, component-wise. Given G_θ, the computation of G_η may be said to be easy in the sense of $G_\eta = G_\theta^{-1}$. Thus, we can easily compute A_{ζ_k} and D_{ζ_k}.

Suppose two probability distributions are given, estimated under different experimental conditions, denoted by $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}^1)$ and $p(\boldsymbol{x}; \hat{\boldsymbol{\xi}}^2)$. Our task is to test whether any higher-order (than the kth order) interaction is significantly different between the two distributions. In this case, we first need to solve the following MLE:

$$(\hat{\boldsymbol{\zeta}}_k^{10}, \hat{\boldsymbol{\zeta}}_k^{20}) = \operatorname{argmax} \sum_j^N \log p(X^j; \boldsymbol{\zeta}_k^1) p(X^j; \boldsymbol{\zeta}_k^2)$$
$$\text{subject to } \boldsymbol{\theta}_{k+}^1 = \boldsymbol{\theta}_{k+}^2. \tag{5.10}$$

The χ^2 test is then given as

$$\begin{aligned}\lambda'_{k+} &= 2ND[\hat{\boldsymbol{\zeta}}_k^{10} : \hat{\boldsymbol{\zeta}}_k^{1}] + 2ND[\hat{\boldsymbol{\zeta}}_k^{20} : \hat{\boldsymbol{\zeta}}_k^{2}] \\ &\approx N \sum_{A,B>k} g_{AB}^{\zeta_k}(\hat{\boldsymbol{\zeta}}_k^{10})(\zeta_A^{10} - \hat{\zeta}_A^{1})(\zeta_B^{10} - \hat{\zeta}_B^{1}) \\ &\quad + N \sum_{A,B>k} g_{AB}^{\zeta_k}(\hat{\boldsymbol{\zeta}}_k^{20})(\zeta_A^{20} - \hat{\zeta}_A^{2})(\zeta_B^{20} - \hat{\zeta}_B^{2}) \\ &\sim \chi^2(2m), \qquad \text{where} \quad m = \sum_{l=k+1}^{n} {}_nC_l.\end{aligned}$$

To relate the neural firing with discrete behavioral choice, denoted by Y, we have the following decomposition in the mutual information.

Theorem 12.

$$I(X, Y) = I_{k+}(X, Y) + I_{k-}(X, Y), \tag{5.11}$$

where we define

$$\begin{aligned}I_{k+}(X, Y) &= E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_k(X \mid y) : \boldsymbol{\zeta}_k(X, y)\right]\right], \\ I_{k-}(X, Y) &= E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_k(X, y) : \boldsymbol{\zeta}_k(X)\right]\right]\end{aligned}$$

and $\boldsymbol{\zeta}_k(X, y) = (\boldsymbol{\eta}_{k-}(X \mid y); \boldsymbol{\theta}_{k+}(X))$.

5.5 Homogeneous Case. This section discusses an approach under the assumption of the homogeneity of neural activities, which is useful for avoiding some practical difficulties (Grün & Diesmann, 2000). Here, homogeneity refers to the assumption of homogeneous neural firing—that the interaction of some kth orders is the same as another among all k tuple neurons. For simplicity, we mostly assume below that the interaction of any kth order is the same as another among all k tuple neurons, which will be referred to as full homogeneous assumption.

It is easy, using Rota's methods in the set theory in relation to principle of inclusion and exclusion (Amari, 2001), to explicitly write down the coordinate transformations between the P and η coordinates and also between the P and θ coordinates. Hence, it is possible in principle to do the coordinate transformation between the η and θ coordinates and also between the k mixed and P coordinates. There are remaining practical difficulties, however, in two aspects. First, the computational complexity in these coordinate transformations increases exponentially. Second, the limitation in the number of samples becomes severe in estimating the values in any coordinate system as the number of neurons increases.

One approach to overcoming these difficulties is to use the homogeneity assumption. Under the full homogeneity assumption, the $2^n - 1$ dimensions in the n-neuron case reduce to n dimensions. The coordinate transformation in this case is given by theorem 13.

Theorem 13.

$$\eta_k = \sum_{l=0}^{n-k} {}_{(n-k)}C_l P_{k+l}, \quad \log P_k = \sum_{l=1}^{k} {}_kC_l \theta_l - \psi, \tag{5.12}$$

where ${}_nC_k$ denotes the binomial coefficient. Equivalently, we also have

$$P_k = \sum_{l=0}^{n-k} {}_{(n-k)}C_l (-1)^l \eta_{k+l}, \quad \theta_k = \sum_{l=0}^{k} {}_kC_l (-1)^{(k-l)} \log P_l. \tag{5.13}$$

All of the results in the previous sections can be rewritten by using these simplified coordinates.

5.6 Interesting Set of Neurons. One important question in this n-neurons case is how to find a set of neurons whose firing shows significant coincident firing. We discuss one practical approach here, mentioning several mathematical tricks to overcome some practical issues. In practice, this question may occur either when (1) we simply want to give an answer to the question by some rigorous tests or (2) when, in facing a vast amount of data, we only want to find some candidates of interesting sets of neurons for further investigation. In this second case, it is perhaps unnecessary to be too rigorous; simplicity is preferred, although some complexity is inevitably involved. A method for this purpose is particularly needed given the increasing number of neurons simultaneously recorded in experiments. This section is formulated accordingly.

One obvious approach is to find a significant kth order interaction ($k < n$) and group a significant k-tuple of the neurons as an interesting set of neurons. A basic procedure is given as follows:

1. Given the population of recorded neurons, calculate the θ coordinates and find nonzero θ of the largest order, denoted by θ_k^A, where A indicates a specific k-tuple and k is the order.
2. There can be multiple θ_k^A in the same kth order, but we discard any interaction between them. Under this circumstance, each A is specified by $\boldsymbol{\zeta}_k = (\boldsymbol{\eta}_{k-}^A; \theta_k^A)$.
3. We set our null hypothesis, $\boldsymbol{\zeta}_k^0 = (\boldsymbol{\eta}_{k-}^A; \theta_k^0)$, and then compute the χ^2 value to test whether θ_k^A is significant against θ_k^0.
4. If A is found to be significant, we nominate A as an interesting set of neurons, denoted by A^*.

5. If A is insignificant, we go down to the lower order. In doing so, we omit the neurons that are already registered in $\{A^*\}$ (i.e., any subset of A^*) from consideration. Given these, when we find nonzero θ of the next largest order, $\theta^A_{k'}$, we repeat from step 2.

In the above procedure, there are some practical concerns to solve. First, the coordinate transformation from $\boldsymbol{\zeta}_k$ to $\boldsymbol{p}$ becomes computationally expensive as the number of neurons increases. In our procedure, this transformation is needed to compute the χ^2 value. The χ^2 value is obtained by using the Fisher metric component at $\boldsymbol{\zeta}^0_k$, which is computed by use of this transformation (i.e., $\boldsymbol{\zeta}^0_k$ to $\boldsymbol{p}^0$). We now recall that the reason we choose to compute the Fisher metric component at $\boldsymbol{\zeta}^0_k$ comes from hypothesis testing formalism and that the χ^2 value is a quadratic approximation of the KL divergence. Therefore, for our testing, we can in fact use the Fisher metric at $\boldsymbol{\zeta}_k$ to compute the χ^2 value,

$$\lambda = 2ND[\boldsymbol{\zeta}^0_k : \boldsymbol{\zeta}_k] \approx Ng(\boldsymbol{\zeta}_k)(\theta^0_k - \theta^A_k)^2 \sim \chi^2(1). \tag{5.14}$$

The value of λ is easy to compute because all coordinates at $\boldsymbol{\zeta}_k$ are easily obtained from the data.

