Information loss associated with imperfect observation and mismatched decoding

Masafumi Oizumi[1]*, Masato Okada[1,2] and Shun-Ichi Amari[1]

[1] Laboratory for Mathematical Neuroscience, RIKEN Brain Science Institute, Wako-City, Saitama, Japan
[2] Department of Complexity Science and Engineering, The University of Tokyo, Kashiwa-City, Chiba, Japan

We consider two types of causes leading to information loss when neural activities are passed and processed in the brain. One is responses of upstream neurons to stimuli being imperfectly observed by downstream neurons. The other is upstream neurons non-optimally decoding stimuli information contained in the activities of the downstream neurons. To investigate the importance of neural correlation in information processing in the brain, we specifically consider two situations. One is when neural responses are not simultaneously observed, i.e., neural correlation data is lost. This situation means that stimuli information is decoded without any specific assumption about neural correlations. The other is when stimuli information is decoded by a wrong statistical model where neural responses are assumed to be independent even when they are not. We provide the information geometric interpretation of these two types of information loss and clarify their relationship. We then concretely evaluate these types of information loss in some simple examples. Finally, we discuss use of these evaluations of information loss to elucidate the importance of correlation in neural information processing.

Keywords: Fisher information, information loss, imperfect observation, mismatched decoding, information geometry, correlated activity, coincidence detection

INTRODUCTION

Neurons in early sensory areas represent the information of various stimuli from the external world by their noisy activities. The noise inherent in neural activities needs to be properly handled by the nervous system for the information to be accurately processed. One simple but powerful means for coping with the neural noise is population coding. Neurophysiological experiments have shown that many neurons with different selectivities respond to particular stimuli. These findings suggest that the nervous system represents information through population activities, which would be helpful for accurate information processing. This coding scheme for stimulus information is known as population coding.

An important feature of population coding is that the activity of neurons is correlated (Gray et al., 1989; Gawne and Richmond, 1993; Zohary et al., 1994; Meister et al., 1995; Lee et al., 1998; Ishikane et al., 2005; Averbeck et al., 2006; Ohiorhenuan et al., 2010; but see Ecker et al., 2010). A crucial question is how much information the nervous system can extract from correlated population activities. In general, it is difficult for the nervous system to maximally extract information when its activities are correlated because two conditions must be satisfied. First, downstream neurons must perfectly observe the responses of the upstream neurons (perfect observation of neural responses). Second, downstream neurons must optimally decode the information from the observed neural responses (optimal decoding of stimulus). In other words, if either the observation or the decoding is imperfect or non-optimal, which are both likely situations in the nervous system, stimuli information is inevitably degraded. In this work, we discuss the amount of information loss associated with imperfect observation and non-optimal decoding.

With regard to non-optimal decoding, several researchers including the authors of this paper have investigated how much information would be lost if neural correlation is ignored in decoding (Nirenberg et al., 2001; Wu et al., 2001; Golledge et al., 2003; Averbeck and Lee, 2006; Oizumi et al., 2010). This type of decoding is called mismatched decoding (Merhav et al., 1994; Oizumi et al., 2010) because the decoding does not match the actual neural activities, i.e., the actual neural activities are correlated but the correlations are ignored in decoding. With regard to imperfect observation, this work is the first to address this issue. As an example of imperfect observation, we specifically consider the situation that the activity of neurons is not simultaneously observed by downstream neurons (**Figure 1**). This is related to whether the coincidence detector plays an important role in information processing in the brain (Abeles, 1982; König et al., 1996). If a large proportion of the total information is lost when the responses of neurons are not simultaneously observed, coincidence detection would be necessary for accurate information processing.

The framework of mismatched decoding was introduced to quantify the importance of correlated activity because this importance could be quantified by the amount of information loss when neural correlation is ignored in decoding (Nirenberg et al., 2001; Wu et al., 2001; Nirenberg and Latham, 2003). Similarly, the importance of correlated activity could be also measured using the amount of information loss due to non-simultaneous observation. We therefore obtain two different measures for the importance of correlated activity by introducing the concepts of imperfect observation and mismatched decoding. We aim to clarify the relationship between the measures and provide a simple guide on how to use them.

FIGURE 1 | Schematic of simultaneous and non-simultaneous observation of neural responses.
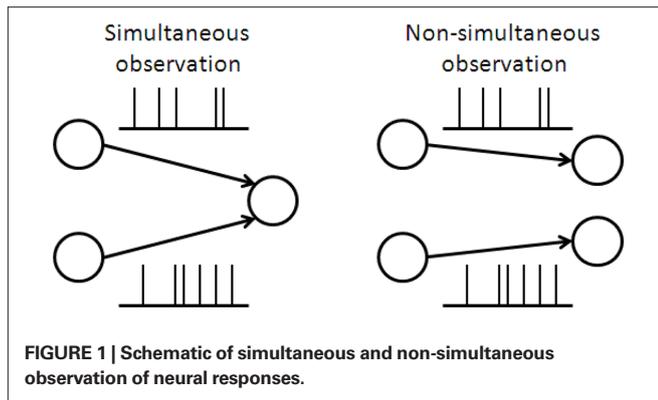
Combining the concepts of imperfect observation and mismatched decoding produces four types of situations where stimulus is inferred: (1) observation is perfect and decoding is optimal; (2) decoding is optimal but observation is imperfect; (3) observation is perfect but decoding is mismatched; and (4) observation is imperfect and decoding is mismatched. We discuss the inferences in these four types of situations from the viewpoint of information geometry (Amari and Nagaoka, 2000) and then clarify their relationships. We also specifically compute the amount of information obtained through these four inference types for neural responses described by the Gaussian model and by a binary probabilistic model.

This paper is organized as follows. First, we introduce the concept of an exponential family and two probability distributions that belong to the exponential family, i.e., the Gaussian distribution and the binary probabilistic model, which have both been intensively investigated as representative models of neural responses (Abbott and Dayan, 1999; Amari, 2001; Nakahara and Amari, 2002). Second, we provide information geometric interpretation of the four types of inference mentioned above and describe how to evaluate the information in each of the four types by using the Fisher information. Third, we compute the amount of information in each of the four types of inference in the Gaussian model and explain the relationship between the inference with imperfect observation and that with mismatched decoding. Fourth, we also compute the amount of information obtained by the four types of inference in simple binary probabilistic models. Finally, we summarize the results and discuss how to use the two measures introduced in this study for quantifying the importance of neural correlation and mention some of the future directions of this work.

EXPONENTIAL FAMILY OF PROBABILITY DISTRIBUTIONS

Let us denote the conditional probability distribution for a neural response $\mathbf{r} = (r_1, r_2, \ldots, r_N)$ over a population of N neurons being evoked by a stimulus s as $p(\mathbf{r}; s)$, where s is a continuous variable. We assume that $p(\mathbf{r}; s)$ belongs to the exponential family. Probability distributions that belong to an exponential family can be written as

$$p(\mathbf{r};s) = \exp\left(\sum_{I=1}^{n} R_I(\mathbf{r})\theta_I(s) - \psi(s) \right), \quad (1)$$

where $R_I(\mathbf{r})$ is a function of neural responses $\mathbf{r}$, $\theta_I(s)$ is a function of stimulus s, and a normalization constant $\psi(s)$ is also a function of stimulus s. $\theta_I(s)$ is called the natural parameter of the exponential family. Two examples of probability distributions are investigated in this paper.

EXAMPLE 1. GAUSSIAN DISTRIBUTION

The number of spikes emitted by a neuron over a fixed time period (time-averaged rate) or the total number of spikes emitted by a population of neurons (population-averaged rate), which is denoted by $\mathbf{r}$, may be described by the Gaussian distribution

$$p(\mathbf{r};s) = \frac{1}{\sqrt{(2\pi)^N \det[\mathbf{C}(s)]}} \exp\left(-\frac{1}{2}(\mathbf{r} - \mathbf{f}(s))^T \mathbf{C}^{-1}(s)(\mathbf{r} - \mathbf{f}(s)) \right), (2)$$

where N is the number of neurons, $\mathbf{f}(s)$ is the average number of spikes, and $\mathbf{C}(s)$ is the covariance matrix. If we rewrite Eq. 2 as follows, we can see that the Gaussian distribution belongs to the exponential family

$$p(\mathbf{r};s) = \exp\left(\mathbf{C}^{-1}(s)\mathbf{f}(s) \cdot \mathbf{r} - \frac{1}{2}\mathbf{C}^{-1}(s) \cdot \overline{\mathbf{r}\mathbf{r}}^T - \frac{1}{2}\log\det[\mathbf{C}(s)] \right.$$
$$\left. - \frac{1}{2}\mathbf{f}^T(s)\mathbf{C}^{-1}\mathbf{f}(s) \right). \quad (3)$$

In this case,

$$\mathbf{R}_1 = \{r_i\} = \{r_1, \cdots, r_N\}, \quad (4)$$

$$\mathbf{R}_2 = \{r_i r_j\} = \{r_1 r_1, r_1 r_2, \cdots, r_{N-1} r_N, r_N r_N\}, \quad (5)$$

$$\boldsymbol{\theta}_1 = \left\{ \sum_{j=1}^{N} C_{ij}^{-1}(s) f_j(s) \right\}, \quad (6)$$

$$\boldsymbol{\theta}_2 = \left\{ -\frac{1}{2} C_{ij}^{-1}(s) \right\}, \quad (7)$$

$$\psi(s) = \frac{1}{2}\log\det[\mathbf{C}(s)] + \frac{1}{2}\mathbf{f}^T(s)\mathbf{C}^{-1}\mathbf{f}(s). \quad (8)$$

