LETTER **Communicated by Kechen Zhang**

Difficulty of Singularity in Population Coding

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

Fisher information has been used to analyze the accuracy of neural population coding. This works well when the Fisher information does not degenerate, but when two stimuli are presented to a population of neurons, a singular structure emerges by their mutual interactions. In this case, the Fisher information matrix degenerates, and the regularity condition ensuring the Cramér-Rao paradigm of statistics is violated. An animal shows pathological behavior in such a situation. We present a novel method of statistical analysis to understand information in population coding in which algebraic singularity plays a major role. The method elucidates the nature of the pathological case by calculating the Fisher information. We then suggest that synchronous firing can resolve singularity and show a method of analyzing the binding problem in terms of the Fisher information. Our method integrates a variety of disciplines in population coding, such as nonregular statistics, Bayesian statistics, singularity in algebraic geometry, and synchronous firing, under the theme of Fisher information.

1 Introduction

A fundamental concern in systems neuroscience is how information is represented in the brain. Population coding is a widely accepted scheme of representation in which the accuracy of encoding is evaluated by Fisher information (Seung & Sompolinsky, 1993; Zemel, Dayan, & Pouget, 1998; Yoon & Sompolinsky, 1999; Zhang & Sejnowski, 1999; Pouget, Dayan, & Zemel, 2000; Wu, Amari & Nakahara, 2002a, 2002b; Wu, Nakahara, & Amari, 2001; Nakahara & Amari, 2002; Nakahara, Wu, & Amari, 2001; Wu, Chen, Niranjan, & Amari, 2003). The Fisher information is not only related to the mean square error of an estimator by the Cramér-Rao theorem, but is an intrinstic quantity to measure how much information is included in the population coding.

Neural Computation **17, 839–858** (2005) © 2005 Massachusetts Institute of Technology

When two similar stimuli are presented at the same time, their responses interact mutually and cause strange phenomena in both the mathematical structure and the behavior of an animal. We explain this phenomenon by a simple example (Ingle, 1968; Amari & Arbib, 1977). When a frog observes a fly in its visual field, it is represented by an excitation pattern of neurons in the visual field (tectum). The pattern has a peak at the place corresponding to the location of the fly (see Figure 1A). Neural firing fluctuates randomly, so that the amount of information included in such population coding is analyzed by Fisher information. However, when two flies appear at the same time, the response pattern of neural excitation has two peaks at the corresponding locations of the flies (see Figure 1B). When the two flies are very close, the two peaks merge into one, and there remains only a noisy pattern of one peak (see Figure 1C). What will happen under such a condition?

Neuroethological study has observed pathological behavior in that the frog cannot distinguish the two flies and sometimes jumps in between the flies, failing to catch either (Ingle, 1968). Mathematically, the Fisher information matrix degenerates in this case, so that the established Cramér-Rao paradigm of statistics does not hold. Hence, a new framework of statistics is required to analyze such a nonregular or singular situation. Looking at the situation in more detail, we find that algebraic singularity exists in the underlying statistical manifold, which is the main cause of difficulty. There are many practical methods to resolve the problem. The frog has its own neural algorithm. Many theoreticians use the Bayesian strategy, and biologically plausible algorithms have been proposed (Pouget, Zhang, Denever, & Latham, 1998; Zemel, et al., 1998; Zemel & Pillow, 2002; Sahani & Dayan, 2003). However, the Bayesian method is not free of the difficulty caused by the singularity, because the performances of Bayesian algorithms are largely affected by the prior distributions in the singular case, which is completely different from the regular case. Computer simulations have been used so far to analyze the estimation errors of proposed algorithms. We need a new theretical framework to carry out the mathematical analysis, in particular the limit of accuracy attainable by the optimal method. The singularity also affects the Bayesian method strongly in a different way from the regular case. This study focuses on such a singular problem of population coding, providing a new framework of analysis. We do not propose a practical procedure to decide the number of stimuli or to decode information giving the estimated locations of stimuli. Instead, our framework is necessary to analyze the behaviors of algorithms mathematically.

For this purpose, we must integrate various disciplines related to the scheme of population coding: neural fields (Amari, 1977), degenerate Fisher information in a nonregular model (Amari & Ozeki, 2001; Fukumizu, 2003), algebraic geometrical singularity (Fukumizu & Amari, 2000; Watanabe, 2001a) that causes such degeneration, and the information-geometrical method (Amari & Nagaoka, 2000; Amari, 1998) to elucidate the whole

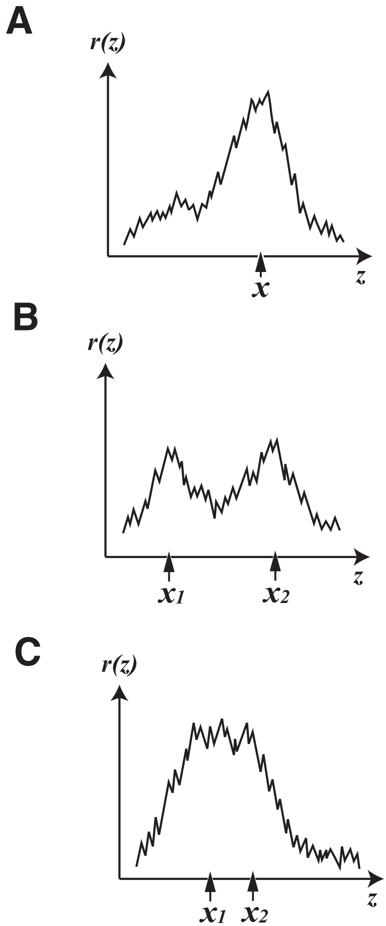

Figure 1: Schematic presentation of firing patterns $r(z)$ in a neural population, where $r(z)$ denotes the firing rate of neurons at location z: given a single stimulus at x (A), given two stimuli at x_1 and x_2 that are far enough apart (B), and given two stimuli close to each other (C).

structure. (See also Dacunha-Castelle & Gassiat, 1997; Lin & Shao, 2003 and Burnashev & Amari, 2002, for recent statistical analysis of the log-likelihood ratio and the maximum likelihood estimator in singular statistical models.)

