2004 International Symposium on Nonlinear
Theory and its Applications (NOLTA2004)
Fukuoka, Japan, Nov. 29 - Dec. 3, 2004

Comments on analysis of neural coding by information geometric measure

Hiroyuki Nakahara

Lab. for Mathematical Neuroscience, RIKEN Brain Science Institute
2-1 Hirosawa, Wako, Saitama, 351-0198, Japan
Email: *hiro@brain.riken.jp*

Abstract—We previously proposed the use of information geometric (IG) measure, as a general method of analysis, to investigate the interaction of spike firing. The IG measure is a general method of handling the probability distribution of random binary vector. In the present paper, while we give a brief introduction of the IG measure, we discuss its characteristics in a plain way. It is hoped that the present paper helps us understand its utility in a more depth.

1. Introduction

The present paper has two aims; First, we like to give here a general, albeit briefly, introduction of information geometric (IG) measure on spike train analysis and second, we explain our perspective on this method of analysis and its implications.

The organization of the present paper is as follows. Section 2 explains the background and motivation. Section 3 introduces the IG measure and some of its essential properties. Then conclusion follows.

2. Background and motivation

One of the central challenges in neuroscience is to understand what and how information is carried by a population of neural firing. The simultaneous (or multi-unit) recording of many neural activities has been becoming widely available. One obvious advantage of such a technique is to save time in data collection; for example, collecting activities of a hundred neurons may be done one day by multi-unit recording, whereas a single unit recording may take a year. Of course, we should be aware of the different nature of data collection between these techniques, which we do not discuss in the present paper but is still worth to be mentioned. After this is said, we like to ask the question; how we can make best of such a massive data other than merely saving time?

This is the motivation behind our previous work; we proposed a method of analysis, called information geometric (IG) measure, and suggested that this method allows us to analyze the higher-order interaction among firing of a neural population [2]. An important characteristic of the IG measure is *model-free*; it is a general framework of analyzing a random binary vector[1]. In fact, due to its generality, it has been also applied to DNA microarray data successfully [3]. The application of the method on a single spike train is also discussed [4].

3. Information geometric measure

Fig 1 describes the data collection in multi-unit recording. In each trial, a number of neurons are recorded simultaneously over a time period. We discretize the time period by bins so small that there is only a single spike or no spike in each bin. By assigning 1 and 0 to a spike and no spike in each bin, respectively, each trial is represented by the matrix (Fig 1 right), filled by 0 or 1, whose size may be indicated as $M \times L$, where M is the number of neurons and L is the number of bins corresponding to the time period of a trial.

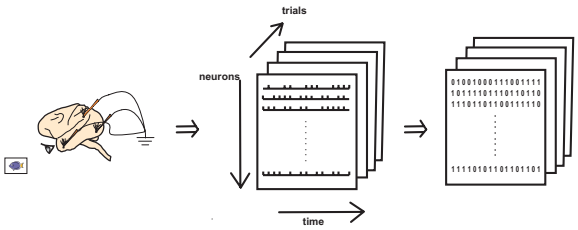

Figure 1: Schematic drawing of data collection in multi-unit recording

Let us put $N = M \times L$. All possible combination of 0s and 1s, i.e. $\{0,1\}^N$, is 2^N. Any probability distribution on this N-dimensional random binary vector is uniquely determined by determining the probabilities of $2^N - 1$ events, since the sum of all probabilities of 2^N events should be one. Let us write 2^N probabilities of each $p(\boldsymbol{x})$ by

$$p_{i_1 \cdots i_N} = \mathrm{Prob}\{X_1 = i_1, \cdots, X_N = i_N\},$$

[1]It can be also used as any random binary or even k-discrete vector

where $i_k = 0, 1$ and subject to $\sum_{i_1,\cdots,i_N} p_{i_1\cdots i_N} = 1$. The set of all the probability distributions $\{p(\boldsymbol{x})\}$ forms a $\left(2^N - 1\right)$-dimensional manifold $\boldsymbol{S}_N$. We call any combination of $\left(2^N - 1\right)$ components of $\{p_{i_1\cdots i_N}\}$ P-coordinates.