EXAMPLE 2. LOG-LINEAR MODEL OF BINARY NEURAL RESPONSE

When we analyze neural responses within a short time period (~1–10 ms, typically), the neural responses are considered stochastic binary variables: $r_i = 1$ when the ith neuron fires within the time bin, and $r_i = 0$ when it does not. The joint distributions of N random binary variables can be generally written in the following form (Amari, 2001; Nakahara and Amari, 2002):

$$p(\mathbf{r};s)$$
$$= \exp\left(\sum_{i=1}^{N} \theta_i(s) r_i + \sum_{i<j} \theta_{ij}(s) r_i r_j + \cdots + \theta_{12\ldots N}(s) r_1 r_2 \cdots r_N - \psi(s) \right), \quad (9)$$

where the normalization constant $\psi(s)$ is given by

$$\psi(s) = -\log p(r_1 = r_2 = \cdots = r_N = 0; s). \quad (10)$$

This probability distribution is clearly in the exponential family form. In this case,

$$\mathbf{R}_1 = \{r_i\} = \{r_1, \cdots, r_N\},$$

$$\mathbf{R}_2 = \{r_i r_j\} = \{r_1 r_1, r_1 r_2, \cdots, r_{N-1} r_N, r_N r_N\},$$

$$\cdots,$$

$$\mathbf{R}_N = \{r_1 r_2 \cdots r_N\} \tag{11}$$

$$\boldsymbol{\theta}_1 = \{\theta_i(s)\},$$

$$\boldsymbol{\theta}_2 = \{\theta_{ij}(s)\},$$

$$\cdots,$$

$$\boldsymbol{\theta}_N = \{\theta_{12\cdots N}(s)\}. \tag{12}$$

Recent investigations have shown that the observed statistics of neural responses can be sufficiently captured by this type of probabilistic model, which contains up to second-order correlation terms (Schneidman et al., 2006; Shlens et al., 2006; Tang et al., 2008), although the importance of higher-order correlations has also been discussed (Amari et al., 2003; Montani et al., 2009; Ohiorhenuan et al., 2010). For simplicity, we consider only the second-order correlations and ignore higher-order correlations in this work, i.e.,

$$\boldsymbol{\theta}_I = 0, \quad (\text{for } I > 2). \tag{13}$$

INFERENCE OF STIMULUS AND FISHER INFORMATION

We consider the inference problem of how accurately the stimulus value s can be estimated when the stochastic neural response $\mathbf{r}$ is given. We assume that neural response $\mathbf{r}$ is observed many times. The neural response at the tth trial is denoted by $\mathbf{r}(t)$ and the number of trials by T. If each neural response is independent and identically distributed, the probability distribution for T observations of neural responses is given by

$$p_T(\mathbf{r}^T; s) = \prod_{t=1}^{T} p(\mathbf{r}(t); s), \tag{14}$$

$$= \exp\left(\sum_{I=1}^{2} \overline{\mathbf{R}}_I(\mathbf{r}) \cdot \boldsymbol{\theta}_I(s) - T\psi(s)\right), \tag{15}$$

where $\overline{\mathbf{R}}_I$ are sufficient statistics for the probability distribution and are given by

$$\overline{\mathbf{R}}_1 = \{\overline{r}_i\} \qquad \overline{r}_i = \frac{1}{T}\sum_{t=1}^{T} r_i(t), \tag{16}$$

$$\overline{\mathbf{R}}_2 = \{\overline{r}_{ij}\} \qquad \overline{r}_{ij} = \frac{1}{T}\sum_{t=1}^{T} r_i(t) r_j(t). \tag{17}$$

We evaluate the accuracy of the estimate by using the Fisher information,

$$g(s) = -E_{p_T}\left[\frac{d^2 l_s}{ds^2}\right] = -\int d\mathbf{r}^T p_T(\mathbf{r}^T; s) \frac{d^2 \log p_T(\mathbf{r}^T; s)}{ds^2}, \tag{18}$$

where $l_s = \log p_T(\mathbf{r}^T; s)$ and E_{p_T} denotes the expectation with respect to the distribution $p_T(\mathbf{r}^T; s)$. Through the Cramér–Rao bound, the Fisher information bounds the average squared decoding error for an unbiased estimate as follows:

$$E_{p_T}\left[(s - \hat{s})^2\right] \geq \frac{1}{g(s)}, \tag{19}$$

where s is the true stimulus value and $\hat{s}$ is the estimate. Since the Fisher information is the lower bound of the mean square error, behavior of the mean square error and the Fisher information could be different in general (Brunel and Nadal, 1998; Yaeli and Meir, 2010). However, the maximum likelihood estimator, which chooses s for an estimate that maximizes likelihood function $p_T(\mathbf{r}^T; s)$, achieves the Cramér–Rao bound (Eq. 19) as $T \to \infty$. A Bayesian estimator, which is generally a biased estimator, can also achieve the Cramér–Rao bound as $T \to \infty$ because it becomes equivalent to the maximum likelihood estimator as $T \to \infty$.

We compute the Fisher information with respect to stimulus s from the Fisher information matrix with respect to the natural parameters,

$$g_{IJ}(\boldsymbol{\theta}) = E_{p_T}\left[\partial_I l_{\boldsymbol{\theta}} \partial_J l_{\boldsymbol{\theta}}\right], \tag{20}$$

$$= -E_{p_T}\left[\partial_I \partial_J l_{\boldsymbol{\theta}}\right], \tag{21}$$

where $l_{\boldsymbol{\theta}} = p_T(\mathbf{r}^T; \boldsymbol{\theta})$, and $\partial_I = \partial/\partial\theta_I$. From the definition of the Fisher information matrix with respect to the natural parameters (Eqs. 20 and 21), the natural parameters are given by

$$g_{IJ}(\theta) = E_{p_T}\left[\left(\overline{R}_I - \eta_I\right)\left(\overline{R}_J - \eta_J\right)\right], \tag{22}$$

$$= T\partial_I \partial_J \psi(\theta), \tag{23}$$

where

$$\eta_I = E_{p_T}\left[\overline{R}_I\right] = E_p\left[R_I\right]. \tag{24}$$

By using $g_{IJ}(\boldsymbol{\theta})$, the Fisher information with respect to the stimulus s can be written as

$$g(s) = \sum_{I,J} g_{IJ}(\boldsymbol{\theta}) \frac{\partial \theta_I}{\partial s} \frac{\partial \theta_J}{\partial s}. \tag{25}$$

The Fisher information $g(s)$ described above determines the accuracy of the estimate of s under two conditions: (1) all sufficient statistics $\overline{R}_I$ are available, and (2) the likelihood function $p(\mathbf{r}; s)$ is exactly known. Regarding the first condition, downstream neurons may not be able to simultaneously access the responses of all upstream neurons. This imperfect observation of neural responses by downstream neurons leads to loss of information. Similarly, regarding the second condition, downstream neurons are unlikely to completely know the likelihood function $p(\mathbf{r}; s)$. Downstream neurons are more likely to only partially know $p(\mathbf{r}; s)$ and to decode the stimulus based on a decoding model $q(\mathbf{r}; s)$, which is not equal to $p(\mathbf{r}; s)$ but partially matches $p(\mathbf{r}; s)$ (Nirenberg et al., 2001; Wu et al.,

2001; Oizumi et al., 2010). This mismatched decoding of stimuli by downstream neurons also results in loss of information. These two types of information loss are evaluated next.

INFORMATION LOSS CAUSED BY IMPERFECT OBSERVATION OF NEURAL RESPONSES

In this work, we specifically consider the situation that second-order sufficient statistics $\bar{\mathbf{R}}_2$ are not accessible to downstream neurons and only first-order sufficient statistics $\bar{\mathbf{R}}_1$ are available to them. This is related to whether coincidence detector neurons are needed to accurately estimate the stimulus. To evaluate the loss of information associated with loss of data, we first marginalize the joint probability distribution $p(\bar{R}_1, \bar{R}_2; s)$ over $\bar{R}_2$:

$$p\left(\bar{R}_1; s\right) = \int d\bar{R}_2\, p\left(\bar{R}_1, \bar{R}_2; s\right). \tag{26}$$

When only $\bar{R}_1$ is observed, the Fisher information with respect to stimulus s is given by

$$g^L(s) = -\int d\bar{R}_1\, p\left(\bar{R}_1; s\right) \frac{d^2 \log p\left(\bar{R}_1; s\right)}{ds^2}. \tag{27}$$

The information loss associated with loss of $\bar{R}_2$ is

$$\Delta g = g(s) - g^L(s). \tag{28}$$

INFORMATION LOSS CAUSED BY MISMATCHED DECODING OF STIMULUS

We evaluate the loss of information when downstream neurons infer the stimulus parameter s based on not the correct probability distribution $p(\mathbf{r}; s)$ but a mismatched probability distribution $q(\mathbf{r}; s)$. We assume that $q(\mathbf{r}; s)$ also belongs to the exponential family and that the maximum likelihood estimation based on $q(\mathbf{r}; s)$ is consistent, i.e.,

$$E_p\left[\frac{dl_q(\mathbf{r}; s)}{ds}\right] = 0, \tag{29}$$

where E_p denotes the expectation with respect to $p(\mathbf{r}; s)$ and $l_q(\mathbf{r}; s) = \log q(\mathbf{r}; s)$. We evaluate the squared decoding error of the maximum likelihood estimation with the mismatched likelihood function $q(\mathbf{r}; s)$ based on T observations of neural responses, $\mathbf{r}^T$. The estimate $\hat{s}_q$ is given by

$$\hat{s}_q = \arg\max_s l_{q_T}\left(\mathbf{r}^T; s\right), \tag{30}$$

$$= \arg\max_s \sum_{t=1}^{T} l_q(\mathbf{r}(t); s). \tag{31}$$

By differentiating $l_{q_T}(\mathbf{r}^T; s)$ with respect to s at $\hat{s}_q$, we obtain the quasi-likelihood estimating equation

$$\sum_{t=1}^{T} \frac{dl_q\left(\mathbf{r}(t); \hat{s}_q\right)}{ds} = 0. \tag{32}$$

When the left-hand side terms of Eq. 32 are expanded at the true value of stimulus parameter s,

$$0 = \sum_{t=1}^{T} \frac{dl_q\left(\mathbf{r}(t); \hat{s}_q\right)}{ds},$$

$$= \sum_{t=1}^{T} \frac{dl_q(\mathbf{r}(t); s)}{ds} + \sum_{t=1}^{T} \frac{d^2 l_q(\mathbf{r}(t); s)}{ds^2}\left(s - \hat{s}_q\right). \tag{33}$$

According to the central limit theorem, the first term on the right-hand side of Eq. 33 converges to a Gaussian distribution with mean 0 and variance $TE_p[(dl_q(\mathbf{r}; s)/ds)^2]$ as $T \to \infty$. From the weak law of large numbers, the coefficient of the second term on the right-hand side of Eq. 33 becomes

$$\sum_{t=1}^{T} \frac{d^2 l_q(\mathbf{r}(t); s)}{ds^2} = TE_p\left[\frac{d^2 l_q(\mathbf{r}; s)}{ds^2}\right]. \tag{34}$$

Taken together, we have

$$E_p\left[\left(\hat{s}_q - s\right)^2\right] = \frac{1}{T} \frac{E_p\left[\left(\frac{dl_q(\mathbf{r}; s)}{ds}\right)^2\right]}{\left(E_p\left[\frac{d^2 l_q(\mathbf{r}; s)}{ds^2}\right]\right)^2}. \tag{35}$$

If we regard the inverse of the squared decoding error as the Fisher information, the Fisher information for mismatched decoding model $q(\mathbf{r}; s)$ is

$$g^*(s) = T \frac{\left(E_p\left[\frac{d^2 l_q(\mathbf{r}; s)}{ds^2}\right]\right)^2}{E_p\left[\left(\frac{dl_q(\mathbf{r}; s)}{ds}\right)^2\right]}. \tag{36}$$

Note that when $q(\mathbf{r}; s) = p(\mathbf{r}; s)$, $g^*(s) = g(s)$.