We begin with the analysis of the Fisher information of population coding in a neural field. The Fourier domain is useful for analysis when the field is homogeneous. Various properties in population coding, such as the effects of the shape of a tuning curve and the correlation of noises, will be clearly understood in this framework (Wu et al., 2002b). We summarize the method in section 2, when a single stimulus is given in a one-dimensional neural field. When two stimuli appear close together at the same time in a one-dimensional field, their mutual interaction becomes very strong. The Fisher information is a matrix in this case, but it degenerates when the two are close to each other. This violates the classic framework of statistics, and one cannot apply the Cramér-Rao theorem because the estimation error grows without limit. It is indeed difficult to define the optimal strategy to judge if there are two targets or one. It is more difficult to identify the intensity and the location of each target and to analyze rigorously the performances of various algorithms. The difficulty is observed in animal behavior (Ingle, 1968; Treue, Hol, & Rauber, 2000; Reichardt & Poggio, 1975). When the Fisher information degenerates, its inverse diverges to infinity. Therefore, we need a careful mathematical analysis.

We demonstrate that the space of parameters to be estimated from the neural excitation pattern has algebraic singularity and analyze its structure in detail. Such singularity causes the degeneration of the Riemannian structure (or the Fisher information) of the space as a black hole does in the real space-time universe. We analyze how the accuracy of estimation degenerates as the two targets draw near to each other. To this end, the Fisher information is decomposed into regular directions and singular directions, and the accuracy of estimation in the respective directions is analyzed. Such analysis gives new insight into modern statistics.

The Bayes inference is another method to give a practical solution even in a singular case. However, we show that there exists a serious problem, because a smooth, positive prior over the parameters induces a singular prior on the singular model. A smooth, positive prior on the space of response functions is not absolutely continuous to a smooth prior on the parameter space. They are mutually singular. We study such a problem, showing that the same difficulty exists in Bayesian framework. From the viewpoint of model selection, the number of stimuli is decided from the data by some criteria such as the Akaike information criterion, the Bayesian information criterion, or the minimum descriptive length criterion. However, for all of these model selection procedures, their statistical validity loses its ground in such a singular case, as we explain below.

We then ask if there is a way to get rid of this difficulty, finding that this is closely related to the binding problem. We demonstrate how synchronous firing resolves this difficulty by using the Fisher information. The difficulty is remarkably relaxed by synfiring, which resolves singularity. The synfiring scheme has been studied extensively, but this is the first time it is analyzed in terms of the Fisher information. We suggest a new method of analyzing a general binding problem in terms of the Fisher information in the framework of information geometry (Amari & Nagaoka, 2000).

2 Fisher Information for a Single Stimulus

We recapitulate the population coding in a neural field, summarizing the analysis of the Fisher information for a single stimulus. Details are found in the work of Wu et al. (2001). Consider a population of neurons located on a line, which we call a one-dimensional neural field, and the position of a neuron is denoted by a coordinate z. Let $r(z)$ be the firing rate of the neurons at around position z. Neurons are excited in response to a stimulus applied from the outside. We consider a set of possible stimuli of which the features are specified by one-dimensional coordinates x. The feature may be the orientation of a bar, the position of an object, or something else. There is a natural one-to-one correspondence between x and z, and a stimulus with feature x is encoded by an activity pattern $r(z)$ over the field such that the neurons at positions $z \approx x$ are excited strongly, with $r(z)$ having a peak at $z \approx x$ (see Figure 1).

More precisely, let $f(z)$ be a unimodal function having a peak at $z = 0$, and when stimulus x is applied, the expected firing rate of the neurons at around z is given by $f(z - x)$. The function $f(z)$ is called the tuning curve, and we use the gaussian function φ for f,

$$\varphi(z) = \frac{1}{\sqrt{2\pi}a} e^{-z^2/2a^2}, \tag{2.1}$$

for the ease of analysis, where a represents the width of the tuning function.

Neuronal firing is stochastic in its nature. Given stimulus x, the observed firing pattern $r(z; x)$ is given by

$$r(z; x) = f(z - x) + \sigma\varepsilon(z), \tag{2.2}$$

where $\varepsilon(z)$ denotes random fluctuations with intensity σ. We assume that $\varepsilon(z)$ is gaussian with mean 0. Firing of neurons at positions z and z' is correlated in general, so that we put

$$\langle\varepsilon(z)\rangle = 0, \tag{2.3}$$

$$\left\langle\varepsilon(z)\varepsilon\left(z'\right)\right\rangle = h\left(z - z'\right), \tag{2.4}$$

where $< \quad >$ denotes expectation and h denotes the covariance function. Here, $h(z - z')$ decreases to 0 as $\left|z - z'\right|$ becomes larger. We assume the following form of the covariance function $h(z)$,

$$h(z) = (1 - \beta)\delta(z) + \beta e^{-z^2/2b^2}, \tag{2.5}$$

where $\delta(z)$ is the delta function, $1 - \beta$ represents the intensity of independent fluctuation proper to each neuron, and b represents the effective range of correlated fluctuation. This form captures a variety of the correlation structure. Our previous articles (Wu et al., 2001, 2002a, 2002b) investigated the properties of the Fisher information for the single stimulus. Although the above $h(z)$ does not include a multiplicative noise structure, it is possible to address such a case with a little modification (Yoon & Sompolinsky, 1999; Eurich & Wilke, 2000; Nakahara & Amari, 2002; Wu, Amari, & Nakahara, 2004), so that for simplicity we use only the above form in this study.

The probability distribution of firing pattern $\mathbf{r} = \{r(z)\}$ given x is then written explicitly as

$$Q(\mathbf{r}|x) = \frac{1}{Z} \exp\left\{ -\frac{1}{2\sigma^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} [r(z) - f(z-x)] h^*(z-z') [r(z') - f(z'-x)] dz dz' \right\}, \tag{2.6}$$

where Z is the normalization factor. The function $h^*(z)$ is the inverse kernel of the correlation function h, defined by

$$\int_{-\infty}^{\infty} h^*(z-z') h(z'-z'') dz' = \delta(z-z''),$$

where δ is the delta function. The Fisher information for the single stimulus x is given by

$$I_F(x) = E\left[-\frac{d^2 \ln Q(\mathbf{r}|x)}{dx^2} \right] - \int Q(\mathbf{r}|x) \frac{d^2 \ln Q(\mathbf{r}|x)}{dx^2} d\mathbf{r}, \tag{2.7}$$

where E denotes expectation. Simple calculation yields (Wu et al., 2001)

$$I_F = \frac{1}{\sigma^2} \int f'(z-x) h^*\left(z-z'\right) f'\left(z'-x\right) dz dz', \tag{2.8}$$

which does not depend on x because of the translation invariance of the neural field. By using the Fourier transform, I_F is given by

$$I_F = \frac{1}{2\pi\sigma^2} \int_{-\infty}^{\infty} \frac{\omega^2 \exp\left(-a^2\omega^2\right)}{H(\omega)} d\omega, \tag{2.9}$$

where $H(\omega)$ is the Fourier transform of $h(z)$,

$$H(\omega) = (1-\beta) + \beta\sqrt{2\pi} b \exp\left(-\frac{b^2\omega^2}{2}\right). \tag{2.10}$$

This form elucidates various characteristics of the Fisher information for a single stimulus (Wu et al., 2001).