There can be various coordinate systems in $\boldsymbol{S}_N$. The two coordinate systems, namely η- and θ-coordinate systems, play an eminent role in information geometric measure [1, 2]. The η-coordinates is given by by the expectation parameters,

$$\eta_{i_1 i_2 \cdots i_k} = E\left[x_{i_1} \cdots x_{i_k}\right], \; k = 1, ..., N \tag{1}$$

which has $2^N - 1$ components. In other words,

$$\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_N) = (\eta_i, \eta_{ij}, \cdots, \eta_{1\cdots N}) \tag{2}$$

forms the η-coordinate system in $\boldsymbol{S}_N$, which is linearly related to $\{p_{i_1\cdots i_N}\}$. This η-coordinates defines m-flat structure in $\boldsymbol{S}_N$.

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

$$\begin{aligned} \log p(\boldsymbol{x}) \quad = \quad & \sum \theta_i x_i + \sum_{i<j} \theta_{ij} x_i x_j + \sum_{i<j<k} \theta_{ijk} x_i x_j x_k \\ & \cdots + \theta_{1\cdots n} x_1 \cdots x_N - \psi, \end{aligned} \tag{3}$$

where the indices of θ_{ijk}, etc. satisfy $i < j < k$, etc and ψ is a normalization term, corresponding to $-\log p(x_1 = x_2 = ... = x_N = 0)$. All θ_{ijk}, etc., together have $2^N - 1$ components and form another coordinate system, called θ-coordinates,

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_N) = (\theta_i, \theta_{ij}, \theta_{ijk}, \cdots, \theta_{12...N}). \tag{4}$$

and corresponding to e-flat structure in $\boldsymbol{S}_N$. It is easy to compute any components of $\boldsymbol{\theta}$ and for example, we can get $\theta_1 = \log \frac{p_{10,\ldots,0}}{p_{00,\ldots,0}}$.

Information geometry assure us that e-flat and m-flat manifolds are dually flat: The η-coordinates and θ-coordinates are dually orthogonal coordinates. The properties of the dual orthogonal coordinates remarkably simplify some apparently complicated issues. Due to the space limitation, we cannot fully describe them here (refer to [2] in details) but mention only a few results. To utilize the property of the dually orthogonal coordinates, it is convenient to define their partitions, called a k-cut;

$$\boldsymbol{\theta} = \left(\boldsymbol{\theta}_{k-} \; ; \; \boldsymbol{\theta}_{k+}\right), \quad \boldsymbol{\eta} = \left(\boldsymbol{\eta}_{k-} \; ; \; \boldsymbol{\eta}_{k+}\right) \tag{5}$$

where $\boldsymbol{\theta}_{k-}$ and $\boldsymbol{\eta}_{k-}$ consist of coordinates whose subindices have no more than k indices, i.e., $\boldsymbol{\theta}_{k-} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_k)$, $\boldsymbol{\eta}_{k-} = (\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \cdots, \boldsymbol{\eta}_k)$, and $\boldsymbol{\theta}_{k+}$ and $\boldsymbol{\eta}_{k+}$ consist of the coordinates whose subindices have more than k indices, i.e., $\boldsymbol{\theta}_{k+} = (\boldsymbol{\theta}_{k+1}, \boldsymbol{\theta}_{k+2}, \cdots, \boldsymbol{\theta}_N)$, $\boldsymbol{\eta}_{k+} = (\boldsymbol{\eta}_{k+1}, \boldsymbol{\eta}_{k+2}, \cdots, \boldsymbol{\eta}_N)$. Then let us define the k-cut mixed coordinate system by

$$\boldsymbol{\zeta}_k = \left(\boldsymbol{\eta}_{k-}; \boldsymbol{\theta}_{k+}\right). \tag{6}$$

Any k-cut mixed coordinate system forms the coordinate system of $\boldsymbol{S}_n$. A change in the $\boldsymbol{\theta}_{k+}$ part preserves the k-marginals of $p(\boldsymbol{x})$ (i.e., $\boldsymbol{\eta}_{k-}$), while a change in the $\boldsymbol{\eta}_{k-}$ part preserves the interactions among more than k variables. These changes are mutually orthogonal.