INFORMATION GEOMETRIC INTERPRETATIONS

We discuss the inference of stimulus when neural responses are only partially observed and that when a mismatched probability distribution is used for decoding, from the information geometric viewpoint (Amari and Nagaoka, 2000). We consider the following four types of inference.

Inference 1 (Perfect observation and matched decoding) The complete data $(\bar{\mathbf{R}}_1, \bar{\mathbf{R}}_2)$ are available and the true probability distribution $p(\mathbf{r}; s)$ is used for decoding.

Inference 2 (Imperfect observation) The true probability distribution $p(\mathbf{r}; s)$ is used for decoding but only partial data $(\bar{\mathbf{R}}_1)$ are available.

Inference 3 (Mismatched decoding) The complete data $(\bar{\mathbf{R}}_1, \bar{\mathbf{R}}_2)$ are available but a mismatched probability distribution $q(\mathbf{r}; s)$ is used for decoding.

Inference 4 (Imperfect observation and mismatched decoding) A mismatched probability distribution $q(\mathbf{r}; s)$ is used for decoding and only partial data $(\bar{\mathbf{R}}_1)$ are available.

We assumed that both the true probability distribution $p(\mathbf{r}; s)$ and mismatched probability distributions $q(\mathbf{r}; s)$ belong to the exponential family of probability distributions S given in Eq. 1. S is specified by n-dimensional natural parameters $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)$. If we take $\boldsymbol{\theta}$ as a coordinate system introduced in set S of probability distributions, we can regard S as an n-dimensional manifold (space). A point in S represents a specific probability distribution determined by the parameters $\boldsymbol{\theta}$. The true statistical model $p(\mathbf{r}; s)$ denoted by M and the mismatched statistical model $q(\mathbf{r}; s)$ denoted

by M^*, both of which are parameterized by a single variable s, are considered as curves in the manifold S, i.e., one-dimensional submanifolds having a coordinate s.

Inference 1 First, we describe Inference 1 from the viewpoint of information geometry (Amari, 1982; Amari and Nagaoka, 2000). Let us denote the observed data $(\bar{\mathbf{R}}_1^o, \bar{\mathbf{R}}_2^o)$ by $\bar{\mathbf{x}}$. $\bar{\mathbf{x}}$ can be considered as a point in S, which we call the observed point. The observed point $\bar{\mathbf{x}}$ is distributed near the point s that represents the true probability distribution when stimulus s is presented, $p(\mathbf{r}; s)$. The deviation of $\bar{\mathbf{x}}$ from the point specified by the true stimulus parameter s can be decomposed into the deviation in the parallel direction to M and the deviation in the orthogonal direction to M. The maximum likelihood estimation corresponds to the minimizer of the Kullback–Leibler divergence between the distribution corresponding to the observed point $\bar{\mathbf{x}}$ (which is not in M in general) and distributions in M,

$$\hat{s} = \arg\min_s KL\left[\bar{\mathbf{x}}\left(\bar{\mathbf{R}}_1^o, \bar{\mathbf{R}}_2^o\right) \| p(\mathbf{r}; s) \right], \qquad (37)$$
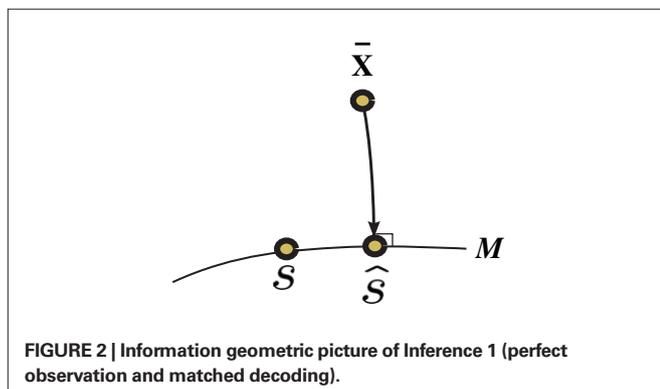
where $\hat{s}$ is the maximum likelihood estimator. The geometric interpretation of the maximum likelihood estimation is the orthogonal projection to M from $\bar{\mathbf{x}}$ (**Figure 2**). The orthogonal projection completely eliminates the deviation of $\bar{\mathbf{x}}$ from s in the orthogonal direction to M but the deviation in the parallel direction to M remains. This remaining deviation corresponds to the Fisher information of M (Eq. 18). If we use other estimators that are not orthogonal projections to M (e.g., the moment estimator), the decoding error necessarily becomes larger than the orthogonal projection.

Inference 2 In the Inference 2 case (Amari, 1995), only $\bar{\mathbf{R}}_1$ is observed. Let us define a submanifold D, which is formed by the set of observed points, where $\bar{\mathbf{R}}_1$ is fixed at the observed value $\bar{\mathbf{R}}_1^o$ but unobserved $\bar{\mathbf{R}}_2$ takes arbitrary values. Submanifold D is called the data submanifold. The maximum likelihood estimation based on partial observed data corresponds to searching for the pair of points $\hat{\mathbf{x}} \in D$ and $\hat{s} \in M$ that minimizes the Kullback–Leibler divergence between D and M (**Figure 3**), i.e.,

$$\min KL(D \| M) = \min_{\mathbf{x} \in D, s \in M} K(\hat{\mathbf{x}} \| s). \qquad (38)$$

The estimated value of s is expressed as

$$\hat{s}_d = \arg\min_s \left[\min_{\bar{R}_2} KL\left[\hat{\mathbf{x}}\left(\bar{R}_1^o, \bar{R}_2\right) \| p(\mathbf{r}; s) \right] \right]. \qquad (39)$$

For a given candidate point in D, $\hat{\mathbf{x}}$, the point in M that minimizes the Kullback–Leibler divergence is given by the orthogonal projection of $\hat{\mathbf{x}}$ to M.

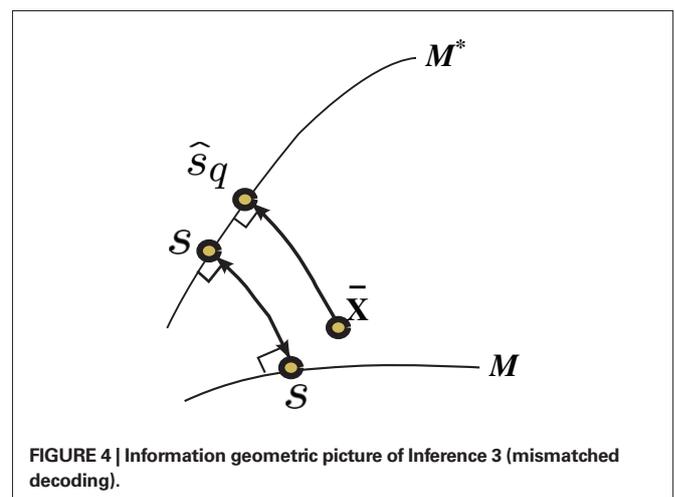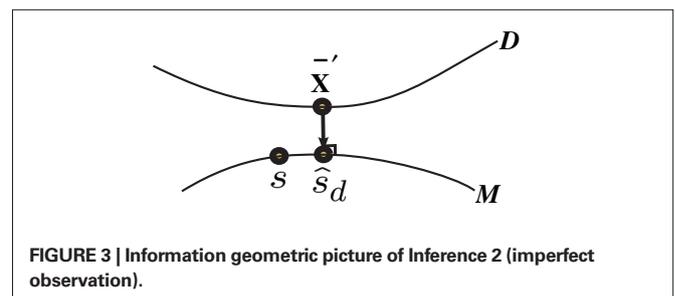**Inference 3** For Inference 3 (**Figure 4**), we assumed that the maximum likelihood estimation based on a mismatched model $q(\mathbf{r}; s)$ is unbiased (Eq. 29). This condition corresponds to the case where the point $p(\mathbf{r}; s)$ in M and the point $q(\mathbf{r}; s)$ in M^*, which both represent a given stimulus parameter s, are the mutually nearest points in S in terms of the Kullback–Leibler divergence, i.e.,

$$q(\mathbf{r}; s) = \arg\min_{q(\mathbf{r}; s')} KL\left[p(\mathbf{r}; s) \| q(\mathbf{r}; s') \right]. \qquad (40)$$

If we differentiate Eq. 40 with respect to s', we obtain Eq. 29. Similar to in Inference 1, the maximum likelihood estimation based on a mismatched model M^* corresponds to the minimization of the Kullback–Leibler divergence between the observed point $\bar{\mathbf{x}}$ and points in M^*,

$$\hat{s}_q = \arg\min_s KL\left[\bar{\mathbf{x}}\left(\bar{R}_1^o, \bar{R}_2^o\right) \| q(\mathbf{r}; s) \right]. \qquad (41)$$

This corresponds to the orthogonal projection of the observed point $\bar{\mathbf{x}}$ to M^* (**Figure 4**). The orthogonal projection to M^* cannot completely eliminate the deviation in the direction perpendicular to M, unless M and M^* are in parallel. Thus, information is inevitably lost depending on the angle between M and M^*. The Fisher information for the mismatched decoding model is given by Eq. 36.

FIGURE 3 | Information geometric picture of Inference 2 (imperfect observation).

FIGURE 2 | Information geometric picture of Inference 1 (perfect observation and matched decoding).

FIGURE 4 | Information geometric picture of Inference 3 (mismatched decoding).

Inference 4 In Inference 4 (**Figure 5**), the maximum likelihood estimation with partial observed data $\bar{R}_1^o$ and a mismatched probability distribution $q(\mathbf{r};s)$ corresponds to searching for two points in the data submanifold D and the mismatched model M^*:

$$\hat{s}_{qd} = \arg\min_s \left[\min_{\bar{R}_2} KL\left[\bar{\mathbf{x}}\left(\bar{R}_1^o, \bar{R}_2\right) \| q(\mathbf{r};s) \right] \right]. \quad (42)$$

RELATIONSHIP BETWEEN INFERENCE WITH PARTIAL OBSERVED DATA AND INFERENCE WITH MISMATCHED DECODING MODEL: GAUSSIAN CASE

In this section, we compute the Fisher information obtained by the four types of inference described in the previous section when the probability distributions are Gaussian. We also discuss the relationship between the inferences.