Second, we should ask how to set θ^0_k, that is, our null hypothesis. Apparently, there are two choices. One is to choose $\theta^0_k = 0$, which is a hypothesis of no purely kth-order interaction. The other is to obtain the value of θ_k under the homogeneous assumption, denoted by $\boldsymbol{\zeta}^*_k = (\boldsymbol{\eta}^*_{k-}; \theta^*_k)$, and set $\theta^0_k = \theta^*_k$. Both choices are feasible, but the underlying assumption differs in each choice. Third, a practical concern is the limited number of trials. One practical solution is to use the homogeneity assumption. For example, when we are interested only in which order should be considered as the order of the interesting set of neurons, we suggest using $\boldsymbol{\zeta}^*_k$ to compare against the null hypothesis of no purely kth-order interaction, that is, $\boldsymbol{\zeta}^0_k = (\boldsymbol{\eta}^*_k; 0)$. As another example, in case we happen to find a specific k-tuple A firing ($\theta^A_k \neq 0$), we may use $\boldsymbol{\zeta}_k = (\boldsymbol{\eta}^A_1; \boldsymbol{\eta}^*_2, \ldots, \boldsymbol{\eta}^*_{(k-1)}; \theta^A_k)$ under the partial homogeneous assumption to compare against the null hypothesis $\boldsymbol{\zeta}^0_k = (\boldsymbol{\eta}^A_1; \boldsymbol{\eta}^*_2, \ldots, \boldsymbol{\eta}^*_{(k-1)}; 0)$.

6 Examples

In this section, we demonstrate our method using artificial data. (More examples, including application of the proposed method to experimental data and also to autocorrelation, are available as a technical report: Nakahara, Amari, Tatsuno, Kang, & Kobayashi, 2002.)

6.1 Example 1: Firing of Two Neurons.

6.1.1 Coincident Firing. In this simulation, we aim to demonstrate a relation between correlation coefficient and θ, and hypothesis testing under different null hypotheses. Figures 2A and 2B give the mean firing frequency of two neurons and their correlation coefficients (COR; the N-JPSTH), respectively (see the legend for spike generation). Period a is the control period. The neural firing in this period presumably indicates the resting level activity, which was set to have a very weak correlation here. The firing is almost independent in both periods b and c, whereas it is not independent in period d, whose COR is larger than that of period a.

Figure 2C shows θ_3, a quantity in our measure to indicate the pairwise interaction. At first glance, the time course of θ_3 may look similar to the COR. A careful inspection, however, shows that they are different (e.g., the relative magnitudes between periods a and d are different for COR and θ). This is because $(\eta_1, \eta_2, \theta_3)$ and $(\eta_1, \eta_2, \mathrm{COR})$ form different coordinate systems, although both θ_3 and COR represent the correlational component.

The KL divergence is used to measure the discrepancy between two probability distributions—one distribution to be examined and the other to be of the null hypothesis. Using the orthogonality of θ_3 with (η_1, η_2), we can decompose the KL divergence into two terms—one representing the discrepancy in the correlational component and the other the discrepancy in the mean firing. We call the former the KL divergence in correlation for convenience.

We first examine the KL divergence in correlation against the null hypothesis of independent firing (i.e., $\theta_3 = 0$), which is the probability distribution with the same marginals of the examined probability but with the independent firing. This KL divergence, denoted by KL1, is indicated by the solid line in Figure 2D. Compared with θ_3, the KL1 takes into account the metric in the probability space. In Figure 2D, we also indicate by the dashed line the corresponding p-value, derived from the $\chi^2(1)$ distribution (e.g., if the p-value reaches 0.95, it is significant with $p < 0.05$). We observe that period d is significant in Figure 2D.

We now examine the KL divergence in correlation against the null hypothesis of the averaged activity (including both averaged mean firing rates and averaged coincident firing) in the control period. In Figure 2E, this KL divergence in correlation, denoted by KL2(1), is indicated by a solid line, while the corresponding p-value, derived from $\chi^2(1)$, is indicated by a dashed line. In Figure 2E, period d is no longer significant under this null hypothesis. Period c now is clearly significant, whereas period b is barely significant, although θ_3 is the same between the two periods (see Figure 2C). This is because the mean firing rates are different between periods b and c, and the Fisher information metric takes the geometrical structure into account—not only the degree of the coincident firing (θ_3) but also the marginal probability (i.e., the mean firing rates).

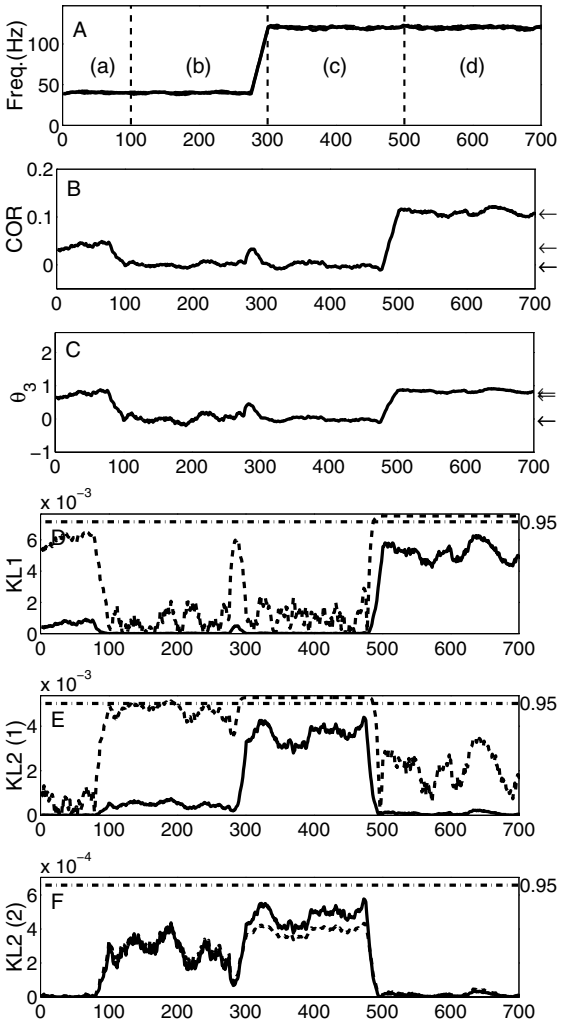
A comparison between Figures 2D and 2E illustrates a simple fact that what should be considered as a significant event depends on what is taken as the null hypothesis. In Figure 2E, where the p-value is from $\chi^2(1)$ formulation, the underlying assumption is that the averaged activity in the control period, which is estimated from data in practice, is regarded as a true average activity in the control period (see section 3.5). When we are interested in comparing significances between two null hypotheses of independent firing and the averaged activity in the control period, we suggest that the comparison of p-values from $\chi^2(1)$ is informative, as in Figures 2D and 2E.