3 Fisher Information for Two Stimuli

Fisher information is useful for evaluating the accuracy of population coding. The Cramér-Rao theorem shows that ${I_F}^{-1}$ provides a lower bound of the mean square error for the estimated feature or location $\hat{x}$. The situation, changes drastically however, when two stimuli are presented at the same time. The response of neurons at z, given two stimuli at locations x_1 and x_2, is modeled by

$$r(z) = f(z; x_1, x_2) + \sigma\epsilon(z), \tag{3.1}$$

where $f(z; x_1, x_2)$ is the tuning curve when two stimuli with features x_1 and x_2 are presented. Let the intensities of the two stimuli be v and $1 - v$, where we assume that the total intensity is regulated to be equal to 1. Then the tuning curve is defined by

$$f(z; x_1, x_2, v) = (1 - v)\varphi(z - x_1) + v\varphi(z - x_2), \tag{3.2}$$

where $0 \le v \le 1$. Experimental results (Treue et al., 2000; Martinez-Trujillo & Treue, 2000) suggest this form of tuning curves. The probability distribution $Q(\mathbf{r}; x_1, x_2, v)$ of neural firing is given as equation 2.6, where $f(z)$ is now set as equation 3.2.

In this setting, we search for the Fisher information in population coding concerning two stimuli, $(1 - v, x_1)$ and (v, x_2). Since the unknown parameters (x_1, x_2, v) are three in number, the Fisher information now becomes a 3×3 matrix. When the parameters are represented by a vector $\boldsymbol{\theta} = (x_1, x_2, v)$,

$$I_F(\boldsymbol{\theta}) = -E\left[\frac{\partial^2 \ln Q(\mathbf{r}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}}\right]. \tag{3.3}$$

We can calculate the Fisher information matrix I_F in a similar way, and its inverse would give the accuracy of the estimated features of the stimuli. When $x_1 = x_2$, the two targets merge into one, and $f(z; x, x, v) = \varphi(z - x)$ does not depend on v, because of $x_1 = x_2$, that is, a single stimulus. However, when the two targets are close, $x_1 \approx x_2$, a pathological situation occurs, and it causes an extreme difficulty in estimating v, causing nonidentifiability of v. Let us first discuss this difficulty in case of $x_1 = x_2$ for ease of understanding. The Fisher information matrix indeed degenerates, and its inverse diverges. Therefore, we cannot apply the Cramér-Rao paradigm of statistics in this case. Such a pathological statistical model is said to be nonregular and has been outside the scope of the standard theory of statistics, except for some pioneering work (Weyl, 1939; Hotelling, 1939) and recent developments (Dacunha-Castelle & Gassiat, 1997; Lia & Shao, 2003; Fukumizu, 2003). The analysis in the pathological case is the main topic of this letter. We also show how the singularity affects the Bayesian method and model selection.

The pathological situation is elucidated by algebraic geometry, because algebraic singularities exist in the parameter space $M = \{(x_1, x_2, v)\}$ of two stimuli (see Figure 2A). First, when $x_1 = x_2 = x$, the tuning curve is the same whatever v is, so that the points on a vertical line,

$$l_x \ : \ x_1 = x_2 = x, \ \ v = \text{arbitrary}, \tag{3.4}$$

have the same effect on the neural field (see Figure 2A, red bold vertical line). Second, when $v = 1$ $(v = 0)$, there is no x_1 (x_2) stimulus, so that x_1 (x_2) cannot be decided. Hence, the points on the line $l_{x,1} : v = 1$, x_1 are arbitrary, and the line $l_{x,0} : v = 0$, x_2 is arbitrary (see Figure 2A, red bold horizontal lines). Thus, the points in the parameter space that have the same effect would be summarized as the same point in the set $\tilde{M}$ consisting of the resultant probability distributions. Mathematically, we divide M by the equivalence relation of the same effect. The parameter space M then collapses along those critical lines, and the resultant reduced $\tilde{M}$ has algebraic singularity, as shown in Figure 2B. The Fisher information matrix degenerates on these critical points.

We now introduce a new method of analysis, focusing on the algebraic singularity. We begin with reparameterizing the features of stimuli as

$$\begin{aligned} u &= x_2 - x_1 \\ w &= (1 - v)x_1 + vx_2. \end{aligned}$$

Here, u indicates the difference between the locations of the two stimuli, and w indicates their center of gravity. Using these, we can rewrite the tuning curve as

$$f(z; u, w, v) = (1 - v)\varphi(z - w + vu) + v\varphi(z - w - (1 - v)u). \tag{3.5}$$

With these parameters (w, u, v), we evaluate the accuracy of coding in the neighborhood of $u \approx 0$. Without loss of generality, we rescale the z-axis to let $a = 1$, where x_1, x_2, and u should be rescaled as well as the width b of the correlation. Using the Chebyshev-Hermite expansion with respect to u of equation 3.5, we get,

$$f(z; u, w, v) = \left(\sum_{n=0}^{\infty} \frac{c_n(v)}{n!} u^n H_n(z - w) \right) \varphi(z - w),$$