ONE-DIMENSIONAL CASE

Before we deal with the multidimensional Gaussian model, we first consider the one-dimensional case as a toy example. We specifically consider the Gaussian distribution with mean $\mu(s) = s$ and variance $\sigma^2(s) = s^2$:

$$p(r) = \frac{1}{\sqrt{2\pi s^2}} \exp\left(-\frac{1}{2s^2}(r-s)^2 \right). \quad (43)$$
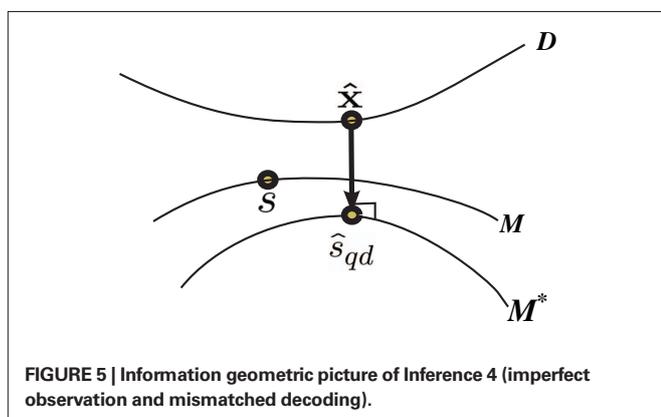
The statistical model $M = \{N(s, s^2)\}$ is expressed as a curve in the manifold $S = \{N(\mu, \sigma^2)\}$ with coordinates of mean μ and variance σ^2 (**Figure 6**). The probability distribution on T observations of r is given by

$$p(\bar{r}, \bar{R}) = \exp\left(T\left(\frac{\bar{r}^2}{s} - \frac{\bar{R}}{2s^2} - \log\left(\sqrt{2\pi s^2}\right) - \frac{1}{2} \right) \right), \quad (44)$$

where $\bar{r}$ and $\bar{R}$ are sufficient statistics

$$\bar{r} = \frac{1}{T} \sum_{t=1}^{T} r(t), \quad (45)$$

$$\bar{R} = \frac{1}{T} \sum_{t=1}^{T} r^2(t). \quad (46)$$

In this model, we compute the Fisher information and the information loss when the data of variance $\bar{R}$ is lost and those when the decoding model whose variance is mismatched with the actual one is used.

Inference 1 When the data of neural responses $\bar{\mathbf{x}} = (\bar{r}, \bar{R})$ are completely observed and the actual statistical model M is used in decoding, the maximum likelihood estimation of s corresponds to the orthogonal projection from $\bar{\mathbf{x}}$ to M (**Figure 6**). In this case, we can compute the Fisher information as

$$I_{F1}\left(s; \bar{r}, \bar{R}\right) = \frac{3T}{s^2}. \quad (47)$$

Inference 2 When data of variance $\bar{R}$ is lost, the data manifold D is given by

$$D = \left\{ N(\mu, \sigma^2) \mid \mu = \bar{r}, \quad \sigma^2 \text{ is arbitrary} \right\}. \quad (48)$$

In this case, the estimated value of s is the intersection point of D and M (**Figure 6**). By using Eq. 27, we can obtain the Fisher information,

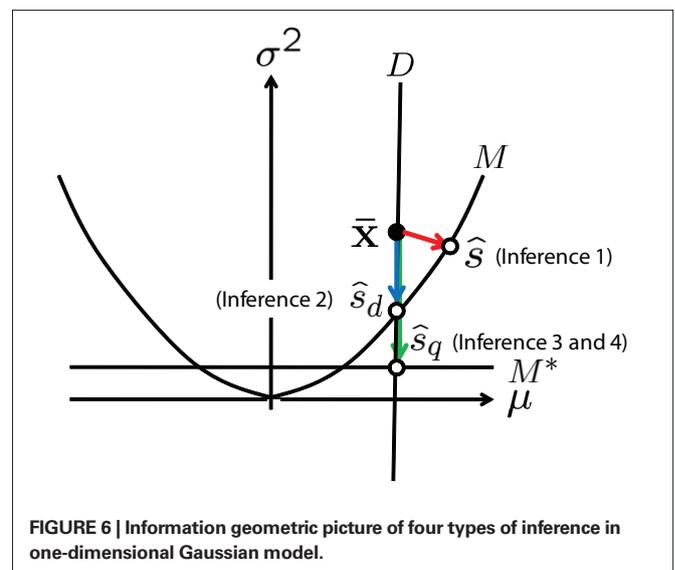$$I_{F2}(s; \bar{r}) = \frac{T+2}{s^2} \sim \frac{T}{s^2}, \quad (49)$$

where we used the fact that the marginal distribution over $\bar{R}$ can be written as

$$p(\bar{r}) = \frac{1}{\sqrt{2\pi\left(\frac{s}{\sqrt{T}}\right)^2}} \exp\left(-\frac{1}{2\left(\frac{s}{\sqrt{T}}\right)^2}(\bar{r} - s)^2 \right). \quad (50)$$

$p(\bar{r})$ can be derived by considering that $\bar{r}$ also obey a Gaussian distribution and that the mean and the variance of $\bar{r}$ are s and s^2/T, respectively.

The information loss is given by

$$\Delta I_{F2}(s) = \frac{2T}{s^2}. \quad (51)$$

FIGURE 5 | Information geometric picture of Inference 4 (imperfect observation and mismatched decoding).

FIGURE 6 | Information geometric picture of four types of inference in one-dimensional Gaussian model.

Inference 3 We specifically consider the inference with the following mismatched decoding model to compare it with the inference when $\bar{R}$ is lost:

$$q(r) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(r-s)^2\right). \quad (52)$$

In this model, the mean is equal to the actual one but the variance is mismatched with the actual one. The maximum likelihood estimation based on the mismatched decoding model M^* corresponds to the orthogonal projection from the observed point $\bar{\mathbf{x}}$ to M^* (**Figure 6**). By using Eq. 36, we obtain the Fisher information

$$I_{F3}(s) = \frac{T}{s^2}. \quad (53)$$

Inference 4 When the mismatched decoding model $q(r)$, where the variance is independent of s, is used for decoding, the data of variance $\bar{R}$ does not affect the results of the inference. Thus, even if $\bar{R}$ is lost, no information is lost in this mismatched decoding. The Fisher information in the Inference 4 case is the same as that in the Inference 3 case:

$$I_{F4}(s) = I_{F3}(s). \quad (54)$$

As Eqs. 49 and 53 show, $I_{F3}(s)$ is equal to $I_{F2}(s)$. We can also easily show that $I_{F3}(s)$ is equal to $I_{F2}(s)$ in one-dimensional cases in general. However, in the multidimensional case, $I_{F3}(s)$ is not equal to $I_{F2}(s)$. In the next section, we explain the general relationship between $I_{F2}(s)$ and $I_{F3}(s)$ in the multidimensional Gaussian model.

MULTIDIMENSIONAL CASE

We next consider the multidimensional Gaussian distribution shown in Eq. 2. The probability distribution for T observations of neural responses $\mathbf{r}$ is given by

$$p(\bar{\mathbf{r}}, \bar{R}) = \exp\left(T(\boldsymbol{\theta}(s)\cdot\bar{\mathbf{r}} + \Theta(s)\cdot\bar{R} - \psi(s))\right), \quad (55)$$

where the sufficient statistics $\bar{\mathbf{r}}$ and $\bar{R}$ are

$$\bar{\mathbf{r}} = \frac{1}{T}\sum_{t=1}^{T} r(t), \quad (56)$$

$$\bar{\mathbf{R}} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{r}(t)\mathbf{r}(t)^T, \quad (57)$$

the natural parameters $\boldsymbol{\theta}$ and Θ are

$$\boldsymbol{\theta}(s) = \mathbf{C}^{-1}(s)\mathbf{f}(s), \quad (58)$$

$$\Theta(s) = -\frac{1}{2}\mathbf{C}^{-1}(s), \quad (59)$$

and the normalization constant $\Psi(s)$ is

$$\Psi(s) = \frac{1}{2}\mathbf{f}(s)\mathbf{C}^{-1}(s)\mathbf{f}(s) + \log\left(\sqrt{\det[\mathbf{C}]}\right) + \log\left(\sqrt{(2\pi)^N}\right). \quad (60)$$

Inference 1 First, let us consider the Fisher information in the Inference 1 case. The Fisher information matrix with respect to the natural parameters is given by

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}}, \bar{R})}{\partial\theta_i\partial\theta_j} = TC_{ij}, \quad (61)$$

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}}, \bar{R})}{\partial\theta_i\partial\Theta_{jk}} = T\left(C_{ji}f_k + C_{ki}f_j\right), \quad (62)$$

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}}, \bar{R})}{\partial\Theta_{ij}\partial\Theta_{kl}} = T\left(2C_{ik}f_lf_j + 2C_{jk}f_lf_i + 2C_{ik}C_{lj}\right). \quad (63)$$

By using the Fisher information matrix with respect to the natural parameters, we can obtain the Fisher information with respect to stimulus s from Eq. 25:

$$\frac{I_{F1}(s;\bar{\mathbf{r}}, \bar{R})}{T} = \mathbf{f}'^T\mathbf{C}^{-1}\mathbf{f}' + \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}'\mathbf{C}^{-1}\mathbf{C}'\mathbf{C}^{-1}\right]. \quad (64)$$

Inference 2 Second, let us consider the Fisher information in the Inference 2 case. We consider the situation that the second-order sufficient statistics $\bar{R}$ in Eq. 55 are lost and only the first-order sufficient statistics $\bar{\mathbf{r}}$ are observed. The marginalized distribution over missing data $\bar{R}$ is given by

$$p(\bar{\mathbf{r}})$$
$$= \frac{1}{\sqrt{(2\pi)^N \det[\mathbf{C}(s)/T]}}\exp\left(-\frac{T}{2}(\bar{\mathbf{r}}-\mathbf{f}(s))^T\mathbf{C}^{-1}(s)(\bar{\mathbf{r}}-\mathbf{f}(s))\right), \quad (65)$$

$$= \exp\left(T(\boldsymbol{\theta}(s)\cdot\bar{\mathbf{r}} + \Theta(s)\cdot\overline{\mathbf{r}\mathbf{r}}^T - \Psi(s))\right), \quad (66)$$

where the natural parameters $\boldsymbol{\theta}$ and Θ are

$$\boldsymbol{\theta}(s) = \mathbf{C}^{-1}(s)\mathbf{f}(s), \quad (67)$$

$$\Theta(s) = -\frac{1}{2}\mathbf{C}^{-1}(s), \quad (68)$$

and the normalization constant $\Psi(s)$ is

$$\Psi(s) = \frac{1}{2}\mathbf{f}(s)\mathbf{C}^{-1}(s)\mathbf{f}(s) + \frac{1}{T}\log\left(\sqrt{\det[\mathbf{C}]}\right) + \frac{1}{T}\log\sqrt{\left(\frac{2\pi}{T}\right)^N},$$

$$\sim \frac{1}{2}\mathbf{f}(s)\mathbf{C}^{-1}(s)\mathbf{f}(s), \quad (69)$$

where we ignored the terms of the order of $1/T$ in the limit of $T \to \infty$. In this case, the Fisher information matrix with respect to the natural parameters is computed as follows:

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}})}{\partial\theta_i\partial\theta_j} = TC_{ij}, \quad (70)$$

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}})}{\partial\theta_i\partial\Theta_{jk}} = T\left(C_{ji}f_k + C_{ki}f_j\right), \quad (71)$$

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}})}{\partial\Theta_{ij}\partial\Theta_{kl}} = T\left(2C_{ik}f_lf_j + 2C_{jk}f_lf_i\right). \quad (72)$$

If we compare the components of the Fisher information matrix when $\overline{\mathbf{R}}$ is missing with those when the data are complete, the information loss due to the missing data is seen to be represented in the components $\frac{\partial^2 \log p(\overline{\mathbf{r}}, \overline{\mathbf{R}})}{\partial \Theta_{ij} \partial \Theta_{kl}}$ (Eqs. 63 and 72). By using the Fisher information matrix with respect to the natural parameters, we can compute the Fisher information with respect to stimulus s as

$$\frac{I_{F2}(s; \overline{\mathbf{r}})}{T} = \mathbf{f}'^{T} \mathbf{C}^{-1} \mathbf{f}'. \tag{73}$$

From Eqs. 64 and 73, we find that the information loss due to the missing data $\overline{\mathbf{R}}$ is

$$\frac{\Delta I_{F2}(s)}{T} = \frac{1}{2} \mathrm{Tr} \left[\mathbf{C}' \mathbf{C}^{-1} \mathbf{C}' \mathbf{C}^{-1} \right]. \tag{74}$$

This information loss solely depends on $\mathbf{C}' \mathbf{C}^{-1}$ and is always positive.

Inference 3 Third, let us consider the Fisher information in the Inference 3 case. In the Inference 2 case, we considered that the second-order sufficient statistics $\overline{\mathbf{R}}$, which are the variance and covariance data of neural responses, are lost. A mismatched probability distribution $q(\mathbf{r}; s)$ that is comparable with the inference when $\overline{\mathbf{R}}$ is lost would be that the mean in $q(\mathbf{r}; s)$ is the same as that in $p(\mathbf{r}; s)$ but the covariance matrix in $q(\mathbf{r}; s)$ does not match the true covariance matrix in $p(\mathbf{r}; s)$. As a simple example, we assume that the covariance matrix in the mismatched probability distribution is a constant matrix $\mathbf{C}_q$ that is independent of s:

$$q(\mathbf{r}; s) = \frac{1}{\sqrt{(2\pi)^N \det[\mathbf{C}_q]}} \exp\left(-\frac{1}{2} (\mathbf{r} - \mathbf{f}(s))^T \mathbf{C}_q^{-1} (\mathbf{r} - \mathbf{f}(s)) \right). \tag{75}$$

In this case, we can show that the maximum likelihood estimation based on the mismatched probability distribution is consistent, i.e., the condition shown in Eq. 29 is satisfied, as follows:

$$E_p[\partial_s l_q(\mathbf{r}; s)] = E_p\left[\mathbf{r}^T \mathbf{C}_q^{-1} \frac{\partial \mathbf{f}}{\partial s} - \mathbf{f}^T \mathbf{C}_q^{-1} \frac{\partial \mathbf{f}}{\partial s} \right],$$

$$= 0. \tag{76}$$

By using Eq. 36, we obtain the Fisher information for the mismatched model $q(\mathbf{r}; s)$:

$$\frac{I_{F3}(s)}{T} = \frac{\left(\mathbf{f}'^{T} \mathbf{C}_q^{-1} \mathbf{f}' \right)^2}{\mathbf{f}'^{T} \mathbf{C}_q^{-1} \mathbf{C} \mathbf{C}_q^{-1} \mathbf{f}'}. \tag{77}$$

Inference 4 and comparison Finally, we consider the Fisher information in the Inference 4 case and compare the four types of inference described above. It is obvious that the Fisher information obtained by Inference 1, I_{F1}, is the largest and the Fisher information obtained by Inference 4, I_{F4}, is the smallest. On the other hand, the relationship between the Fisher information obtained by Inference 2, I_{F2}, and that obtained by Inference 3, I_{F3}, is not clear, i.e.,

$$I_{F1} > I_{F2}, I_{F3} > I_{F4}. \tag{78}$$

However, the relationship between I_{F2} and I_{F3} can be clarified by considering the Fisher information obtained by Inference 4, I_{F4}. We assumed that a mismatched model that is related to the loss of

covariance data, $\overline{\mathbf{R}}$, is the statistical model whose covariance matrix is a constant matrix. In this case, the vector of natural parameters Θ in Eq. 55, which is coupled with $\overline{\mathbf{R}}$, does not depend on s. Thus, when we use a mismatched model $q(\mathbf{r}; s)$ whose covariance matrix is independent of s, the inference does not change even if the data about covariance $\overline{\mathbf{R}}$ are lost. Thus, Inferences 3 and 4 result in the same estimate of s and the same Fisher information:

$$I_{F4}(s) = I_{F3}(s). \tag{79}$$

To summarize, the relationship between the Fisher information in each of the four inference cases is

$$I_{F1} > I_{F2} > I_{F3} = I_{F4}. \tag{80}$$

INFORMATION LOSS IN LOG-LINEAR MODEL OF BINARY NEURAL RESPONSE

In this section, we evaluate the information loss associated with loss of data and mismatched decoding in the log-linear model of binary neural response.

TWO-NEURON MODEL

As the simplest example, we first consider the two-neuron model,

$$p(r_1, r_2) = \exp(\theta_1(s) r_1 + \theta_2(s) r_2 + \Theta(s) r_1 r_2 - \log Z(s)), \tag{81}$$

where $Z(s)$ is a normalization constant,

$$Z(s) = \sum_{r_1, r_2 = 0, 1} \exp\left(\sum_{i=1}^{2} \theta_i(s) r_i + \Theta(s) r_1 r_2 \right). \tag{82}$$

The probability distribution for T observations of neural responses $\mathbf{r}$ can be written as

$$p(\overline{r}_1, \overline{r}_2, \overline{R}_{12}) = W(\overline{r}_1, \overline{r}_2, \overline{R}_{12}) \exp\left(T(\theta_1 \overline{r}_1 + \theta_2 \overline{r}_2 + \Theta \overline{R}_{12} - \log Z) \right),$$

$$= \exp\left(T\left(\theta_1 \overline{r}_1 + \theta_2 \overline{r}_2 + \Theta \overline{R}_{12} + \frac{\log W(\overline{r}_1, \overline{r}_2, \overline{R}_{12})}{T} - \log Z \right) \right), \tag{83}$$

where $\overline{r}_1$, $\overline{r}_2$, and $\overline{R}_{12}$ are sufficient statistics,

$$\overline{r}_1 = \frac{1}{T} \sum_{t=1}^{T} r_1^{(t)}, \tag{84}$$

$$\overline{r}_2 = \frac{1}{T} \sum_{t=1}^{T} r_2^{(t)}, \tag{85}$$

$$\overline{R}_{12} = \frac{1}{T} \sum_{t=1}^{T} r_1^{(t)} r_2^{(t)}, \tag{86}$$

and $W(\overline{r}_1, \overline{r}_2, \overline{R}_{12})$ is the number of configurations of $(\mathbf{r}^{(1)}, \mathbf{r}^{(2)}, \ldots, \mathbf{r}^{(T)})$ where the sufficient statistics take the specific values $\overline{r}_1$, $\overline{r}_2$, and $\overline{R}_{12}$. $W(\overline{r}_1, \overline{r}_2, \overline{R}_{12})$ can be expressed as

$$W(\overline{r}_1, \overline{r}_2, \overline{R}_{12})$$

$$= \frac{T!}{(T\overline{R}_{12})! \left(T(\overline{r}_1 - \overline{R}_{12}) \right)! \left(T(\overline{r}_2 - \overline{R}_{12}) \right)! \left(T(1 - \overline{r}_1 - \overline{r}_2 + \overline{R}_{12}) \right)!}. \tag{87}$$

In the limit of $T \to \infty$, by using Stirling's formula, $W(\overline{r}_1, \overline{r}_2, \overline{R}_{12})$ can be approximated as

$$
\begin{aligned}
\log W = -T\Big(&\overline{R}_{12} \log \overline{R}_{12} + (\overline{r}_1 - \overline{R}_{12}) \log (\overline{r}_1 - \overline{R}_{12}) \\
&+ (\overline{r}_2 - \overline{R}_{12}) \log (\overline{r}_2 - \overline{R}_{12})) \\
&+ (1 - \overline{r}_1 - \overline{r}_2 + \overline{R}_{12}) \log (1 - \overline{r}_1 - \overline{r}_2 + \overline{R}_{12})\Big).
\end{aligned} \quad (88)
$$

Inference 1 We first compute the Fisher information in the Inference 1 case. The Fisher information matrix with respect to the natural parameters is given by

$$
-\frac{\partial^2 \log p(\overline{r}_1, \overline{r}_2, \overline{R}_{12})}{\partial \theta_i \partial \theta_j} = T(\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle), \quad (89)
$$

$$
-\frac{\partial^2 \log p(\overline{r}_1, \overline{r}_2, \overline{R}_{12})}{\partial \Theta^2} = T(\langle r_1 r_2 \rangle - \langle r_1 r_2 \rangle^2), \quad (90)
$$

$$
-\frac{\partial^2 \log p(\overline{r}_1, \overline{r}_2, \overline{R}_{12})}{\partial \theta_i \partial \Theta} = T(\langle r_1 r_2 \rangle - \langle r_i \rangle \langle r_1 r_2 \rangle), \quad (91)
$$

where $\langle x \rangle = E_p[x] = \sum_r x p(r_1, r_2)$. By using the Fisher information matrix with respect to the natural parameters, we can obtain the Fisher information with respect to stimulus s from Eq. 25:

$$
\begin{aligned}
&\frac{I_{F1}(s; \overline{r}_1, \overline{r}_2, \overline{R}_{12})}{T} \\
&= \sum_{i=1}^{2} \langle r_i \rangle (1 - \langle r_i \rangle) \left(\frac{\partial \theta_i}{\partial s}\right)^2 + \langle r_1 r_2 \rangle (1 - \langle r_1 r_2 \rangle) \left(\frac{\partial \Theta}{\partial s}\right)^2 \\
&+ 2(\langle r_1 r_2 \rangle - \langle r_1 \rangle \langle r_2 \rangle) \frac{\partial \theta_1}{\partial s} \frac{\partial \theta_2}{\partial s} + 2 \sum_{i=1}^{2} \langle r_1 r_2 \rangle (1 - \langle r_i \rangle) \frac{\partial \theta_i}{\partial s} \frac{\partial \Theta}{\partial s}.
\end{aligned} \quad (92)
$$