In contrast, since we estimate the average activity in the control period from data in practice, a more proper and more conservative hypothesis testing is to use a different formulation of the likelihood ratio test. This results in using the p-value associated with $\chi^2(2)$ (see section 3.5). In Figure 2F, we indicate by a solid line the KL2(2), the KL divergence in the correlation that corresponds to this $\chi^2(2)$ formulation. The dashed line in-

Figure 2: *Facing page*. Example of a two-neuron case to detect the significant pairwise correlation. The spikes of two neurons were generated such that whether a spike exists or not in each bin (1 ms bin width is used) was probabilistically determined in each trial (where the number of all trials was 2000), given an assumed probability $(\eta_1, \eta_2, \eta_{12})$ in each period a–d. Period a (0–100 ms) is with $(\eta_1, \eta_2, \eta_{12}) = (0.04, 0.04, 0.0031)$. Period b (100–300 ms) is with (0.04, 0.04, 0.0016). Period c (300–500 ms) is with (0.12, 0.12, 0.0144). Period d (500–700 ms) is with (0.12, 0.12, 0.0260). To estimate the probabilities from artificially sampled data, averaged values in each bin were obtained over all trials, and the value in each bin was then finally determined by smoothing over several bins (set as 25 ms). (A) Mean firing frequency for two neurons. Because the mean firing frequency of the two neurons was the same, the two lines are superimposed with very little fluctuation. (B) Correlation coefficient. (C) θ_3 $(= \theta_{12})$. (D) KL divergence in correlation against the null hypothesis of independent firing and the corresponding p-value, derived from $\chi^2(1)$, are indicated by solid and dashed lines, respectively (see the text). (E) KL divergence in correlation against the null hypothesis of the averaged activity in the control period and the corresponding p-value, from $\chi^2(1)$, are indicated by solid and dashed lines, respectively. (F) Using the formulation that the firing in the control period and in other periods is from the same correlation level (see section 3.5), the p-value, derived from $\chi^2(2)$, is indicated by dashed line, while the sum of corresponding KL divergences is indicated by solid line. Arrows at the right-hand side of B and C indicate the true values (note that some arrows are superimposed since they are close each other). In all examples in this section, the values of the η coordinates are provided in each figure legend. The η coordinates can be easily converted to the P coordinates used to generate sampled data, by which we obtained the estimated values of any coordinates in figures, and given the P coordinates, it is simple to compute true θ values (e.g., equations 4.3–4.6 for Figure 4). In all examples, the estimated θ values reasonably match with the true ones (e.g., Figure 6C).

dicates the p-value, now from $\chi^2(2)$. We can observe that the p-value in Figure 2F gives a more conservative estimate compared with the p-value in Figure 2E.

6.1.2 Mutual Information Between Firing and Behavior. Figure 3 shows the decomposition of mutual information (MI) between firing and behavior, using artificial data. We assumed only two choices for the behavior, denoted by s1 and s2. Figures 3A and 3B show the mean firing frequency with respect to s1 and s2, respectively. The mean firing of both neurons is the same between s1 and s2 in period a, assumed to be the control period. In periods b



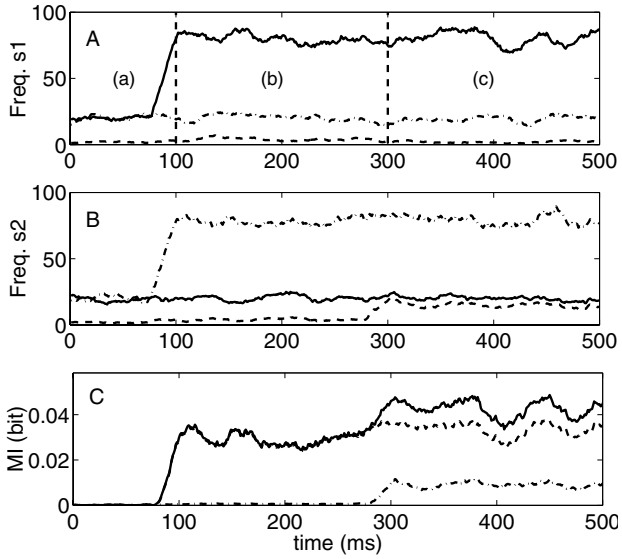


Figure 3: Example of a two-neuron case to obtain mutual information (MI) between firing and behavior and its decomposition. The spikes of two neurons were generated and estimated in a similar manner to Figure 2. The number of stimulus condition is assumed to be two, denoted by s1 and s2, and the number of trials per stimulus was 500. In period a (0–100 ms), assumed probabilities $(\eta_1, \eta_2, \eta_{12})$ for s1 and s2 were given as $(0.02, 0.02, 0.002)$ and $(0.02, 0.02, 0.002)$, respectively. In period b, (100–300 ms), $(0.08, 0.02, 0.004)$ and $(0.02, 0.08, 0.004)$; in period c, (300–500 ms), $(0.08, 0.02, 0.002)$ and $(0.02, 0.08, 0.015)$. (A, B) Mean firing frequency of the two neuron with respect to s1 (A) and s2 (B), respectively. A solid line indicates the mean firing of one neuron and a dashed-dotted line the mean firing of the other neuron. A dashed line indicates the mean firing of the coincident firing. (C) MI. The (total) MI is indicated by a solid line and is decomposed into two terms: the MI by the modulation of the mean firing rate (dashed line) and the MI by the modulation of the pairwise correlation (dashed-dotted line).

and c, assumed to be the test periods, we have set the mean firing of the two neurons as a somewhat mirror image between s1 and s2. The mean firing of each neuron stays the same in the test periods. Notably, however, the mean coincident firing increases in period c only when s2 is given (period c in Figure 3B).

In Figure 3C, we show the (total) MI between firing and behavior and its decomposition. The total MI (the solid line in Figure 3C) exists in periods b and c. Its magnitude is larger in period c than in period b, although the mean firing of each neuron stays the same in both periods. This is because

the coincident firing is modulated by the behavioral choices only in period c. This observation can be directly examined by looking into the decomposed MIs (see the figure legend). In brief, we observe that the increase in the total MI in period c comes almost exclusively from the part of the MI by the modulation of the coincident firing (indicated by the dashed-dot line in Figure 3C).

6.2 Example 2: Firing of Three Neurons.

6.2.1 Inspection of Triplewise Interaction. Figure 4A gives $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \boldsymbol{\eta}_3)$, where the firing is assumed to be homogeneous for simplicity. Period a is assumed to be the control period. Figure 4B shows COR. The COR in the control period is almost zero, while the CORs in periods b and c are almost the same as each other, both being different from zero. Yet when we are more careful in looking into the interaction, using θ_{ij} (see Figure 4C) and θ_{123} (see Figure 4D), we see that the nature of the interaction is largely different between periods b and c.

The triplewise interaction, θ_{123}, shown in Figure 4D is nearly zero in periods a and b. Hence, θ_{ij} in Figure 4C indicates the purely pairwise correlation in these periods. On the other hand, since θ_{123} is not zero in period c, θ_{ij} in this period does not represent the purely pairwise correlation any more, although it is still correct to say that the pairwise correlation is different between periods b and c by simply observing that θ_{ij}'s are different in the two periods. The fact that θ_{123} is not zero in period c indicates that the purely triplewise interaction exists in this period, where θ_{123} is negative so that the triplewise interaction is negative.