Let us illustrate some utilities of the mixed coordinates by simple examples [2]. Suppose we have an estimated probability distribution of N neurons in a *test* period of an experimental task (e.g. a period of showing an orientation bar to inspect neurons in early visual cortex), denoted by $p(\boldsymbol{x}; \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ represents the parameters (or the values of coordinate components) in a general form and a probability distribution of our null hypothesis (e.g. a probability of spontaneous firing in a control or resting period), $p(\boldsymbol{x}; \boldsymbol{\xi}^0)$. Then, using the k-cut mixed coordinates, we obtain the decomposition,

$$D[\boldsymbol{\xi}^0 : \boldsymbol{\xi}] = D[\boldsymbol{\zeta}^0 : \boldsymbol{\zeta}'_k] + D[\boldsymbol{\zeta}'_k : \boldsymbol{\zeta}], \tag{7}$$

where $\boldsymbol{\zeta}^0$ and $\boldsymbol{\zeta}$ are the mixed coordinates, corresponding to $\boldsymbol{\xi}^0$ and $\boldsymbol{\xi}$, respectively, and we define $\boldsymbol{\zeta}'_k = (\boldsymbol{\eta}^0_{k-}; \boldsymbol{\theta}_{k+})$. In this decomposition, $D[\boldsymbol{\zeta}^0 : \hat{\boldsymbol{\zeta}}'_k]$ represents the discrepancy between $\boldsymbol{\xi}^0$ and $\hat{\boldsymbol{\xi}}$ in the interactions higher than the k-th order and $D[\hat{\boldsymbol{\zeta}}'_k : \hat{\boldsymbol{\zeta}}]$ equal to and lower than the k-th order. Thus, by this decomposition, the term representing each discrepancy can be separated. Furthermore, this decomposition allows us to convert each divergence to p-values of χ^2 test (not shown here). In this manner, we can examine not only the pairwise interaction against null hypothesis of any correlated firing (in control period), whereas most previous studies were concerned with the pairwise interaction against the null hypothesis of independent firing, but also which higher-order interaction is statistically significantly different from that of the control period in a systematic manner.

This type of decomposition also leads to the decomposition of the mutual information between neural firing and behavior [2];

$$I(X, Y) = I_{k+}(X, Y) + I_{k-}(X, Y), \tag{8}$$

where we define $I_{k+}(X, Y) = E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_k(X|y) : \boldsymbol{\zeta}_k(X, y)\right]\right]$, $I_{k-}(X, Y) = E_{p(Y)}\left[D\left[\boldsymbol{\zeta}_k(X, y) : \boldsymbol{\zeta}_k(X)\right]\right]$ and $\boldsymbol{\zeta}_k(X, y) = (\boldsymbol{\eta}_{k-}(X|y); \boldsymbol{\theta}_{k+}(X))$. Note that examining if there exists significant coincident firing is related to but different from examining if the significant coincident firing conveys the information of behavior. Thus, having the decomposition of the mutual information under the same framework, like above, is useful in quantitatively relating the test of coincident firing with the test of behavioural information conveyed.

In [2], we pointed out that although many studies investigate the correlation of two neuron firing again

the null hypothesis of independent firing, it is rather more appropriate to test it again the null hypothesis of spontaneous correlated firing (e.g. in control period), or in general, against the null hypothesis of any correlated firing. The IG measure easily lets us do so. It can be also extended to the test of any higher-order interaction (of any number of neurons) against any null hypothesis. Under the same framework, the IG measure lets us decompose the information of behavior to the terms, conveyed by modulation of different order interactions among neurons.