Inference 2 We next compute the Fisher information when the data of neural correlation $\overline{R}_{12}$ are lost. When T is finite, it is difficult to marginalize the probability distribution $p(\overline{r}_1, \overline{r}_2, \overline{R}_{12})$ over $\overline{R}_{12}$ because there are many possible $\overline{R}_{12}$ when specific values of $\overline{r}_1$ and $\overline{r}_2$ are given. However, in the $T \to \infty$ case, we only need to consider the most probable $\overline{R}_{12}$ when $\overline{r}_1$ and $\overline{r}_2$ are given. By differentiating the argument of the exponential function in Eq. 83 with respect to $\overline{R}_{12}$, we can obtain the equation for the most probable $\overline{R}_{12}$:

$$
\Theta - \log \frac{\overline{R}_{12}(1 - \overline{r}_1 - \overline{r}_2 + \overline{R}_{12})}{(\overline{r}_1 - \overline{R}_{12})(\overline{r}_2 - \overline{R}_{12})} = 0. \quad (93)
$$

We denote the solution of this equation by $\overline{R}_{12}^*$. Note that $\overline{R}_{12}^*$ depends on Θ but does not depend on θ_1 and θ_2. The marginalized probability distribution is written as

$$
\begin{aligned}
&p(\overline{r}_1, \overline{r}_2) \\
&= \exp\left(T\left(\theta_1 \overline{r}_1 + \theta_2 \overline{r}_2 + \Theta \overline{R}_{12}^* + \frac{\log W(\overline{r}_1, \overline{r}_2, \overline{R}_{12}^*)}{T} - \log Z \right) \right).
\end{aligned} \quad (94)
$$

By using Eq. 94, we can compute the Fisher information when $\overline{R}_{12}$ is lost as

$$
\frac{I_{F2}(s; \overline{r}_1, \overline{r}_2)}{T} = \frac{I_{F1}(s; \overline{r}_1, \overline{r}_2, \overline{R}_{12})}{T}
$$

$$
- E_{p(\overline{r}_1, \overline{r}_2)} \left[\frac{\partial^2}{\partial \Theta^2} \left(\Theta \overline{R}_{12}^* + \frac{1}{T} \log W(\overline{r}_1, \overline{r}_2, \overline{R}_{12}^*) \right) \left(\frac{\partial \Theta}{\partial s}\right)^2, \right] \quad (95)
$$

where we used the fact that $\overline{R}_{12}^*$ only depends on Θ. The information loss is given by

$$
\begin{aligned}
\frac{\Delta I_{F2}(s)}{T} &= E_{p(\overline{r}_1, \overline{r}_2)} \left[\frac{\partial^2}{\partial \Theta^2} \left(\Theta \overline{R}_{12}^* + \frac{1}{T} \log W(\overline{r}_1, \overline{r}_2, \overline{R}_{12}^*) \right) \right] d\left(\frac{\partial \Theta}{\partial s}\right)^2, \\
&= \left(\frac{1}{\langle r_1, r_2 \rangle} + \frac{1}{\langle r_1 \rangle - \langle r_1 r_2 \rangle} + \frac{1}{\langle r_2 \rangle - \langle r_1 r_2 \rangle} \right. \\
&\left. + \frac{1}{1 - \langle r_1 \rangle - \langle r_2 \rangle + \langle r_1 r_2 \rangle} \left(\frac{\partial \Theta}{\partial s}\right)^2 \right).
\end{aligned} \quad (96)
$$

This information loss only depends on $\partial \Theta / \partial s$. Thus, when Θ is independent of s, there is no information loss even if data $\overline{R}_{12}$ are lost.

Inference 3 We next compute the Fisher information when neural correlation is ignored in decoding. We consider the mismatched decoding model $q(r_1, r_2)$ that is the product of the marginal probability distributions of the actual distribution $p(r_1, r_2)$:

$$
q(r_1, r_2; s) = \prod_{i=1}^{2} p(r_i),
$$

$$
= \prod_{i=1}^{2} \exp(\theta_i^D(s) r_i - \log Z_i^D(s)), \quad (97)
$$

where $p(r_i)$ is the marginalized distribution over r_j, i.e., $p(r_i) = \sum_{r_j} p(r_i, r_j)$, and the normalization constant $Z_i^D(s)$ is given by

$$
Z_i^D(s) = 1 + \exp(\theta_i^D(s)). \quad (98)
$$

From the definition of $q(r_1, r_2)$, the averaged values of r_1 and r_2 over the mismatched model $q(r_1, r_2)$ are equal to those over the actual model $p(r_1, r_2)$. Thus, the following relationship holds between $\theta_i^D(s)$ and the natural parameters in the actual probability distribution $p(r_1, r_2)$:

$$
\langle r_i \rangle = \frac{1}{Z(s)} \sum_{r_1, r_2} r_i \exp(\theta_1 r_1 + \theta_2 r_2 + \Theta r_1 r_2),
$$

$$
= \frac{\exp(\theta_i^D(s))}{1 + \exp(\theta_i^D(s))}. \quad \text{(for } i = 1, 2\text{)} \quad (99)
$$

The maximum likelihood estimation based on this mismatched decoding model $q(r_1, r_2)$ is shown to be consistent as follows:

$$
E_p \left[\frac{\partial \log q(r_1, r_2)}{\partial s} \right] = E_p \left[\sum_{i=1}^{2} \frac{\partial \theta_i^D(s)}{\partial s} \left(r_i - \frac{\exp(\theta_i^D(s))}{1 + \exp(\theta_i^D(s))} \right) \right], \quad (100)
$$

$$
= 0. \quad (101)
$$

When the estimation by a mismatched decoding model is consistent, the Fisher information obtained by the mismatched decoding model can be computed by Eq. 36. The Fisher information is given by

$$
\frac{I_{F3}(s)}{T} = \frac{\left(\sum_{i=1}^{2}\left(\frac{\partial\theta_i^D}{\partial s}\right)^2\langle r_i\rangle(1-\langle r_i\rangle)\right)^2}{\sum_{i=1}^{2}\left(\frac{\partial\theta_i^D}{\partial s}\right)^2\langle r_i\rangle(1-\langle r_i\rangle)+2\frac{\partial\theta_1^D}{\partial s}\frac{\partial\theta_2^D}{\partial s}(\langle r_1 r_2\rangle-\langle r_1\rangle\langle r_2\rangle)}, \quad (102)
$$

where

$$
\frac{\partial\theta_i^D}{\partial s} = \frac{\partial\theta_i}{\partial s}\langle r_i\rangle(1-\langle r_i\rangle)+\frac{\partial\theta_j}{\partial s}(\langle r_1 r_2\rangle-\langle r_1\rangle\langle r_2\rangle)
$$
$$
+\frac{\partial\Theta}{\partial s}\langle r_1 r_2\rangle(1-\langle r_i\rangle). \quad \text{(for } i,j=1,2) \quad (103)
$$

As discussed in the previous section, $I_{F3}(s)$ is always smaller than $I_{F2}(s)$. Although the information loss associated with the loss of data $\bar{R}$ only depends on $\partial\Theta/\partial s$, the information loss associated with ignoring correlation in decoding depends not only on $\partial\Theta/\partial s$ but also on $\partial\theta_1/\partial s$ and $\partial\theta_2/\partial s$. Thus, when neural correlation is ignored in decoding, the information is lost even if Θ is independent of s.

As a special case, if $\theta_1=\theta_2$, which means $\langle r_1\rangle=\langle r_2\rangle=\langle r\rangle$, the information loss only depends on $\partial\Theta/\partial s$ and is given by

$$
\frac{\Delta I_{F3}(s)}{T} = \frac{\langle r_1 r_2\rangle(\langle r\rangle-\langle r_1 r_2\rangle)(1-2\langle r\rangle+\langle r_1 r_2\rangle)}{\langle r\rangle-2\langle r\rangle^2+\langle r_1 r_2\rangle}\left(\frac{\partial\Theta}{\partial s}\right)^2. \quad (104)
$$

In this case, $\Delta I_{F3}(s)$ is equal to $\Delta I_{F2}(s)$ (Eq. 96). This is because if $\theta_1=\theta_2=\theta$, only two parameters, namely θ and Θ, are in the statistical model. This is the same situation as in the one-dimensional Gaussian case illustrated in **Figure 6**, where $\Delta I_{F3}(s)$ is also equal to $\Delta I_{F2}(s)$.

HOMOGENEOUS N NEURON MODEL

We next consider the case with a large number of neurons. In this case, the Fisher information cannot be analytically computed in general. To restrict ourselves to dealing with an analytically tractable model, we here only deal with a probabilistic model of a homogeneous neural population. In this model, $\theta_i(s)=\theta(s)$ for any i and $\theta_{ij}(s)=\Theta(s)$ for any pair of i and j, i.e.,

$$
p(\mathbf{r};s) = \exp\left(\theta(s)\sum_{i=1}^{N}r_i+\frac{\Theta(s)}{N}\sum_{i<j}r_i r_j-\log Z(s)\right), \quad (105)
$$

where the normalization constant $Z(s)$ is given by

$$
Z(s) = \text{Tr}\exp\left(\theta(s)\sum_{i=1}^{N}r_i+\frac{\Theta(s)}{N}\sum_{i<j}r_i r_j\right), \quad (106)
$$

where Tr stands for the sum over all possible combinations of the neuron state variables $(r_1, r_2,\ldots,r_N)$. The probability distribution of T observations is given by

$$
p(\bar{r},\mathbf{R}) = \exp\left(T\left(\theta(s)\sum_{i=1}^{N}\bar{r}_i+\frac{\Theta(s)}{N}\sum_{i<j}\bar{R}_{ij}-\log Z(s)\right)\right), \quad (107)
$$

where $\bar{r}=\{\bar{r}_i\}$ and $\mathbf{R}=\{\bar{R}_{ij}\}$ are sufficient statistics,

$$
\bar{r}_i = \frac{1}{T}\sum_{t=1}^{T}r_i(t), \quad (108)
$$

$$
\bar{R}_{ij} = \frac{1}{T}\sum_{t=1}^{T}r_i(t)r_j(t). \quad (109)
$$

Inference 1 We first compute the Fisher information when data are complete and the decoding is optimal. As Eq. 23 shows, the Fisher information can be computed if we evaluate $\log Z$. For analytical tractability, we consider the limit of $N\to\infty$. In this case, Z can be calculated as

$$
Z = \text{Tr}\exp\left(\theta\sum_i r_i+\frac{\Theta}{N}\sum_{i<j}r_i r_j\right),
$$
$$
= \text{Tr}\exp\left(\left(\theta-\frac{\Theta}{2N}\right)\sum_i r_i+\frac{\Theta}{2N}\left(\sum_i r_i\right)^2\right),
$$
$$
= \int_{-\infty}^{\infty}dm\exp\left(N\left(\frac{\Theta}{2}m^2+\theta m\right)+\log W(m)\right), \quad (110)
$$

where m is

$$
m = \frac{1}{N}\sum_{i=1}^{N}r_i \quad (111)
$$

and $W(m)$ is the number of states where m takes a certain value, which is given by

$$
W(m) = \frac{N!}{(Nm)!(N(1-m))!}. \quad (112)
$$

In the limit of $N\to\infty$, by using Stirling's formula, $W(m)$ can be approximated as

$$
\log W(m) = -N(m\log m+(1-m)\log(1-m)). \quad (113)
$$

We denote the argument of the exponential function in Eq. 110 by F, where

$$
F = N\left(\frac{\Theta}{2}m^2+\theta m-m\log m-(1-m)\log(1-m)\right). \quad (114)
$$