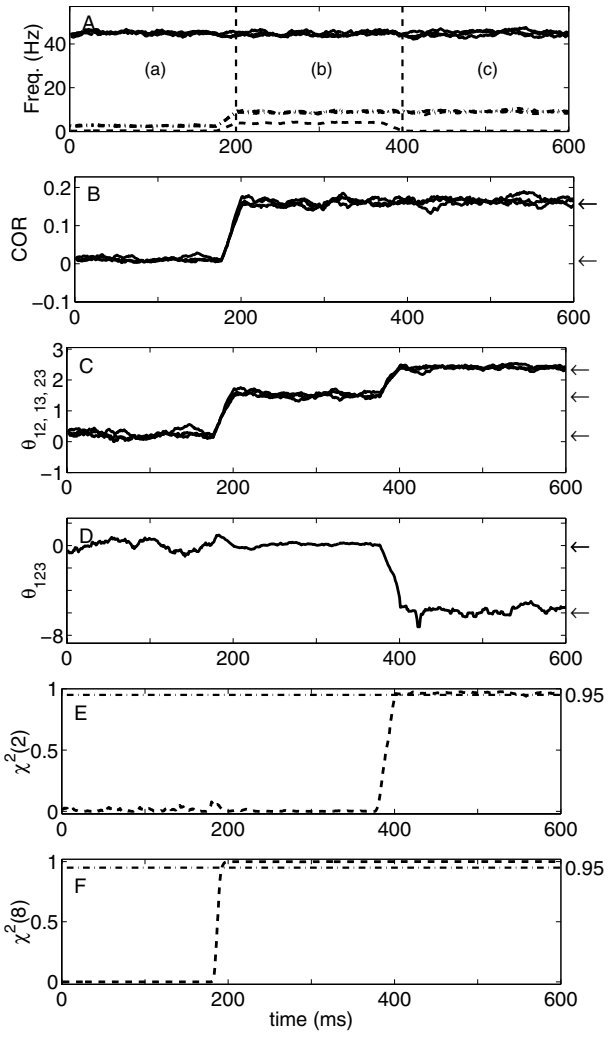
We can make the above observations more quantitative. In Figure 4E, the p-value, derived from $\chi^2(2)$, is to measure the triplewise interaction against the null hypothesis of the activity in the control period. Here we used the decomposition of k-cut $= 2$, which separates the triplewise interaction from other orders, which take the mean firing rates and the pairwise interaction together (see section 4.1). We observe that the triplewise coincident firing becomes significant only in period c.

Figure 4F indicates the p-value from $\chi^2(8)$, which is to measure the triplewise and pairwise interaction together. Here we used the decomposition of k-cut $= 1$ that separates the mean firing rates (first order) from other orders, which take the pairwise and triplewise interactions together (see section 4.2). We can see that both periods b and c now become significant.

6.2.2 Mutual Information Between Behavior and Firing. Figure 5 shows the decomposition of the MI between firing and behavior. In each of two stimulus conditions, denoted by s1 and s2, the neuron firing is assumed to be homogeneous for simplicity. Figures 5A and 5B show the mean frequency of single-neuron firing, pairwise firing, and triplewise firing with respect to s1 and s2, respectively. Period a was assumed to be the control period,

while periods b through d were assumed to be test periods. For s2, the mean frequency is the same over all periods. For s1, compared with period a, only the mean frequency of single-neuron firing is different in period b; only the mean frequency of pairwise firing is different in period c; and only the mean frequency of triplewise firing is different in period d.

The total MI is shown by the solid line in Figures 5C and 5D, while the decomposed MIs by k-cut $= 2$ and $= 1$ are shown in Figures 5C and 5D, respectively. In period b, Figure 5D indicates that most of the behavioral information is carried by modulation of single-neuron firing (the dashed



line in the figure), although some information is also carried by modulation of the other orders together (dotted line; see the legend). By inspection of period b in Figure 5C, we can further observe that the behavioral information is mostly carried by the modulation of taking single-neuron firing and pairwise firing together (the dashed line, which is almost the same as the solid line, i.e., the total MI) but is not really carried by triplewise firing (dotted line).

We can inspect periods c and d in a similar manner. In brief, Figures 5C and 5D together indicate that most behavioral information is carried by the modulation of taking pairwise and triplewise firing together in period c and is carried by the modulation of triplewise firing in period d.

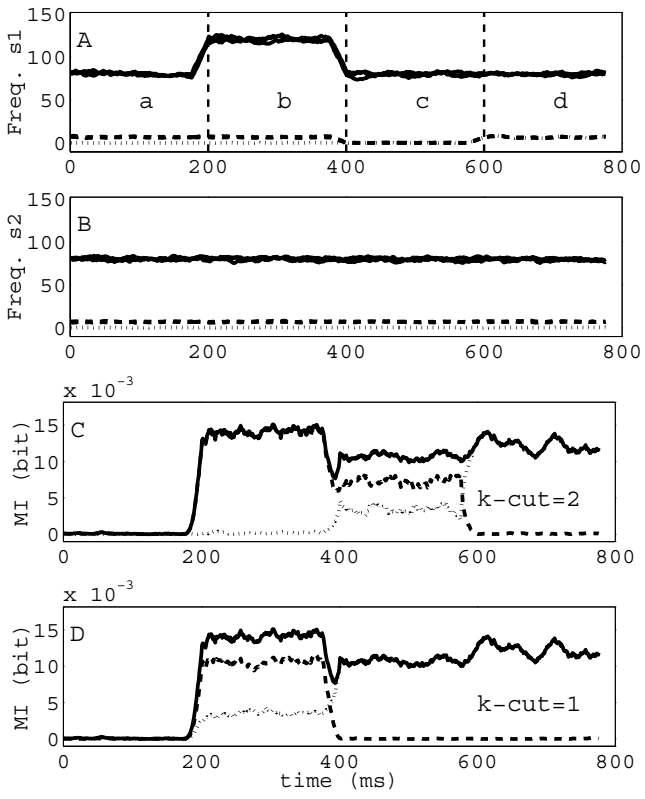
6.3 Example 3: Interesting Set of Neuron Firing. Here, we demonstrate our method to find an interesting set of neuron firing and also illustrate some practical issues in data analysis. There are two starting assumptions for this demonstration. First, the number of simultaneously recorded neurons is assumed to be 10. Second, the number of trials is assumed to be 1200 here. The number of trials is a severe limitation in real data for detecting higher-order interaction, and we consider 1200 as conceivable, or at least not impossible.

We follow the procedure set out in section 5.6 under the assumption of full homogeneous firing and ask whether any significant order of firing exists against the null hypothesis of independent firing at the questioned order. Note that it will be very difficult to investigate our data faithfully, because the assumed number of trials is 1200 and the dimension of 10 neurons is $2^{10} - 1 = 1023$. The full homogeneous assumption reduces the dimension of n neuron firing from $(2^n - 1)$ to n dimension so that it circumvents the undersampling problem.

Figure 4: *Facing page*. Example of a three-neuron case to detect the significant triplewise interaction. The spikes of three neurons were generated and estimated in a similar manner to Figure 2. (A) The η coordinates. Since we treated a homogeneous case here for simplicity, $\boldsymbol{\eta} = (\eta_i, \eta_{ij}, \eta_{ijk})$ is shown from top to bottom, being superimposed with the same order of η coordinates. (B) Correlation coefficients of three pairs of neural firing are shown, being superimposed. (C) The second order, $\theta_{ij} = \{\theta_{12}, \theta_{13}, \theta_{23}\}$. (D) The third order, θ_{123}. (E) The p-value, from $\chi^2(2)$, to indicate triplewise coincident firing against the null hypothesis of the average activity in the control period. (F) The p-value, from $\chi^2(8)$, to indicate pairwise and triplewise coincident firing together against the null hypothesis of the average activity in the control period. As for spike data generation, the number of trials is set as 2000, and the spike probability is assumed to be homogeneous in each period; $\boldsymbol{\eta} = (\eta_i, \eta_{ij}, \eta_{ijk}) = (0.0450, 0.00258, 0.00020)$ in period a, $\boldsymbol{\eta} = (0.0450, 0.0090, 0.0040)$ in period b, and $\boldsymbol{\eta} = (0.0449, 0.0090, 0.0001)$ in period c. The arrows at the right-hand side in *B*, *C*, and *D* indicate true values (see the Figure 1 legend).