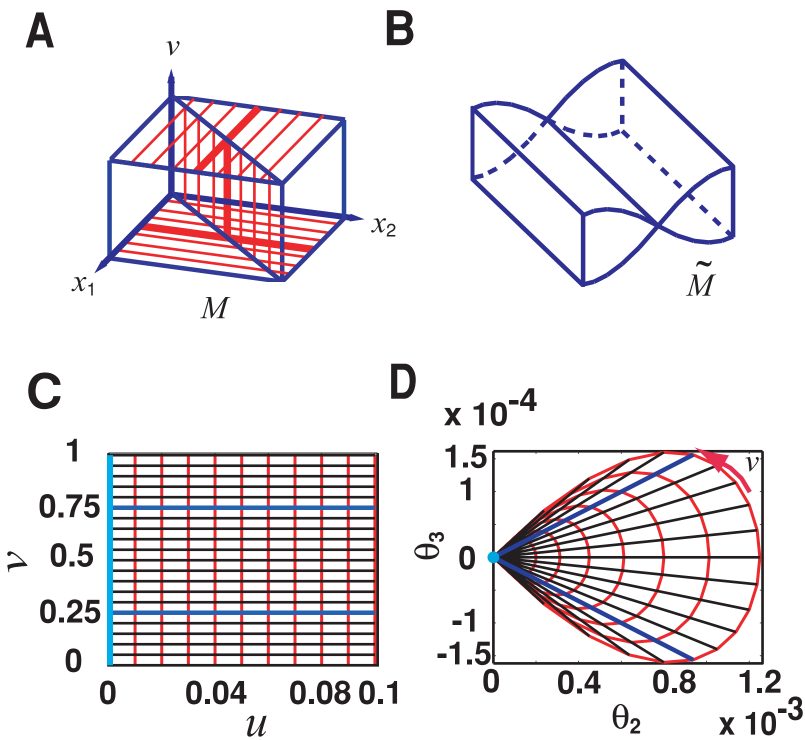

Figure 2: (A) Parameter space $M = \{(x_1, x_2, v)\}$. The critical set consists of three surfaces with red lines. The three red thicker lines correspond to a set of those parameters that are equivalent to one stimuli and give the same probability distribution. (B) The reduced space $\tilde{M}$, where the equivalent set of parameters is reduced to one point. (C) Regular (u, v)-plane, where the part $u < 0, v = 0.5 + v_0$, is equivalent to $u > 0$, $v = 0.5 - v_0$ by interchanging x_1 and x_2. (D) The (u, v)-plane embedded in regular space (θ_1, θ_2) by a cusp-type singular map. Note that the line $u = 0$ in (C) (the light blue line) is mapped to a point in D.

where $c_n(v) \equiv (1 - v)(-v)^n + v(1 - v)^n$ and $H_n(z)$ is the Chebyshev-Hermite polynomials. We approximate $f(z; u, w)$ by taking the summation up to 3, having

$$f(z; u, w, v) = \left\{1 + \frac{1}{2}c_2(v)u^2 H_2(z - w) + \frac{1}{6}c_3(v)u^3 H_3(z - w)\right\} \varphi(z - w),$$

$$c_0 = 1,\ c_1 = 0,\ c_2 = v(1 - v),\ c_3 = -v(1 - v)(2v - 1)$$

$$H_0(z) = 1,\ H_1(z) = z,\ H_2(z) = z^2 - 1,\ H_3(z) = z^3 - 3z.$$

Let us now introduce the set of new parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$ in which f is written as

$$f(z; \boldsymbol{\theta}) = \{1 + \theta_2 H_2(z - \theta_1) + \theta_3 H_3(z - \theta_1)\}\, \varphi(z - \theta_1), \tag{3.6}$$

and consider a new statistical model S parameterized by $\boldsymbol{\theta}$ in which the probability distributions are given by equation 2.6 and f is given by equation 3.6. Our original statistical model M is parameterized by $\boldsymbol{\xi} = (w, u, v)$ and is included in S by the mapping

$$\boldsymbol{\theta} = \left(w, \frac{c_2(v)u^2}{2}, \frac{c_3(v)u^3}{6}\right). \tag{3.7}$$

The mapping (3.7) from $\boldsymbol{\xi}$ to $\boldsymbol{\theta}$ is singular, because its Jacobian determinant,

$$|J| = \left|\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\xi}}\right|, \tag{3.8}$$

vanishes at $u = 0$.

We adopt the strategy of calculating the Fisher information $I_F^S(\boldsymbol{\theta})$ of S, which is regular in the coordinates $\boldsymbol{\theta}$. The singular $I_F(\boldsymbol{\xi})$ of M is given through the singular mapping equation 3.7. This method elucidates the singular structure of M. We first evaluate $I_F^S(\boldsymbol{\theta})$ at around $\theta_2 = \theta_3 = 0$, which corresponds to $u = 0$. The elements of $I_F^S(\boldsymbol{\theta}) = (I_{kl})$ are calculated as

$$I_{kl} = \frac{1}{\sigma^2} \int \partial_k f(z; \boldsymbol{\theta}) h^*\left(z - z'\right) \partial_l f\left(z'; \boldsymbol{\theta}\right) dz dz',$$

where $\partial_k f = (\partial/\partial\theta_k) f$. In terms of the Fourier transform, they are

$$I_{kl} = \frac{1}{2\pi\sigma^2} \int F_k(\omega) H^*(\omega) \overline{F_l(\omega)} d\omega, \tag{3.9}$$

where $F_k(\omega) = \partial F(\omega)/\partial\theta_k$ and $F(\omega)$ is the Fourier transform of $f(z)$. The Fisher information matrix is explicitly calculated as

$$I_F^S(\boldsymbol{\theta}) = \begin{bmatrix} g_1 & 0 & -g_2 \\ 0 & g_2 & 0 \\ -g_2 & 0 & g_3 \end{bmatrix},$$

where

$$g_k = \frac{1}{2\pi\sigma^2} \int \frac{\omega^{2k} \exp(\omega^2)}{H(\omega)} d\omega > 0.$$

For ease of further study, we modify the parameters from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}^*$,

$$\theta_1^* = \theta_1 + \frac{g_2}{g_3}\theta_3, \tag{3.10}$$

$$\theta_2^* = \theta_2, \quad \theta_3^* = \theta_3. \tag{3.11}$$

Then the Fisher information matrix is diagonalized,

$$I_F^S(\boldsymbol{\theta}^*) = \begin{bmatrix} g_1 & 0 & 0 \\ 0 & g_2 & 0 \\ 0 & 0 & \bar{g}_3 \end{bmatrix},$$

where $\bar{g}_3 = g_3 - g_2^2/g_1 > 0$.

It is now easy to obtain the Fisher information matrix $I_F(\boldsymbol{\xi})$ in terms of the parameters $\boldsymbol{\xi} = (w, u, v)$, which represent the features of the stimuli. Let J be the Jacobian of the transformation from $\boldsymbol{\xi}$ to $\boldsymbol{\theta}^*$,

$$J = \frac{\partial \boldsymbol{\theta}^*}{\partial \boldsymbol{\xi}} = \begin{bmatrix} 1 & (1/2)g_2 g_3^{-1} c_3 u^2 & (1/6)g_2 g_3^{-1} c_3' u^3 \\ 0 & c_2 u & (1/2)c_2' u^2 \\ 0 & (1/2)c_3 u^2 & (1/6)c_3' u^3 \end{bmatrix}, \tag{3.12}$$

where $c_i' = dc_i(v)/dv$. Then the Fisher information $I_F(\boldsymbol{\xi})$ is obtained from

$$I_F(\boldsymbol{\xi}) = J^T I_F^S(\boldsymbol{\theta}^*) J. \tag{3.13}$$

4 Algebraic Singularity and Accuracy of Estimation

Although the Fisher information $I_F^S(\boldsymbol{\theta}^*)$ is regular, $I_F(\boldsymbol{\xi})$ is singular at $u = 0$, because the Jacobian J of the transformation is singular at $u = 0$, $|J| = 0$. In other words, the present model M is embedded in a regular space as a singular subset $\tilde{M}$, and the singularity cannot be resolved by coordinate transformation. Let us highlight the role of the singular mapping. Since the structure of the field is homogeneous along the w-axis, we focus on the mapping between (u, v) and

$$(\theta_2^*, \theta_3^*) = \left(\frac{v(1-v)u^2}{2}, \frac{v(1-v)(1-2v)u^3}{6} \right). \tag{4.1}$$