In the limit of $N\to\infty$, the integral in Eq. 110 can be approximated as

$$
\int_{-\infty}^{\infty}dm\exp(F) = \exp(F^*), \quad (115)
$$

where F^* is the maximum of the function F. From $\partial F/\partial m=0$, the value of m that maximizes the function F is the solution of the self-consistent equation

$$
m = \frac{\exp(\theta+\Theta m)}{1+\exp(\theta+\Theta m)}. \quad (116)
$$

We set the solution of Eq. 116 to $m^*(\theta,\Theta)$. The Fisher information matrix with respect to the natural parameters θ and Θ is given by

$$
-\frac{\partial^2 F^*}{\partial\theta^2} = N\frac{m^*(1-m^*)}{1-\Theta m^*(1-m^*)}, \quad (117)
$$

$$-\frac{\partial^2 F^*}{\partial\theta\partial\Theta} = N\frac{m^{*2}(1-m^*)}{1-\Theta m^*(1-m^*)}, \qquad (118)$$

$$-\frac{\partial^2 F^*}{\partial\Theta^2} = N\frac{m^{*3}(1-m^*)}{1-\Theta m^*(1-m^*)}. \qquad (119)$$

By using the Fisher information matrix, we obtain the Fisher information with respect to stimulus s from Eq. 25:

$$\frac{I_{F1}(s)}{N} = \frac{m^*(1-m^*)}{1-\Theta m^*(1-m^*)}\left(\frac{\partial\theta}{\partial s}\right)^2 + 2\frac{m^{*2}(1-m^*)}{1-\Theta m^*(1-m^*)}\frac{\partial\theta}{\partial s}\frac{\partial\Theta}{\partial s}$$
$$+ \frac{m^{*3}(1-m^*)}{1-\Theta m^*(1-m^*)}\left(\frac{\partial\Theta}{\partial s}\right)^2. \qquad (120)$$

Inferences 2 and 3 We next consider the situation where the data of correlation, $\bar{R}$, are lost and a mismatched decoding model that ignores neural correlation $q(\mathbf{r}; s)$ is used. We first consider the inference with a mismatched decoding model (Inference 3). This model $q(\mathbf{r}; s)$ is defined from the actual probability distribution $p(\mathbf{r}; s)$ as follows:

$$q(\mathbf{r};s) = \prod_{i=1}^{N} p(r_i),$$

$$= \prod_{i=1}^{N}\exp(\theta_D(s)r_i - \log Z_D(s)), \qquad (121)$$

where the normalization constant Z_D is given by

$$Z_D(s) = 1 + \exp(\theta_D(s)). \qquad (122)$$

The relationship between m^* in the previous section and θ_D is

$$m^* = \frac{\exp(\theta_D)}{1+\exp(\theta_D)}. \qquad (123)$$

By comparing Eq. 123 with Eq. 116, we obtain

$$\theta_D = \theta + \Theta m^*. \qquad (124)$$

Similar to in the previous section, the maximum likelihood estimation with the independent decoding model $q(\mathbf{r}; s)$ can be shown to be consistent as follows:

$$E_p[\partial_s l_q(\mathbf{r};s)] = E_p\left[\left(\sum_{i=1}^{N} r_i\right)\frac{\partial\theta_D}{\partial s} - N\frac{\exp(\theta_D)}{1+\exp(\theta_D)}\frac{\partial\theta_D}{\partial s}\right],$$
$$= 0. \qquad (125)$$

By using Eq. 36, we can compute the Fisher information obtained by the independent decoding model as

$$\frac{I_{F3}(s)}{N} = \frac{m^*(1-m^*)}{1-\Theta m^*(1-m^*)}\left(\frac{\partial\theta}{\partial s}\right)^2 + 2\frac{m^{*2}(1-m^*)}{1-\Theta m^*(1-m^*)}\frac{\partial\theta}{\partial s}\frac{\partial\Theta}{\partial s}$$
$$+ \frac{m^{*3}(1-m^*)}{1-\Theta m^*(1-m^*)}\left(\frac{\partial\Theta}{\partial s}\right)^2. \qquad (126)$$

Since $I_{F1}(s) = I_{F3}(s)$ (see Eqs. 120 and 126), there is no information loss when the independent model is used for decoding in a homogeneous neural population (Wu et al., 2001). As for the Gaussian model, the Fisher information when the data of neural correlation $\bar{R}$ are lost, I_{F2}, is larger than I_{F3} and is smaller than I_{F1}. Since $I_{F1}(s) = I_{F3}(s)$ in this case, $I_{F2}(s)$ is also equal to $I_{F1}(s)$. Thus,

$$I_{F1}(s) = I_{F2}(s) = I_{F3}(s). \qquad (127)$$

In summary, no information is lost in a homogeneous neural population even when the data of neural correlation, $\bar{R}$, are lost or when the mismatched model that ignores neural correlation is used for decoding.

DISCUSSION

In this work, we introduced a novel framework for investigating information processing in the brain where we studied information loss caused by two situations: imperfect observations and mismatched decoding. By evaluating the information loss caused by non-simultaneous observations of neural responses, we can quantify the importance of correlated activity. This can also be quantified by similarly evaluating the information loss caused by mismatched decoding that ignores neural correlation. We discussed these two types of loss by giving the information geometric interpretations of inferences with partially observed data and those with a mismatched decoding model and elucidated their relationship. We showed that the information loss associated with ignoring correlation in decoding is always larger than that caused by non-simultaneous observations of neural responses. This is because the inference based on an independent decoding model with complete data is equivalent to the inference based on an independent decoding model with "partial" data where the data of neural correlation are lost, which is naturally worse than the inference based on a correct decoding model with the partial data. This also can be intuitively understood by considering that decoding without the data of correlation considers all possible models of neural correlations, including the correct one, whereas decoding with a mismatched model locks it within the wrong domain.

Taking account of the relationship between the two inference methods, we give a simple guide on how to use the two different measures for quantifying the importance of correlation. To address the importance of coincidence detection by downstream neurons without making any specific assumption about the decoding process in higher-order areas in the brain, we should evaluate the information loss caused by non-simultaneous observations. The information loss quantified in this way can be used as the lower bound on the information conveyed by correlated activity. In contrast, the information loss quantified by using the independent decoding model can be used as the upper bound on the information conveyed by correlated activity because neural correlation is ignored not only in the observations but also in the decoding in this quantification. In summary, we consider that both measures should be computed when quantifying the importance of correlations and should be used as the lower bound and upper bound for the importance of correlations.

We considered the case that the stimulus is represented as a continuous variable. In this case, the Fisher information is a suitable measure for quantifying the maximal amount of information that can be extracted from neural responses. When considering a set

of discrete stimuli, which is mostly the case for neurophysiological experiments, we should use the mutual information instead of the Fisher information. Similar to the Fisher information, the mutual information obtained by the inference with partially observed data can be computed by marginalizing the probability distribution of neural responses over the sufficient statistics that correspond to the lost data. In addition, the mutual information obtained by the mismatched decoding model can be computed using information derived by Merhav et al. (Merhav et al., 1994; Latham and Nirenberg, 2005; Amari and Nakahara, 2006; Oizumi et al., 2010). Thus, when we use the mutual information, we can also quantify the importance of correlation with the concepts of missing data and mismatched decoding. We analyzed the binary probabilistic model with a homogeneous neural population and found that neural

correlation does not carry any information in this particular case. It is important to study how much information correlated activity could carry in an inhomogeneous neural population such as the hypercolumn model in V1. Another interesting future direction would be to evaluate the information conveyed by higher-order correlation (Amari et al., 2003); we considered only pair-wise correlation in this paper. These issues remain as future work.

ACKNOWLEDGMENTS

REFERENCES

Abbott, L. F., and Dayan, P. (1999). The effect of correlated variability on the accuracy of a population code. *Neural. Comput.* 11, 91–101.

Abeles, M. (1982). Role of the cortical neuron: integrator or coincidence detector? *Isr. J. Med. Sci.* 18, 83–92.

Amari, S. (1982). Differential geometry of curved exponential families – curvatures and information loss. *Ann. Stat.* 10, 357–385.

Amari, S. (1995). Information geometry of the EM and em algorithms for neural networks. *Neural Netw.* 8, 1379–1408.

Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Trans. Neural Netw.* 47, 1701–1711.

Amari, S., and Nagaoka, H. (2000). *Method of Information Geometry*. New York: Oxford University Press.

Amari, S., and Nakahara, H. (2006). Correlation and independence in the neural code. *Neural Comput.* 18, 1259–1267.

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

Averbeck, B. B., Latham, P. E., and Pouget, A. (2006). Neural correlations, population coding and computation. *Nat. Rev. Neruosci.* 7, 358–366.

Averbeck, B. B., and Lee, D. (2006). Effects of noise correlations on information encoding and decoding. *J. Neurophysiol.* 95, 3633–3644.

Brunel, N., and Nadal, J. (1998). Mutual information, Fisher information, and population coding. *Neural. Comput.* 10, 1731–1757.

Ecker, A., Berens, P., Keliris, G., Bethge, M., Logothetis, N., and Tolias, A. (2010).

Decorrelated neuronal firing in cortical microcircuits. *Science* 327, 584–587.