We specified probability distributions over three periods of a trial to generate spike data (see the figure legend) and call them seed probabilities for convenience. Figure 6A indicates the estimated mean firing frequency of a single neuron (i.e., the first-order mean firing frequency) given generated sampled data, and Figure 6B indicates the estimated mean firing frequencies from the second to the fifth order, which appear from top to bottom in the figure. Period a is assumed to be the control period, in which the seed probability is independent firing. While the first-order mean firing frequency is the same in periods b and c as in period a, the mean firing frequencies in higher order are somewhat different, yet do not look so different in Figure 6B. We will see, however, that its intrinsic probabilistic structure is significantly different.

We can compute the exact p-values in $\chi^2(1)$ from the seed probabilities (see equation 5.14), some of which (only the relevant ones) are shown in Figure 6C. No significant order exists in period a, because the firing in that period is independent. In period b, the p-value of the fourth-order firing exceeds the 0.95 significance level. In period c, while the p-value of the fourth order drops far below, the p-value of the tenth order exceeds the significance level, and the p-value of the seventh order stays around the significance level.



In Figure 6D, we show corresponding p-values estimated from the sample data for the fourth and seventh orders. The indicated significant periods in both orders (see Figure 6D) overall follow the exact values (see Figure 6C). Although the tenth-order firing should be significant (as in Figure 6C), theoretically at least, it could not be observed in our sampled data due to a sampling problem. Indeed, all of the eighth-, ninth-, and tenth-order firings could not be observed in our sampled data (i.e., $\hat{P}_8^{(10)} = \hat{P}_9^{(10)} = \hat{P}_{10}^{(10)} = 0$), and we started our procedure from the seventh order in the sampled data. This kind of situation will most likely be encountered in real data analysis and relates to the sampling problem even under the full homogeneous assumption. This example also indicates that the significant coincident firing, even if it exists, may not be detected due to the sampling problem. Finally, the other orders do not reach the significant level by both exact and estimated values (results not shown).

7 Discussion

In this study, we investigated the nature of an information-geometric measure and its application to spike data analysis. By using the dual orthogonality of the natural and expectation parameters, we have shown that we can systematically investigate neuronal firing patterns, considering not only the second-order but also higher-order interactions, and we provided a method of hypothesis testing. For this purpose, we used the log-linear model (e.g., equation 3.4 for the two-neuron case and equation 4.2 for the three-neuron case). In this aspect, our approach shares some features with previous work (Martignon et

Figure 5: *Facing page*. Example of a three-neuron case to obtain mutual information (MI) between firing and behavior and its decomposition. The spikes of three neurons were generated and estimated in a similar manner to Figure 2. (A, B) Mean firing frequency of the three neurons with respect to s1 (A) and s2 (B). In each figure, the mean firings of a single neuron, of the pairwise coincident firing, and of the triplewise coincident firing are indicated from top to bottom. (C) MI and its decomposition by k-cut $= 2$. The total MI is indicated by a solid line. The two decomposed MIs, one by modulation of taking the mean firing rate and the pairwise interaction together and the other by modulation of the triplewise interaction, are indicated by dashed and dotted lines, respectively. (D) MI and its decomposition by k-cut $= 1$. The total MI is indicated by a solid line. The two decomposed MIs, one by modulation of the mean firing rate and the other by modulation of taking the pairwise and triplewise interactions together, are indicated by dashed and dotted lines, respectively. As for spike data generation, the number of trials per stimulus was 1000, and in each stimulus, the spike probability is assumed to be homogeneous. For s2, assumed probabilities are $(\eta_i, \eta_{ij}, \eta_{ijk}) = (0.08, 0.00704, 0.00041)$ over all the periods, a–d. For s1, assumed probabilities are $(\eta_i, \eta_{ij}, \eta_{ijk}) = (0.08, 0.00704, 0.00041)$ in period a. Compared with these probabilities, η_i changed to 0.12 in period b, η_{ij} changed to 0.00049 in period c, and η_{ijk} changed to 0.00700 in period d.

al., 1995, 2000; Del Prete & Martingon, 1998; Deco, Martignon, & Laskey, 1998),but our study explicitly uses the above orthogonality so that the methods become more transparent, more systematic, and easier. As a reference, we simply mention that the method we propose here can be related to a maximum entropy principle (Jaynes, 1982) and to testing goodness of fit to a maximum entropy distribution with some constraints, which is also related to a derivation of different types of free energies in statistical physics.

In the case of two neurons, we first showed that our method is able to test whether any pairwise correlation in one period is significantly different from that in another period, where the correlation, as the null hypothesis, is not necessarily zero. Second, the method is shown to be able to relate behavior directly with neuronal firing, using their mutual information (MI). The MI is decomposed into two types of information, conveyed by the mean firing rate and coincident firing, respectively. Third, the method is extended to the three neurons' case, where we described the details and went one step further to the general case of n neurons, where we proposed several approaches to meet practical concerns (also see Gütig, Rotter, & Aertsen, 2001). Fourth, we demonstrated the merits of the method with artificial data.

The notion of the third-order and higher-order interactions may be difficult to understand at a glance. For example, even with three neurons, investigation of only the pairwise interaction cannot fully determine their interaction, whatever measures are used. In addition, when we question whether the three neurons bind features of a single object together (Singer & Gray, 1995) or fire together, it seems more natural to seek whether triplewise interaction exists. Hence, to question functions of even only the three neurons' interaction, it is more conclusive and more robust to investigate both pairwise and triplewise interactions. Thus, to explore the functions of neural interaction in general, we consider that a method of analysis that can fully investigate any order interaction in relation to behavior is required, with the method as simple, flexible, and systematic as possible to use. It was a motivation of this study.

This study discussed coincident firing, or simultaneous firing, that is, $t_i = t_j$ (for any i, j) in $X = (X_1(t_1), \ldots, X_i(t_i), \ldots, X_n(t_n))$. This was only for simplicity of presentation. Our method is applicable to a more general case—for any set of $\{t_i\}_{i=1}^n$. It is then intriguing to ask how many specific firing patterns, or synfire chains, become significant and how much behavioral information these significant firing patterns carry (Richmond et al., 1990; McClurkin & Optican, 1996; McClurkin, Zarbock, & Optican, 1996; Abbott et al., 1996; Oram et al., 1999, 2001; Baker and Lemon, 2000). We are also pursuing this question using experimental data from the prefrontal and dorsal extrastriate visual cortices (Nakahara et al., 2001).

This study represented spike firing patterns by a binary vector; in other words, we dissected continuous time into short time bins. Obviously, we should be careful about how much information we lose by this dissection (Grün, Diesmann, Grammont, Riehle, & Aertsen, 1999) or by the use of a short time window (Panzeri, Schultz, et al., 1999; Panzeri, Treves, et al., 1999; Panzeri & Schultz, 2001). The later studies use a Taylor series expansion of MI with respect to time. In contrast, the decomposition of different order interactions and MI in