3.1. Different cases

The IG measure itself is general and can be applied to any data of N-dimensional random binary vector, if done appropriately.

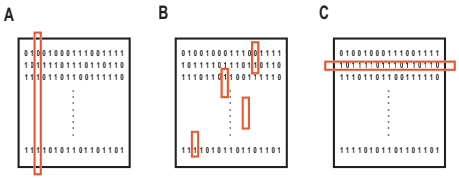

Figure 2: Different ways of creating N-dimensional random binary vector

Fig 2 shows the several cases in multi-unit recording data. In Fig 2 A, a time bin is chosen and fixed as the same for all neurons and then we consider the probability distribution over all neurons at the fixed bin. This corresponds to let $N = M$ (while setting $L = 1$). In [2], we primarily discussed this case only for presentation's simplicity (while briefly mentioning that we can also pick different time bins for different neurons as in Fig 2 B). In this case, we still have $N = M$. Alternatively, picking a single neuron, we can let $N = L$ (while setting $M = 1$) (Fig 2 C) and this is the case that the N-dimensional binary vector represents a spike train of a single neuron.

3.2. Generality of the IG measure, experimental limitation, and different models

Clearly, we can apply the same statistical tools on the distribution of N-dimensional random binary vector to all cases (Fig 2 A-C). This is a good news and is due to the generality of the IG measure. Furthermore, the IG measure treats probability distribution of N-dimensional random binary vector most thoroughly, since it handles the probability distribution with respect to the *full*, i.e. 2^N-1, components of coordinates in $\boldsymbol{S}_n$. Thus, in principle, any statistical question on probability distribution of N-dimensional random binary vector can be addressed by the IG measure. This is also a good news. For example, identifying the probability distribution of neural firing corresponds to identifying 'a point' in a $(2^N - 1)$-dimensional probability space (i.e. $\boldsymbol{S}_n$).

After these have been said, we must also mention that we face a severe limitation, once we start applying the IG measure to multi-unit recording data. That is the experimental limitation on the number of samples. Given N-dimensional random binary vector, we have to deal with the components of the coordinates, whose number is $2^N - 1$. Clearly, this number of coordinates can easily go beyond the currently available number of samples in most experiments. Then, applying the IG measure to such a data may look impossible. Is the IG measure useless in practical data analysis? What shall we do?

To consider this issue, it is worth first considering why this becomes the issue in using the IG measure and also how any other methods of analysis, or different models of neural firing (Fig 2 AB) and/or a single spike train (Fig 2 C), deal with same issue.

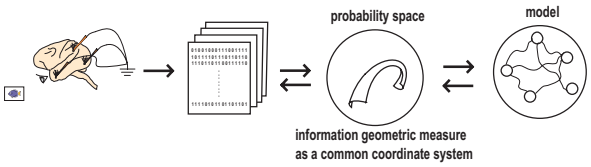

Figure 3: IG measure as a common coordinate system

The IG measure makes it so explicit what the full probability space is and this is why this becomes an issue. Any method cannot escape from the same issue, indeed, and must deal with the same issue, because they deal with the same data, data of N-dimensional random binary vector. Any methods, or models, are actually reducing the dimensionality of the probability space to search by their assumptions or hypotheses, regardless of whether they are mentioned explicitly or not. Most methods/models directly goes to this point first; each one, explicitly or (often implicitly or without mentioning), assumes what properties of data they consider as essential, then inspect some (but not full, in most cases) properties of data accordingly and claims itself better, if not best, than others in most of cases.

Which method/model is better? One advantage of the IG measure is that it can be used as a common coordinate system, since it treats the full space in a systematic manner. It can lay out the common ground for the comparison of different models or results by different methods of analysis (Fig 3). Each reduced search space of different models becomes a subspace in the full space. Then, any models or any assumptions are casted under the common framework.