Figure 2C shows the u-v plane, where the v-axis (indicated by a thick blue vertical line in the figure) corresponding to $u = 0$ degenerates to a single point in the $\boldsymbol{\theta}$-plane, causing the singularity. Horizontal (black) and vertical (red) lines simply indicate the grids of v and u values. Figure 2D shows the regular θ_2-θ_3 plane in which the grids of (u, v) in Figure 2C are embedded.

Note that Figure 2C corresponds to a slice vertically cut from Figure 2A. Now consider the line $v = 1/2 + v_0$ (solid horizontal blue line) in Figure 2C. It corresponds to the stimuli of intensity $1/2 + v_0$, of which distance u changes arbitrarily. Note that when $u < 0$, it corresponds to the stimuli of intensity $1/2 - v_0$, where stimulus 1 and stimulus 2 are interchanged. Hence, we consider the two blue horizontal lines as one component. The line is mapped to a cusp line (black or blue pair) in Figure 2D. The line $v_0 = 0$ corresponds to the horizontal line $\theta_3^* = 0$ in the right figure, and as v_0 increases, the angle of the corresponding cusp becomes larger. As v_0 tends to $1/2$, the cusp shrinks and reduces to the origin. This gives rise to the singularity corresponding to $v = 0$ or 1, another type of singularity, which we do not address in this article. Note that the space of $\boldsymbol{\theta}^*$ itself corresponds to a regular statistical model with a nondegenerate Fisher information matrix. The space of $\boldsymbol{\xi}$ is embedded in the $\boldsymbol{\theta}^*$ space by the mapping with the singularity of the cusp structure (see Figure 2D).

Let $\hat{\boldsymbol{\xi}} = \left(\hat{w}, \hat{u}, \hat{v}\right)$ be the maximum likelihood estimator. The errors $\Delta\hat{w} = \hat{w} - w$, $\Delta\hat{u} = \hat{u} - u$, $\Delta\hat{v} = \hat{v} - v$ are given by

$$\Delta\hat{\boldsymbol{\xi}} = \frac{\partial \boldsymbol{\xi}}{\partial \boldsymbol{\theta}^*}\Delta\hat{\boldsymbol{\theta}}^* = J^{-1}\Delta\hat{\boldsymbol{\theta}}^*. \tag{4.2}$$

The covariance matrix of the error is given by $I_F^{-1}(\boldsymbol{\xi})$, but the inversion of I_F is complicated, so we evaluate the error directly by using equation 3.13.

Since three components of $\Delta\hat{\boldsymbol{\theta}}^*$ are independently distributed, we obtain, by evaluating J^{-1} carefully, the asymptotic evaluation of the variances of the errors $\Delta\hat{\boldsymbol{\xi}}$,

$$V(\Delta\hat{w}) \approx \frac{g_3}{g_1\bar{g}_3}, \tag{4.3}$$

$$V(\Delta\hat{u}) \approx k_1 u^{-4} + k_2 u^{-2}, \tag{4.4}$$

$$V(\Delta\hat{v}) \approx k_3 u^{-6}, \tag{4.5}$$

where

$$k_1 = 36(1-2v)^2 v^{-2}(1-v)^{-2}\bar{g}_3{}^{-1}, \tag{4.6}$$

$$k_2 = 4\left(6v^2 - 6v + 1\right)^2 v^{-2}(1-v)^{-2} g_2^{-1}, \tag{4.7}$$

$$k_3 = 144\bar{g}_3^{-1}. \tag{4.8}$$

When $v = 1/2, k_1$, the coefficient of u^{-4} vanishes. The above results highlight the pathological nature of estimation:

- For w, the Fisher information is of order 1 even when $u \to 0$. Hence, one can estimate the center of gravity of two stimuli accurately.

- When $v \neq 1/2$, the variance of $\Delta\hat{u}$ diverges to infinity in the order of $1/u^4$ as $u \to 0$. The estimation of v is worse, because its variance diverges in the order of $1/u^6$. These show the difficulty in estimating them accurately.
- When $v = 1/2$, that is, the intensities of two stimuli are the same, the dominant term in equation 4.6 vanishes, so that the variance of $\Delta\hat{u}$ is of order $1/u^2$. Hence, estimation is carried out more accurately under such a situation.

Turning to the estimation of the original locations x_1 and x_2, we have

$$\hat{x}_1 = \hat{w} - \hat{v}\hat{u}, \tag{4.9}$$

$$\hat{x}_2 = \hat{w} + (1 - \hat{v})\hat{u}, \tag{4.10}$$

so that terms of $O\left(1/u^6\right)$ dominate in their error variances. In conclusion, one can estimate the center of gravity accurately, but it is very difficult to estimate their intensities and difference of locations, or the respective locations x_1 and x_2, as u tends to 0. We have obtained the order of divergence, u^{-4} and u^{-6}.

More rigorous mathematical analysis when $v = 1/2$ is given in Burnashev and Amari (2002) by using the technique of large deviation.

5 Model Selection and Singularity

We have discussed the behavior of the maximum likelihood estimator (MLE) when two stimuli are presented. When the two are located closely, the estimation of the two locations is difficult. One may decide the number of stimuli (one or two in the present case) after the estimation of the parameters, but one may decide the number before estimation based on the encoded pattern $\boldsymbol{r}$ and then perform estimation. We apply different statistical models: one including single stimulus and the other including two stimuli. By comparing the results, one can decide the number. This is the problem of model selection.

There are two typical methods frequently used in model selection. One is the Akaike information criterion (AIC), which evaluates the Kullback-Leibler divergence between the true distribution and the estimated distribution based on the two models. The other is the Bayesian information criterion (BIC), which evaluates the posterior probabilities of the two models. The minimum descriptive length (MDL) criterion gives practically the same criterion as the BIC, although the underlying philosophy is different.