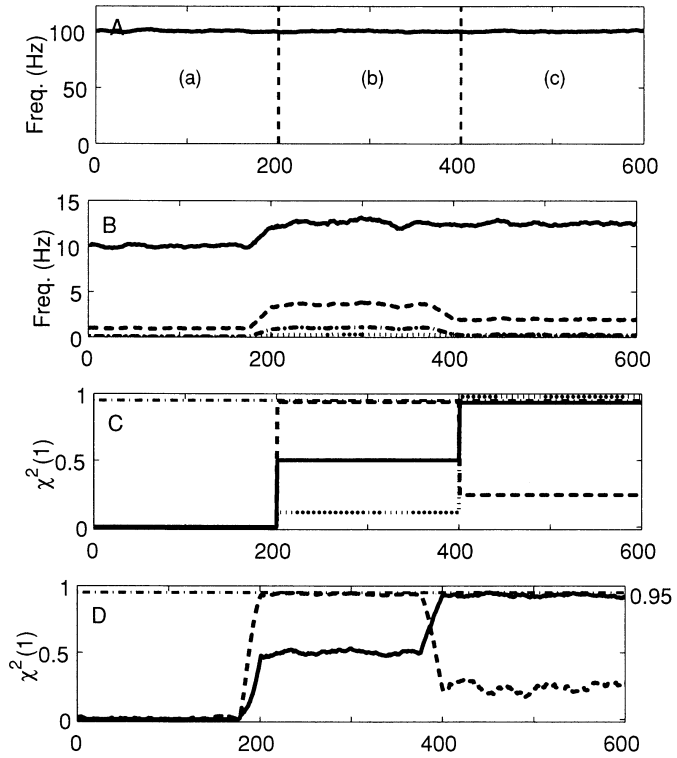


Figure 6: Example of a 10-neuron case to find an interesting set of neurons. The spikes of 10 neurons were generated and estimated in a similar manner to Figure 2. The number of trials was set as 1200. The spike probabilities are assumed to be homogeneous so that we define only 10-dimensional coordinates, which we indicate here by η coordinates. In all periods (a–c), η_1 is fixed as 0.010. In period a, the firing is assumed to be independent, that is, $\eta_k = \eta_1^k$ $(k = 2, \ldots, 10)$. In period b, $\eta_k = 0.0125 \times (0.285)^{k-2}$ $(k = 2, \ldots, 10)$. In period c, $\eta_2 = 0.0125$, $\eta_3 = 0.0020$, and $\eta_k = 0.0003 \times (0.5)^{k-4}$ $(k = 4, \ldots, \ldots, 10)$, to which there are very small adjustments. (A) Mean firing frequency (of the first order). (B) Mean firing frequencies of the second, third, fourth, and fifth orders, which appear from top to bottom. (C) Theoretically expected p-values in $\chi^2(1)$ for the fourth (indicated by a dashed line), seventh (a solid line), and tenth orders (dotted line). (D) Estimated p-values for the fourth (dashed line) and seventh orders (by solid line).

our study is exact given a fixed time bin width, regardless of its size. Mathematically speaking, coincident firing depends on time bin width. The effect of a given bin width in analyzing data is an important subject but not addressed in this study. In addition, we did not address bias correction of the estimates. MI is known to be overestimated given the limited number of sam-

ples. A correcting procedure is needed to estimate MI in our method, too. We can use, in principle, the previously proposed procedure (Optican, Gawne, Richmond, & Joseph, 1991; Kjaer, Hertz, & Richmond, 1994; Treves & Panzeri, 1995; Golomb, Hertz, Panzeri, Treves, & Richmond, 1997; Panzeri & Treves, 1996).

We discussed the method only in cases where no event has zero probability (i.e., $p_A \neq 0$), for which our theory is mathematically rigorously valid. However, in cases where some event has zero probability, we need some care. Since the dimension of binary vector increases rapidly, this issue may be a concern. There are at least two cases. In the first case, where we know a certain firing structure (e.g., for some specific sets of neurons), we can incorporate its knowledge in our estimation. Hence, if we know a priori that some probabilities are zero, the method is applicable with a little modification. In the second case where some zero probabilities are found in estimates due to a limited number of samples, we emphasize that it is a limitation of data but not of the method of analysis. For example, we cannot estimate all 1023 coordinate components of 10 neurons given only 800 samples. Yet we may still want to overcome such a situation in some ways. As one solution, we discussed using the assumption of homogeneous firing. Most studies in literature simply ignore the higher-order interaction, say, triplewise and higher, and focus on analyzing the first-order and pairwise interaction. This corresponds to assuming $\theta_k = 0$ ($k \geq 3$) (i.e., a partial homogeneous assumption). Certainly, the method can be used with such an assumption. Furthermore, although the decomposition was shown only for the mixed coordinates of a type $\zeta_k = (\eta_{k-}; \theta_{k+})$ in our study, the decomposition property has more generality. Any combination of subsets η_A and θ_B that spans the probability space of S_n can be used similarly. Extension of the method from binary to k-discrete random variable vectors is also possible. In addition, we note that other approaches are also available, including some bootstrap methods and a Bayesian framework (Del Prete & Martingon, 1998; Martignon et al., 2000).

Finally, our measure is model free, so it can be used without any assumption of the underlying neuron models and their connections. At the same time, even if the significant events were detected by the measure, the measure itself cannot determine the underlying neural interaction. This is a different issue but important in its own right, and we are interested in it too (Amari, Nakahara, Wu, & Sakai, in press). Overall, the method presented in the article is just a measure. Its ultimate test is whether it will lead to exciting findings in neuroscience (Nakahara et al., 2001).

Acknowledgments

We thank the anonymous reviewers for their helpful comments. H. N. is grateful to K. Siu and K. Kobayashi for technical assistance; M. Tatsuno for technical assistance and his corrections of some proofs; T. Poggio, Y. Kubuta, R. Gütig,

K. Anderson, E. Miller, and I. Kitori for their comments. H. N. is supported by Grants-in-Aid for Scientific Research on Priority Areas (C) of the Ministry of Education, Japan.

References

Abbott, L. F., Rolls, E. T., & Tovee, M. J. (1996). Representational capacity of face coding in monkeys. *Cerebral Cortex, 6*(3), 498–505.

Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.

Abeles, M., Bergman, H., Margalit, E., & Vaadia, E. (1993). Spatiotemporal firing patterns in the frontal cortex of behaving monkeys. *Journal of Neurophysiology, 70*(4), 1629–1638.

Abeles, M., & Gerstein, G. L. (1988). Detecting spatiotemporal firing patterns among simultaneously recorded single neurons. *Journal of Neurophysiology, 60*(3), 909–924.

Aertsen, A., & Arndt, M. (1993). Response synchronization in the visual cortex. *Current Opinion in Neurobiology, 3*(4), 586–594.

Aertsen, A. M. H. J., Gerstein, G. L., Habib, M. K., & Palm, G. (1989). Dynamics of neuronal firing correlation: Modulation of "effective connectivity." *Journal of Neurophysiology, 61*(5), 900–917.

Amari, S. (1982). Differential geometry of curved exponential families—curvature and information loss. *Annals of Statistics, 10*, 357–385.

Amari, S. (1985). *Differential geometrical methods in statistics*. Berlin: Springer-Verlag.

Amari, S. (1987). Differential geometry of a parametric family of invertible linear systems—Riemannian metric, dual affine connections and divergence. *Mathematical Systems Theory, 20*, 53–82.

Amari, S. (2001). Information geometry on hierarchical decomposition of stochastic interactions. *IEEE Transaction on Information Theory, 47*, 1701–1711.

Amari, S., & Han, T. S. (1989). Statistical inference under multi-terminal rate restrictions—a differential geometrical approach. *IEEE Transaction on Information Theory, IT-35*, 217–227.

Amari, S., Kurata, K., & Nagaoka, H. (1992). Information geometry of Boltzmann machines. *IEEE Transaction on Neural Networks, 3*(2), 260–271.

Amari, S., & Nagaoka, H. (2000). *Methods of information geometry*. New York: Oxford University Press.

Amari, S., Nakahara, H., Wu, S., & Sakai, Y. (in press). Synfiring and higher-order interactions in neuron pool.

Baker, S. N., & Lemon, R. N. (2000). Precise spatiotemporal repeating patterns in monkey primary and supplementary motor areas occur at chance levels. *Journal of Neurophysiology, 84*, 1770–1780.