Having a common (and good) coordinate system is

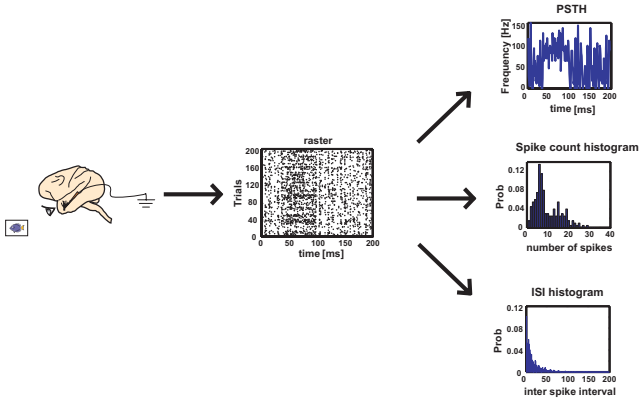

Figure 4: Raster is the basic data of spike train of a single neuron

important. To illustrate, consider spike train data of a single neuron (Fig 4). The raster plot contains all the information of spike trains, if the single-unit recording is done to record only spike occurrence, and is a most basic data format. The raster is then converted to different formats; the peristimulus histogram (PSTH), the spike count distribution, the inter spike interval (ISI) distribution and so on. In other words, for example, the spike count distribution is only a partial information extracted from the raster. It is true that the spike counts are often of primary interest, however, it is also true that we may miss some properties, possibly important, if we only inspect the spike count distribution in data. The "inverse" problem, recovering the probability distribution of N-dimensional random binary vector from the spike count distribution, is ill-posed.

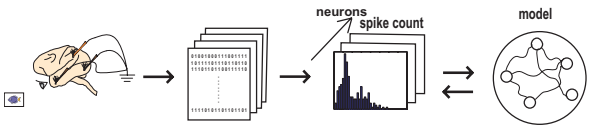

Figure 5: Multi-unit recording and spike count distribution

For example, if we assume the Poisson process as underlying process and regard the spike count distribution as the Poisson distribution, then we can recover the underlying process, by using the PSTH as additional information. The assumption here is so critical that it significantly reduces the space to search. Without such an assumption, we cannot recover the underlying process from the spike count distribution. Question is if this assumption is correct. Consider another example. If we begin with assuming that all what matters in firing of neural population is spike count distribution, our approach becomes as shown in Fig 5. If this assumption is correct, no worry. If not, we are destined to make a wrong conclusion. If we like to examine if this assumption is correct, we must be able to examine other properties in data.

Thus, we see that assumptions made in each model/method are critical and then difficult to be examined, if properties of data are inspected only within each model/method. The assumptions are important since they are the statement of what properties of data are regarded as essential by each model/method. In order to compare different models/methods, the IG measure is useful as the common coordinate system and capable of locating any models/methods in the full probability space (Fig 3).

On the other hand, any methods, including the IG measure, must incorporate some assumptions to reduce the search space in practice, simply because we never get the number of samples sufficient for locating the probability in the full space. Depending upon our interest, e.g. interaction among neurons (Fig 2 AB) or single spike train (Fig 2 C), we must take account of most appropriate assumptions or hypotheses. It is then important to elucidate what assumptions are required for data of our interest and what subspace they occupy in the full probability space. The IG measure allows us to do so in a transparent way.

4. Conclusion

We have introduced the information geometric (IG) measure and discussed its nature.

References

[1] S. Amari "Information geometry on hierarchical decomposition of stochastic interactions," *IEEE Trans on IT*, pp.1701–1711, 2001.

[2] H. Nakahara and S. Amari. "Information geometric measure for neural spikes," *Neural Computation*, pp.2269–2316, 2002.

[3] H. Nakahara et al, "Gene interaction in dna microarray data is decomposed by information geometric measure," *Bioinformatics*, 19:1124–1131, 2003.

[4] H. Nakahara, S. Amari, and B. Richmond "A comparison of descriptive models of a single spike train by information geometric measure," *in preparation.*