One may apply these criteria in the singular case, obtaining a practical answer to decide the number of stimuli. However, all the criteria (AIC, BIC, MDL) are based on the statistical analysis of the behavior of the MLE under each model. In these analyses, the estimator is assumed to be asymptotically subject to the gaussian distribution with the covariance matrix equal to the inverse of the Fisher information matrix. Since the Fisher information is degenerate in the present singular case, the MLE is not subject to the gaussian distribution even asymptotically. So one should be very careful in applying these criteria in hierarchical models including singularity, because no mathematical justification exists.

6 Bayesian Inference and Singularity

The Bayesian inference presumes an adequate prior distribution over the parameters of the model. One may use a parametric family of priors, including hyperparameters, thus introducing a hierarchical structure of Bayesian inference.

Given observed data $\mathbf{r}$, one can calculate the posterior probability of the parameters under the model

$$p(\boldsymbol{\xi}|\mathbf{r}) = \frac{\pi(\boldsymbol{\xi})p(\mathbf{r}|\boldsymbol{\xi})}{p(\mathbf{r})}, \tag{6.1}$$

where $\pi(\boldsymbol{\xi})$ is the prior distribution and $p(\mathbf{r})$ is the marginal distribution of $\mathbf{r}$. When one wants to combine the problem of model selection, where we presume a number of models $M_1, M_2, \cdots$, and the posterior is

$$p(\boldsymbol{\xi}_i, M_i|\mathbf{r}) = \frac{\pi(M_i)\,\pi(\boldsymbol{\xi}_i|M_i)\,p(\mathbf{r}|\boldsymbol{\xi}_i)}{p(\mathbf{r})}, \tag{6.2}$$

where $\pi(M_i)$ is the prior of the model M_i with parameters $\boldsymbol{\xi}_i$. One may use the marginal distribution $p(M_i|\mathbf{r})$ for model selection (BIC). The MAP estimator is the one that maximizes the posterior distribution. This is a point estimator, and it is known that the MAP estimator includes the same amount of Fisher information as the MLE. The Bayesian predictive distribution is obtained by using the posterior distributions of parameters to give a distribution of new data. The predictive distribution does not belong to the model, but is optimal when the prior is true. In general, it gives a good estimator, provided the prior is not extreme. The amount of Fisher information included in it is larger than that included in the MAP or MLE estimator, but the difference is small or only higher order, which can be revealed by calculating the curvature direction of the ancillary statistics (Komaki, 1996).

When a smooth, positive prior is used, all the Bayesian estimators are asymptotically the same in a regular statistical model. One choice is the Jeffreys noninformative prior, which uses the square root of the determinant of the Fisher information matrix, and its optimality is guaranteed by information theory (Clarke & Barron, 1990). Hence, the performance of the

Bayesian inference is asymptotically guaranteed to be good, and if one can choose an adequate prior based on prior knowledge, its performance is better. However, when one chooses an extreme prior, say diverging to infinity at some points or zero on some region, the result of the inference is biased and strongly affected by the chosen prior.

When the statistical model is singular, as is our model, a serious problem arises. The Jeffreys prior converges to 0 in the critical region of parameters. Hence, this is an extreme case. One usually assumes a smooth prior, in particular, the uniform prior, on the model M. Instead of the parameter space M, we may consider the space $\tilde{M}$ of behaviors, which are the space of all the tuning functions $f(z, \xi)$. A smooth prior on $\tilde{M}$ looks very natural to be assumed, but it becomes singular on M because the critical set on M corresponds to an infinite number of the same probability distributions in $\tilde{M}$. Therefore, the uniform (or smooth) prior counts a probability distribution belonging to $\tilde{M}$ infinitely many times in M. If we assume the uniform (or smooth) prior on the reduced model $\tilde{M}$ of behaviors, then such multiplicity disappears, as the Jeffreys prior does. However, such a prior becomes 0 on the critical region of the original parameter space. Therefore, the prior becomes singular in a singular statistical model. There has been little study on the Bayesian inference in a singular statistical model except for the algebraic geometry theory of Watanabe (2001a, 2001b, 2001c).

We can observe that the uniform prior favors a smaller model (one stimulus in our case), so that it suppresses automatically larger models, avoiding overfitting. When the true distribution is in a smaller model, it gives a better performance. Watanabe and his colleagues have studied the Bayesian inference in singular models by using algebraic geometry, which we do not recapitulate here.

7 Synfiring Solution

How can the brain solve the difficulty in the singular situation? We propose a possible solution in the scheme of synchronous firing. Given two stimuli representing different objects, the synfire hypothesis (Abeles, 1991; Diesmann, Gewaltig, & Aertsen, 1999; von der Malsburg, 1999; Singer, 1999) poses that the responses due to one object are synchronized, while those by the other fire synchronously in another phase. For simplicity, we assume there are only two phases, phase 1 and phase 2, and object 1 excites neurons in phase 1 with degree α and in phase 2 with the complementary degree $\bar{\alpha} = 1 - \alpha$. As for object 2, α and $\bar{\alpha}$ are interchanged. Hence, the responses of the neural field in phases 1 and 2 are, respectively,

$$r_1(z) = \alpha\bar{v}\varphi(z - x_1) + \bar{\alpha}v\varphi(z - x_2) + \sigma\varepsilon_1(z) \tag{7.1}$$

$$r_2(z) = \bar{\alpha}\bar{v}\varphi(z - \theta_1) + \alpha v\varphi(z - \theta_2) + \sigma\varepsilon_2(z), \tag{7.2}$$

where $\bar{v} = 1 - v$ and $\varepsilon_i(z)$ are independent noises. The degree of synchrony is represented by α, and there is no synchrony when $\alpha = 1/2$. For

$$\tilde{\mathbf{r}} = (\mathbf{r}_1(z), \mathbf{r}_2(z)), \tag{7.3}$$

the probability of $\tilde{\mathbf{r}}$ is the product of $Q(\mathbf{r}_1(z))$ and $Q(\mathbf{r}_2(z))$ of equation 2.6,

$$Q(\tilde{\mathbf{r}}) = Q(\mathbf{r}_1)\, Q(\mathbf{r}_2), \tag{7.4}$$

because r_1 and r_2 are independent since ε_1 and ε_2 are independent. Hence, the Fisher information is the sum of those in phases 1 and 2.