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

Golledge, H. D., Panzeri, S., Zheng, F., Pola, G., Scannell, J. W., Giannikopoulos, D. V., Mason, R. J., Tovée, M. J., and Young, M. P. (2003). Correlations, feature-binding and population coding in primary visual cortex. *Neuroreport* 14, 1045–1050.

Gray, C. M., König, P., Engel, A. K., and Singer, W. (1989). Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* 338, 334–337.

Ishikane, H., Gangi, M., Honda, S., and Tachibana, M. (2005). Synchronized retinal oscillations encode essential information for escape behavior in frogs. *Nat. Neurosci.* 80, 1087–1095.

König, P., Engel, A. K., and Singer, W. (1996). Integrator or coincidence detector? The role of the cortical neuron revisited. *Trends Neurosci.* 19, 130–137.

Latham, P. E., and Nirenberg, S. (2005). Synergy, redundancy, and independence in population codes, revisited. *J. Neurosci.* 25, 5195–5206.

Lee, D., Port, N. L., Kruse, W., and Georgopoulos, A. P. (1998). Variability and correlated noise in the discharge of neurons in motor and parietal areas of the primate cortex. *J. Neurosci.* 18, 1161–1170.

Meister, M., Lagnado, L., and Baylor, D. A. (1995). Concerted signaling by retinal ganglion cells. *Science* 270, 1207–1210.

Merhav, N., Kaplan, G., Lapidoth, A., and Shamai Shitz, S. (1994). On information rates for mismatched

decoders. *IEEE Trans. Inf. Theory* 40, 1953–1967.

Montani, F., Ince, R. A. A., Senetore, R., Arabzadeh, E., Diamond, M., and Panzeri, S. (2009). The impact of higher-order interactions on the rate of synchronous discharge and information transmission in somatosensory cortex. *Philos. Trans. R. Soc. A* 367, 3297–3310.

Nakahara, H., and Amari, S. (2002). Information-geometric measure for neural spikes. *Neural. Comput.* 14, 2269–2316.

Nirenberg, S., Carcieri, S. M., Jacobs, A. L., and Latham, P. E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature* 411, 698–701.

Nirenberg, S., and Latham, P. E. (2003). Decoding neural spike trains: how important are correlations? *Proc. Natl. Acad. Sci. U.S.A.* 100, 7348–7353.

Ohiorhenuan, I. E., Mechler, F., Purpura, K. P., Schmid, A. M., Hu, Q., and Victor, J. D. (2010). Sparse coding and high-order correlations in fine-scale cortical networks. *Nature* 466, 617–622.

Oizumi, M., Ishii, T., Ishibashi, K., Hosoya, T., and Okada, M. (2010). Mismatched decoding in the brain. *J. Neurosci.* 30, 4815–4826.

Schneidman, E., Berry, M. J. II, Segev, R., and Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature* 440, 1007–1012.

Shlens, J., Field, G. D., Gauthier, J. L., Grivich, M. I., Petrusca, D., Sher, A., Litke, A. M., and Chichilnisky, E. J. (2006). The structure of multi-neuron firing patterns in primate retina. *J. Neurosci.* 260, 8254–8266.

Tang, A., Jackson, D., Hobbs, J., Chen, W., Smith, J. L., Patel, H., Prieto, A.,

Petrusca, D., Grivich, M. I., Sher, A., Hottowy, P., Dabrowski, W., Litke, A. M., and Beggs, J. M. (2008). A maximum entropy model applied to spatial and temporal correlations from cortical networks in vitro. *J. Neurosci.* 28, 505–518.

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

Yaeli, S., and Meir, R. (2010). Error-based analysis of optimal tuning functions explains phenomena observed in sensory neurons. *Front. Comput. Neurosci.* 4:130. doi: 10.3389/fncom.2010.00130

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

APPENDIX

CALCULATION OF FISHER INFORMATION

In this appendix, we provide derivations for Eqs. 61–64, 70–73, and 89–91. First, we derive a useful relationship which holds for the exponential family. From the definition of the exponential family (Eq. 2), we obtain

$$\frac{\partial \log p(\mathbf{r};s)}{\partial \theta_I} = R_I(\mathbf{r}) - \frac{\partial^2 \psi(s)}{\partial \theta_I}. \tag{128}$$

Taking the average of this equation, we find

$$E_p\big[R_I(\mathbf{r})\big] = \frac{\partial \psi(s)}{\partial \theta_I}. \tag{129}$$

By using this relationship, we can derive Eqs. 61–63 as follows,

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \theta_i \partial \theta_j} = T \frac{\partial^2 \psi(s)}{\partial \theta_i \partial \theta_j},$$
$$= T \frac{\partial f_i}{\partial \theta_j},$$
$$= T C_{ij}, \tag{130}$$

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \theta_i \partial \Theta_{jk}} - T \frac{\partial (f_i f_k + C_{jk})}{\partial \theta_i},$$
$$= T\left(\frac{\partial f_i}{\partial \theta_i} f_k + \frac{\partial f_k}{\partial \theta_i} f_j\right),$$
$$= T(C_{ji} f_k + C_{ki} f_j), \tag{131}$$

and

$$-\frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \Theta_{ij} \partial \Theta_{kl}} = T \frac{\partial (f_i f_j + C_{ij})}{\partial \Theta_{kl}},$$
$$= \left(\frac{\partial f_i}{\partial \Theta_{kl}} f_j + \frac{\partial f_j}{\partial \Theta_{kl}} f_i + \frac{\partial C_{ij}}{\partial \Theta_{kl}}\right),$$
$$= T(2C_{ik} f_l f_j + 2C_{jk} f_l f_i + 2C_{ik} C_{lj}). \tag{132}$$

where we used Eqs. 58 and 59. We compute the Fisher information with respect to stimulus s by the following equation,

$$I_F(s) = -\sum_{i,j} \frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \theta_i \partial \partial_j} \frac{\partial \theta_i(s)}{\partial s} \frac{\partial \theta_j(s)}{\partial s}$$
$$-2\sum_{i,j,k} \frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \theta_i \partial \Theta_{jk}} \frac{\partial \theta_i(s)}{\partial s} \frac{\partial \Theta_{jk}(s)}{\partial s}$$
$$-\sum_{i,j,k,l} \frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \Theta_{ij} \partial \Theta_{kl}} \frac{\partial \Theta_{ij}(s)}{\partial s} \frac{\partial \Theta_{kl}(s)}{\partial s}. \tag{133}$$

By using Eqs. 130–132, we have

$$-\sum_{i,j} \frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \theta_i \partial \theta_j} \frac{\partial \theta_i(s)}{\partial s} \frac{\partial \theta_j(s)}{\partial s}$$
$$= T\sum_{i,j,k,l} C_{ij}(C_{ik}^{-1} f_k' + C_{ik}'^{-1} f k)(C_{jl}^{-1} f_l' + C_{jl}'^{-1} f_l),$$
$$= T(\mathbf{f'}^T \mathbf{C}^{-1} \mathbf{f'} + 2\mathbf{f}^T \mathbf{C'}^{-1} \mathbf{C} \mathbf{C}^{-1} \mathbf{f'} + \mathbf{f}^T \mathbf{C'}^{-1} \mathbf{C} \mathbf{C'}^{-1} \mathbf{f}), \tag{134}$$

$$-2\sum_{i,j,k} \frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \theta_i \partial \Theta_{jk}} \frac{\partial \theta_i(s)}{\partial s} \frac{\partial \Theta_{jk}(s)}{\partial s}$$
$$= 2T\sum_{i,j,k,l} \left(C_{ji} f_k + C_{ki} f_j\right)(C_{il}^{-1} f_l' + C_{il}'^{-1} f_l)\left(-\frac{1}{2} C_{jk}'^{-1}\right),$$
$$= T(2\mathbf{f}^T \mathbf{C'}^{-1} \mathbf{C} \mathbf{C}^{-1} \mathbf{f'} + 2\mathbf{f}^T \mathbf{C'}^{-1} \mathbf{C} \mathbf{C'}^{-1} \mathbf{f}), \tag{135}$$

$$-\sum_{i,j,k,l} \frac{\partial^2 \log p(\bar{\mathbf{r}},\bar{\mathbf{R}})}{\partial \Theta_{ij} \partial \Theta_{kl}} \frac{\partial \Theta_{ij}(s)}{\partial s} \frac{\partial \Theta_{kl}(s)}{\partial s}$$
$$-T\sum_{i,j,k,l} (2C_{ik} f_l f_j + 2C_{jk} f_l f_i + 2C_{ik} C_{lj})(-\frac{1}{2} C_{ij}'^{-1})(-\frac{1}{2} C_{kl}'^{-1}),$$
$$= T\left(\frac{1}{2} \mathrm{Tr}\big[\mathbf{C'} \mathbf{C}^{-1} \mathbf{C'} \mathbf{C}^{-1}\big] + \mathbf{f}^T \mathbf{C'}^{-1} \mathbf{C} \mathbf{C'}^{-1} \mathbf{f}\right). \tag{136}$$

Taken together, we can obtain Eq. 64. By similar computations, we can also obtain Eqs. 70–73.

Eqs. 88–91 can be derived as follows,

$$-\frac{\partial^2 \log p(\bar{r_1},\bar{r_2},\bar{R}_{12})}{\partial \theta_i \partial \theta_j}$$
$$= T\frac{\partial}{\partial \theta_j}\left(\frac{1}{Z} \sum_{r_1,r_2=0,1} r_i \exp\left(\sum_{i=1}^2 \theta_i(s)r_i + \Theta(s)r_1 r_2\right)\right)$$
$$= T\big(\langle r_i r_j \rangle - \langle r_i \rangle \langle r_j \rangle\big). \tag{137}$$

$$-\frac{\partial^2 \log p(\bar{r_1},\bar{r_2},\bar{R}_{12})}{\partial \Theta^2}$$
$$= T\frac{\partial}{\partial \Theta}\left(\frac{1}{Z} \sum_{r_1,r_2=0,1} r_1 r_2 \exp\left(\sum_{i=1}^2 \theta_i(s)r_i + \Theta(s)r_1 r_2\right)\right)$$
$$= T\big(\langle r_1 r_2 \rangle - \langle r_1 r_2 \rangle^2\big), \tag{138}$$

$$-\frac{\partial^2 \log p(\bar{r_1},\bar{r_2},\bar{R}_{12})}{\partial \theta_i \partial \Theta}$$
$$= T\frac{\partial}{\partial \theta_i}\left(\frac{1}{Z} \sum_{r_1,r_2=0,1} r_1 r_2 \exp\left(\sum_{i=1}^2 \theta_i(s)r_i + \Theta(s)r_1 r_2\right)\right)$$
$$= T\big(\langle r_1 r_2 \rangle - \langle r_i \rangle \langle r_1 r_2 \rangle\big). \tag{139}$$