Barndorff-Nielsen, O. (1978). *Information and exponential families in statistical theory*. New York: Wiley.

Bialek, W., Rieke, F., de Ruyter van Steveninck, R. R., & Warland, D. (1991). Reading a neural code. *Science, 252*(5014), 1854–1857.

Bohte, S. M., Spekreijse, H., & Roelfsema, P. R. (2000). The effects of pair-wise and higher order correlations on the firing rate of a postsynaptic neuron. *Neural Computation, 12*(1), 153–179.

Brenner, N., Strong, S., Koberle, R., Bialek, W., & de Ruyter van Steveninck, R. (2000). Synergy in a neural code. *Neural Computation, 12*(7), 1531–1552.

Cox, D. R., & Hinkley, D. V. (1974). *Theoretical statistics*. London: CRC Press.

Csiszár, I. (1967). On topological properties of f-divergence. *Studia Scientiarum Mathematicaram Hungarica, 2*, 329–339.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Annals of Probability, 3*, 146–158.

Deadwyler, S. A., & Hampson, R. E. (1997). The significance of neural ensemble codes during behavior and cognition. *Annual Review of Neuroscience, 20*, 217–244.

Deco, G., Martignon, L., & Laskey, K. B. (1998). Comparing different measures of spatio-temporal patterns in neural activity. In *ICANN98* (pp. 943–948). Skövde, Sweden.

Del Prete, V., & Martingon, L. (1998). Methods to estimate couplings and correlations in the activity of ensembles of neurons. SISSA preprint, 115/98.

Engel, A. K., König, P., Kreiter, A. K., Schillen, T. B., & Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neuroscience, 15*(6), 218–226.

Gawne, T. J., & Richmond, B. J. (1993). How independent are the messages carried by adjacent inferior temporal cortical neurons? *Journal of Neuroscience, 13*(7), 2758–2771.

Georgopoulos, A. P., Schwartz, A. B., & Kettner, R. E. (1986). Neuronal population coding of movement direction. *Science, 233*(4771), 1416–1419.

Gerstein, G. L., & Aertsen, A. M. (1985). Representation of cooperative firing activity among simultaneously recorded neurons. *Journal of Neurophysiology, 54*(6), 1513–1528.

Gerstein, G. L., Bedenbaugh, P., & Aertsen, A. M. (1989). Neuronal assemblies. *IEEE Transaction on Biomedical Engineering, 36*(1), 4–14.

Gochin, P. M., Colombo, M., Dorfman, G. A., Gerstein, G. L., & Gross, C. G. (1994). Neural ensemble coding in inferior temporal cortex. *Journal of Neurophysiology, 71*(6), 2325–2337.

Golomb, D., Hertz, J., Panzeri, S., Treves, A., & Richmond, B. (1997). How well can we estimate the information carried in neuronal responses from limited samples? *Neural Computation, 9*(3), 649–665.

Grün, S. (1996). *Unitary joint-events in multiple-neuron spiking activity: Detection, significance, and interpretation*. Thun, Frankfurt on Main: Verlag Harri Deutsch.

Grün, S., & Diesmann, M. (2000). Evaluation of higher-order coincidences in multiple parallel processes. *Society for Neuroscience Abstracts, 26*, 828.3.

Grün, S., Diesmann, M., & Aertsen, A. (2002a). "Unitary events" in multiple single-neuron spiking activity: I: Detection and significance. *Neural Computation, 14*(1), 43–80.

Grün, S., Diesmann, M., & Aertsen, A. (2002b). "Unitary events" in multiple single-neuron spiking activity: II: Nonstationary data. *Neural Computation, 14*(1), 81–120.

Grün, S., Diesmann, M., Grammont, F., Riehle, A., & Aertsen, A. (1999). Detecting unitary events without discretization of time. *Journal of Neuroscience Methods, 94*(1), 67–79.

Gütig, R., Aertsen, A., & Rotter, S. (2002). Statistical significance of coincident spikes: Count-based versus rate-based statistics. *Neural Computation, 14*(1), 121–154.

Gütig, R., Rotter, S., & Aertsen, A. (2001). Analysis of higher-order neuronal interactions based on conditional inference. Manuscript submitted for publication.

Hikosaka, O., & Wurtz, R. (1983). Visual and oculomotor functions of monkey substantia nigra pars reticulata. I. Relation of visual and auditory responses to saccades. *Journal of Neurophysiology, 49*, 1230–1253.

Ito, H., & Tsuji, S. (2000). Model dependence in quantification of spike interdependence by joint peri-stimulus time histogram. *Neural Computation, 12*, 195–217.

Jaynes, E. T. (1982). On the rationale of maximum entropy methods. *Proc. IEEE, 70*, 939–952.

Kitazawa, S., Kimura, T., & Yin, P. B. (1998). Cerebellar complex spikes encode both destinations and errors in arm movements. *Nature, 392*(6675), 494–497.

Kjaer, T. W., Hertz, J. A., & Richmond, B. J. (1994). Decoding cortical neuronal signals: Network models, information estimation, and spatial tuning. *Journal of Computational Neuroscience, 1*, 109–139.

Kudrimoti, H. S., Barnes, C. A., & McNaughton, B. L. (1999). Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and EEG dynamics. *Journal of Neuroscience, 19*(10), 4090–4101.

Laubach, M., Wessberg, J., & Nicolelis, M. A. L. (2000). Cortical ensemble activity increasingly predicts behavior outcomes during learning of a motor task. *Nature, 405*, 567–571.

Lehmann, E. L. (1983). *Theory of point estimation*. London: Chapman & Hall.

Lisman, J. E. (1997). Bursts as a unit of neural information: Making unreliable synapses reliable. *Trends in Neuroscience, 20*(1), 38–43.

MacLeod, K., Bäcker, A., & Laurent, G. (1998). Who reads temporal information contained across synchronized and oscillatory spike trains? *Nature, 395*, 693–698.

Martignon, L., Deco, G., Laskey, K., Diamond, M., Freiwald, W. A., & Vaadia, E. (2000). Neural coding: Higher-order temporal patterns in the neurostatistics of cell assemblies. *Neural Computation, 12*(11), 2621–2653.

Martignon, L., Von Hasseln, H., Grun, S., Aertsen, A., & Palm, G. (1995). Detecting higher-order interactions among the spiking events in a group of neurons. *Biological Cybernetics, 73*(1), 69–81.

Maynard, E. M., Hatsopoulos, N. G., Ojakangas, C. L., Acuna, B. D., Sanes, J. N., Normann, R. A., & Donoghue, J. P. (1999). Neuronal interactions improve cortical population coding of movement direction. *Journal of Neuroscience, 19*(18), 8083–8093.

McClurkin, J. W., Gawne, T. J., Optican, L. M., & Richmond, B. J. (1991). Lateral geniculate neurons in behaving primates. II. Encoding of visual information

in the temporal shape of the response. *Journal of Neurophysiology, 66*(3), 794–808.

McClurkin, J. W., & Optican, L. M. (1996). Primate striate and prestriate cortical neurons during discrimination. I. Simultaneous temporal encoding of information about color and pattern. *Journal of Neurophysiology, 75*(1), 481–495.

McClurkin, J. W., Zarbock, J. A., & Optican, L. M. (1996). Primate striate and prestriate cortical neurons during discrimination. II. Separable temporal codes for color and pattern. *Journal of Neurophysiology, 75*(1), 496–507.

Nadasdy, Z., Hirase, H., Czurko, A., Csicsvari, J., & Buzsaki, G. (1999). Replay and time compression of recurring spike sequences in the hippocampus. *Journal of Neuroscience, 19*(21), 9497–9507.