We calculate the Fisher information $I_F^1(\boldsymbol{\xi})$ included in $\mathbf{r}_1$ directly at $u = 0$. It is given by

$$I_F^1(\boldsymbol{\xi}) = \begin{bmatrix} (\alpha\bar{v}+\bar{\alpha}v)^2 g_1 & -(\alpha-\bar{\alpha})(\alpha\bar{v}+\bar{\alpha}v)v\bar{v}g_1 & 0 \\ -(\alpha-\bar{\alpha})(\alpha\bar{v}+\bar{\alpha}v)v\bar{v}g_1 & (\alpha-\bar{\alpha})^2(v\bar{v})^2 g_1 & 0 \\ 0 & 0 & (\alpha-\bar{\alpha})^2 g_0 \end{bmatrix}. \tag{7.5}$$

Its determinant vanishes so that $\mathbf{r}_1$ has singular information. The $I_F^2(\boldsymbol{\xi})$ in phase 2 is given by interchanging α and $\bar{\alpha}$ in I_F^1. The total Fisher information is

$$\begin{aligned} I_F(\boldsymbol{\xi}) &= I_F^1(\boldsymbol{\xi}) + I_F^2(\boldsymbol{\xi}) \\ &= \begin{bmatrix} cg_1 & (\alpha-\bar{\alpha})^2(v-\bar{v})g_1 & 0 \\ (\alpha-\bar{\alpha})^2(v-\bar{v})g_1 & 2(\alpha-\bar{\alpha})^2(v\bar{v})^2 g_1 & 0 \\ 0 & 0 & 2(\alpha-\bar{\alpha})^2 g_0 \end{bmatrix}, \end{aligned} \tag{7.6}$$

$c = \left(\alpha^2 + \bar{\alpha}^2\right)\left(v^2 + \bar{v}^2\right) + 4\alpha\bar{\alpha}v\bar{v}$, which is nonsingular even at $u = 0$, except for the case with $\alpha = \bar{\alpha} = 1/2$.

The effect of synchrony is dramatic even when the degree of synchrony is weak. The population coding is no longer singular, keeping sufficient information to distinguish the two objects. We next discuss a possible mechanism of inducing synchrony.

8 Remarks on Binding Different Features

A more general aspect of the binding problem is discussed in terms of correlated or synchronous firing. Let an object O have two different features, say, location and orientation, represented by analog features x and x', respectively. In a typical example in the binding problem, x represents the shape parameter and x' the color of an object. Assume that two objects O_1 and O_2 are presented, with respective features (x_1, x_1') and (x_2, x_2'). If there exists a neural field representing the features (x, x') jointly, when one object (x, x') is applied, the tuning function $f(z, z')$ in the two-dimensional field with coordinates (z, z') has a peak at (x, x') of the presented object, and the neural response is its noisy version $r(z, z') = f(z - x, z' - x') + \sigma\varepsilon(z, z')$.

When two objects are presented simultaneously, the response is

$$r(z, z') = (1 - v)\varphi\left(z - x_1, z' - x_1'\right) + v\varphi\left(z - x_2, z' - x_2'\right) + \sigma\varepsilon(z, z'). \tag{8.1}$$

Hence, r has two peaks at (x_1, x_1') and (x_2, x_2'), and there is no difficulty for binding x_i with x_i', provided the two peaks are separated. When they are close, the above type of singularity occurs.

It is not plausible to assume the existence of a field combining any two features. There will be, in general, two different fields, Z and Z'—one representing x and the other x'. Given two objects, two peaks are aroused in field Z at x_1 and x_2, and two peaks in field Z' at x_1' and x_2', in the respective response patterns:

$$r(z) = f\ (z; x_1, x_2) + \sigma\varepsilon(z), \tag{8.2}$$

$$r'\left(z'\right) = f\ \left(z'; x_1', x_2'\right) + \sigma\varepsilon'\left(z'\right). \tag{8.3}$$

The problem is to find which peak, x_1 or x_2, is bound to which peak, x_1' or x_2'. Synchronization is a solution, as we have analyzed.

Synchrony is generated by the correlated firing of neurons. A simple mechanism causing pairwise and higher-order correlations has already been shown (Amari, Nakahara, Wu, & Saka, 2003). We extend that model to be applicable in our case.

Here, we assume spiking neurons in a neural field Z. Let $p(z)$ be the potential of the neuron at z, and the neuron fires when $p(z)$ exceeds a threshold. The potential is given by the tuning function disturbed by fluctuating noises,

$$p(z) = \frac{1}{2}\varphi\left(z - x_1\right) + \frac{1}{2}\varphi\left(z - x_2\right) + \sigma\varepsilon(z), \tag{8.4}$$

where the noise term is decomposed as

$$\varepsilon(z) = \varphi\left(z - x_1\right)\varepsilon_1(z) + \varphi\left(z - x_2\right)\varepsilon_2(z) + \varepsilon_0(z). \tag{8.5}$$

Here, ε_1, ε_2, and $\varepsilon_0(z)$ are independent and gaussian random variables. The noise terms ε_1 and ε_2 are multiplicative, representing the Poisson character of spiking neurons. When $\varepsilon_1, \varepsilon_2$, and ε_0 are temporally independent (or temporal correlations decay to 0 quickly), we have a model of spiking neurons.

The potential in field Z' has a similar nature,

$$p'(z') = (1 - v)\varphi\left(z' - x_1'\right) + v\varphi\left(z' - x_2'\right) + \sigma\varepsilon'(z'), \tag{8.6}$$

where the noise in field Z' is also decomposed as

$$\varepsilon'(z') = \varphi\left(z' - x_1'\right)\varepsilon_1(z') + \varphi\left(z' - x_2'\right)\varepsilon_2(z') + \varepsilon_0'(z'). \tag{8.7}$$

It is plausible that noises ε_1 and ε_2 originate from the objects O_1 and O_2, respectively, and are common to fields Z and Z'. It is analyzed in Amari et al. (2003) that synchrony emerges in such a model. This can solve the binding problem. More detailed analysis using the Fisher information is in our future program.

9 Discussion

This article investigated the Fisher information when two stimuli are presented together. Since ordinary statistical analysis fails in this case due to the degeneracy of the Fisher information, we introduced a novel method using algebraic singularity, and this method allowed us to examine not only the Fisher information but also the algebraic structure of the degeneracy. We analyzed the asymptotic property of how the information decays as the locations of the two stimuli near each other, although we do not propose any decoding algorithms.

We then searched for a possible neural mechanism to resolve singularity. This is given by the synchronization of the neural firing in population coding. We analyzed how the Fisher information behaves under synchronization. We finally discussed a possible mechanism of generating synchronization in relation to the binding problem. Our method synthesizes various disciplines from many fields.

Acknowledgments

H.N. is supported by Grants-in-Aid 14780614 and 14017106 from the MEXT, Japan.

References

Abeles, M. (1991). *Corticonics*. Cambridge: Cambridge University Press.

Amari, S. (1977). Dynamics of pattern formation in lateral-inhibition type neural fields. *Biological Cybernetics*, *27*, 77–87.

Amari, S. (1998). Natural gradient works efficiently in learning. *Neural Computation, 10*, 251–276.

Amari, S., & Arbib, M. A. (1977). Competition and cooperation in neural nets. In J. Metzler (Ed.), *Systems neuroscience* (pp. 119–165). San Diego: Academic Press.

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

Amari, S., Nakahara, H., Wu, S., & Sakai, Y. (2003). Synchronous firing and higher-order interactions in neuron pool. *Neural Computation, 15*, 127–142.

Amari, S., & Ozeki, T. (2001). Differential and algebraic geometry of multilayer perceptrons. *IEICE Transactions on Fundamentals of Electronics Communications and Computer Sciences, E84-A*(1), 31–38.

Burnashev, M. V., & Amari, S. (2002). On some singularities in parameter estimation problems. *Problems of Information Transmission, 38*, 328–346

Clarke, B. S., & Barron, A. R. (1990). Information theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory, 36*, 453–471

Dacunha-Castelle, D., & Gassiat, E. (1997). Testing in locally conic models, and application to mixture models. *Probability and Statistics, 1*, 285–317.

Diesmann, M., Gewaltig, M-O., & Aertsen, A. (1999). Stable propagation of synchronous spiking in cortical neural networks. *Nature, 402*, 529–533.

Eurich, C. W., & Wilke, S. D. (2000). Multidimensional encoding strategy of spiking neurons. *Neural Computation, 12*(7), 1519–1529.

Fukumizu, K. (2003). Likelihood ratio of unidentifiable models and multilayer neural networks. *Annals of Statistics, 31*, 833–851.

Fukumizu, K., & Amari, S. (2000). Local minima and plateaus in hierarchical structures of multilayer perceptrons. *Neural Networks, 13*(3), 317–327.

Hotelling, H. (1939). Tubes and spheres in *n*-spaces, and a class of statistical problems. *Amer. J. Math., 61*, 440–460.

Ingle, D. (1968). Visual releasers of pre-catching behavior in frogs and toads. *Brain Behavior and Evolution, 1*, 500–518.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika, 83*, 299–313.

Liu, X., & Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *Annals of Statistics, 31*(3), 807–832.

Martínez-Trujillo, J. C. M., & Treue, S. (2000). Attention modulates tuning curves from two stimuli inside the receptive field of MT/MST cells in the macaque monkey. *Society for Neuroscience Abstracts, 26*, 251.16.

Nakahara, H., & Amari, S. (2002). Attention modulation of neural tuning through peak and base rate in correlated firing. *Neural Networks, 15*, 41–55.

Nakahara, H., Wu, S., & Amari, S. (2001). Attention modulation of neural tuning through peak and base rate. *Neural Computation, 13*, 2031–2048.

Pouget, A., Dayan, P., & Zemel, R. (2000). Information processing with population codes. *Nature Neuroscience Reviews, 1*, 125–132.

Pouget, A., Zhang, K., Deneve, S., & Latham, P. E. (1998). Statistically efficient estimation using population coding. *Neural Computation, 10*, 373–401.

Reichardt, W., & Poggio, T. (1975). A theory of the pattern induced flight orientation of the fly, *Muska domestica*, II. *Biological Cybernetics, 18*, 69–80.

Sahani, M., & Dayan, P. (2003). Doubly distributional population codes: Simultaneous representation of uncertainty and multiplicity. *Neural Computation*, *15*, 2255–2279.

Seung, H. S., & Sompolinsky, H. (1993). Simple models for reading neuronal population codes. *Proc. Nat. Acad. Sci. USA*, *90*, 10749–10753.

Singer, W. (1999). Neuronal synchrony: A versatile code for the definition of relations? *Neuron*, *24*, 49–65.

Treue, S., Hol, K., & Rauber, H. J. (2000). Seeing multiple directions of motion-physiology and psychophysics. *Nature Neuroscience*, *3*, 270–276.

von der Malsburg, C. (1999). The what and why of binding: The modeler's perspective. *Neuron*, *24*, 95–104.

Watanabe, S. (2001a). Algebraic analysis for non-identifiable learning machines. *Neural Computation*, *13*, 899–933.

Watanabe, S. (2001b). Algebraic geometrical methods for hierarchical learning machines. *Neural Networks*, *14*(8), 1409–1060.

Watanabe, S. (2001c). Algebraic information geometry for learning machines with singularities. In T. K. Leen, T. G. Dieterrich, & V. Tresp (Eds.), *Advances in neural information processing systems*, *13* (pp. 329–336) Cambridge, MA: MIT Press.

Weyl, H. (1939). On the volume of tubes. *Amer. J. Math.*, *61*, 461–472.

Wu, S., Amari, S., & Nakahara, H. (2002a). Asymptotic behaviors of population codes. *Neurocomputing*, *44–46*, 697–702.

Wu, S., Amari, S., & Nakahara, H. (2002b). Population coding and decoding in a neural field: A computational study. *Neural Computation*, *14*, 999–1026.

Wu, S., Amari, S., & Nakahara, H. (2004). Information processing in a neuron ensemble with the multiplicative correlation structure. *Neural Networks*, *17*(2), 205–214.

Wu, S., Chen, D., Niranjan, M., & Amari, S. (2003). Sequential Bayesian decoding with a population of neurons. *Neural Computation*, *15*(5), 993–1012.

Wu, S., Nakahara, H., & Amari, S. (2001). Population coding with correlation and an unfaithful model. *Neural Computation*, *13*, 775–797.

Yoon, H., & Sompolinsky, H. (1999). The effect of correlations on the Fisher information of population codes. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing*, *II* (pp. 167–173). Cambridge, MA: MIT Press.

Zemel, R. S., Dayan, P., & Pouget, A. (1998). Probabilistic interpretation of population codes. *Neural Computation*, *10*, 403–430.

Zemel, R. S., & Pillow, J. (2002). A probabilistic network model of population responses. In R. P. N. Rao, B. A. Olshausen, & M. S. Lewicki (Eds.), *Probabilistic models of the brain: Perception and neural function* (pp. 223–242). Cambridge, MA: MIT Press.

Zhang, K., & Sejnowski, T. J. (1999). Neural tuning: To sharpen or broaden. *Neural Computation*, *11*, 75–84.

Received November 24, 2003; accepted August 31, 2004.