Nagaoka, H., & Amari, S. (1982). *Differential geometry of smooth families of probability distributions* (Tech. Rep.). Tokyo: University of Tokyo.

Nakahara, H., & Amari, S. (2002). Information-geometric decomposition in spike analysis. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems, 14*. Cambridge, MA: MIT Press.

Nakahara, H., Amari, S., Tatsuno, M., Kang, S., & Kobayashi, K. (2002). *Examples of applications of information geometric measure to neural data* (Tech. Rep. No. 02-1). Available on-line: www.mns.brain.riken.go.jp/nakahara/papers/TR_IGspike.ps, TR_IGspike.pdf.

Nakahara, H., Amari, S., Tatsuno, M., Kang, S., Kobayashi, K., Anderson, K. C., Miller, E. K., & Poggio, T. (2001). Information geometric measures for spike firing. *Society for Neuroscience Abstracts, 27*, 821.46.

Nawrot, M., Aertsen, A., & Rotter, S. (1999). Single-trial estimation of neuronal firing rates: From single-neuron spike trains to population activity. *Journal of Neuroscience Methods, 94*(1), 81–92.

Nicolelis, M. A. L., Ghazanfar, A. A., Faggin, B. M., Votaw, S., & Oliveira, L. M. O. 1997. Reconstructing the engram: Simultaneous, multisite, many single neuron recordings. *Neuron, 18*, 529–537.

Optican, L. M., Gawne, T. J., Richmond, B. J., & Joseph, P. J. (1991). Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biological Cybernetics, 65*(5), 305–310.

Optican, L. M., & Richmond, B. J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *Journal of Neurophysiology, 57*(1), 162–178.

Oram, M. W., Hatsopoulos, N. G., Richmond, B. J., & Donoghue, J. P. (2001). Excess synchrony in motor cortical neurons provides redundant direction information with that from coarse temporal measures. *Journal of Neurophysiology, 86*(4), 1700–1716.

Oram, M. W., Wiener, M. C., Lestienne, R., & Richmond, B. J. (1999). Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *Journal of Neurophysiology, 81*(6), 3021–3033.

Palm, G. (1981). Evidence, information, and surprise. *Biological Cybernetics, 42*, 57–68.

Palm, G., Aertsen, A. M., & Gerstein, G. L. (1988). On the significance of correlations among neuronal spike trains. *Biological Cybernetics, 59*(1), 1–11.

Panzeri, S., & Schultz, S. R. (2001). A unified approach to the study of temporal, correlational, and rate coding. *Neural Computation, 13*(6), 1311–1349.

Panzeri, S., Schultz, S. R., Treves, A., & Rolls, E. T. (1999). Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London Series B; Biological Science, 266*(1423), 1001–1012.

Panzeri, S., & Treves, A. (1996). Analytical estimates of limited sampling biases in different information measures. *Network, 7*, 87–107.

Panzeri, S., Treves, A., Schultz, S., & Rolls, E. T. (1999). On decoding the responses of a population of neurons from short time windows. *Neural Computation, 11*(7), 1553–1577.

Parker, A. J., & Newsome, W. T. (1998). Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience, 21*, 227–277.

Pauluis, Q., & Baker, S. N. (2000). An accurate measure of the instantaneous discharge probability, with application to unitary joint-even analysis. *Neural Computation, 12*(3), 647–669.

Perkel, D. H., Gerstein, G. L., & Moore, G. P. (1967). Neuronal spike trains and stochastic point processes: II: Simultaneous spike trains. *Biophysics Journal, 7*, 419–440.

Prut, Y., Vaadia, E., Bergman, H., Haalman, I., Slovin, H., & Abeles, M. (1998). Spatiotemporal structure of cortical activity: Properties and behavioral relevance. *Journal of Neurophysiology, 79*(6), 2857–2874.

Rao, C. R. (1945). Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of Calcutta Math. Soc., 37*, 81–91.

Reinagel, P., & Reid, C. R. (2000). Temporal coding of visual information in the thalamus. *Journal of Neuroscience, 20*, 5392–5400.

Richmond, B. J. & Gawne, T. J. (1998). The relationship between neuronal codes and cortical organization. In H. B. Eichebaum, J. L. Davis, & H. Eichenbaum (Eds.), *Neuronal ensembles strategies for recording and decoding* (pp. 57–79). New York: Wiley-Liss.

Richmond, B. J., Optican, L. M., & Spitzer, H. (1990). Temporal encoding of two-dimensional patterns by single units in primate primary visual cortex. I. Stimulus-response relations. *Journal of Neurophysiology, 64*(2), 351–369.

Riehle, A., Grün, S., Diesmann, M., & Aertsen, A. (1997). Spike synchronization and rate modulation differentially involved in motor cortical function. *Science, 278*, 1950–1953.

Rolls, E. T., Treves, A., & Tovee, M. J. (1997). The representational capacity of the distributed encoding of information provided by populations of neurons in primate temporal visual cortex. *Experimental Brain Research, 114*(1), 149–162.

Roy, A., Steinmetz, P. N., & Niebur, E. (2000). Rate limitations of unitary event analysis. *Neural Computation, 12*(9), 2063–2082.

Salinas, E., & Sejnowski, T. J. (2001). Correlated neuronal activity and the flow of neural information. *Nature Reviews Neuroscience, 2*(8), 539–550.

Samengo, I., Montagnini, A., & Treves, A. (2000). Information-theoretical description of the representational capacity of N independent neuron. *Society for Neuroscience Abstracts, 26*, 739.1.

Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience, 18*, 555–586.

Steinmetz, P. N., Roy, A., Fitzgerald, P. J., Hsiao, S. S., Johnson, K. O., & Niebur, E. (2000). Attention modulates synchronized neuronal firing in primate somatosensory cortex. *Nature, 404*(6774), 187–190.

Stuart, A., Ord, K., & Arnold, S. (1999). *Kendall's advanced theory of statistics, Vol. 2: A classical inference and the linear model*. London: Arnold.

Sugase, Y., Yamane, S., Ueno, S., & Kawano, K. (1999). Global and fine information coded by single neurons in the temporal visual cortex. *Nature, 400*(6747), 869–873.

Tetko, I. V., & Villa, A. E. P. (1992). Fast combinatorial methods to estimate the probability of complex temporal patterns of spikes. *Biological Cybernetics, 76*, 397–407.

Tovee, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *Journal of Neurophysiology, 70*(2), 640–654.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation, 7*, 399–407.

Tsukada, M., Ishii, N., & Sato, R. (1975). Temporal pattern discrimination of impulse sequences in the computer-simulated nerve cells. *Biological Cybernetics, 17*, 19–28.

Vaadia, E., Haalman, I., Abeles, M., Bergman, H., Prut, Y., Slovin, H., & Aertsen, A. (1995). Dynamics of neuronal interactions in monkey cortex in relation to behavioural events. *Nature, 373*, 515–518.

Victor, J. D., & Purpura, K. P. (1997). Metric-space analysis of spike trains: Theory, algorithms and application. *Network: Computation in Neural Systems, 8*, 127–164.

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science, 261*(5124), 1055–1058.

Zhang, K., Ginzburg, I., McNaughton, B. L., & Sejnowski, T. J. (1998). Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. *Journal of Neurophysiology, 79*, 1017–1044.

Zohary, E., Shadlen, M. N., & Newsome, W. T. (1994). Correlated neuronal discharge rate and its implications for psychophysical performance. *Nature*, *370*(6485), 140–143.

Received September 6, 2001; accepted April 23, 2002.