

# A new analytical method for characterizing nonlinear visual processes with stimuli of arbitrary distribution: Theory and applications

**Ryusuke Hayashi\***

Systems Neuroscience Group,  
Human Informatics Research Institute,  
National Institute of Advanced Industrial Science and  
Technology, Tsukuba, Ibaraki, Japan



**Osamu Watanabe\***

Graduate School of Engineering,  
Muroran Institute of Technology,  
Muroran, Hokkaido, Japan

**Hiroki Yokoyama\***

Tamagawa University Brain Science Institute,  
Machida, Tokyo, Japan



**Shin'ya Nishida**

Human Information Science Laboratory,  
NTT Communication Science Laboratories,  
Nippon Telegraph and Telephone Corporation,  
Atsugi, Kanagawa, Japan



Characterization of the functional relationship between sensory inputs and neuronal or observers' perceptual responses is one of the fundamental goals of systems neuroscience and psychophysics. Conventional methods, such as reverse correlation and spike-triggered data analyses are limited in their ability to resolve complex and inherently nonlinear neuronal/perceptual processes because these methods require input stimuli to be Gaussian with a zero mean. Recent studies have shown that analyses based on a generalized linear model (GLM) do not require such specific input characteristics and have advantages over conventional methods. GLM, however, relies on iterative optimization algorithms and its calculation costs become very expensive when estimating the nonlinear parameters of a large-scale system using large volumes of data. In this paper, we introduce a new analytical method for identifying a nonlinear system without relying on iterative calculations and yet also not requiring any specific stimulus distribution. We demonstrate the results of numerical simulations, showing that our noniterative method is as accurate as GLM in estimating nonlinear parameters in many cases and outperforms conventional, spike-triggered data analyses. As an example of the application of our method to actual psychophysical data, we investigated how different

spatiotemporal frequency channels interact in assessments of motion direction. The nonlinear interaction estimated by our method was consistent with findings from previous vision studies and supports the validity of our method for nonlinear system identification.

## Introduction

Reverse correlation (RC) and spike-triggered average (STA) techniques have been extensively used to characterize neural systems (Ringach & Shapley, 2004; Simoncelli, Paninski, Pillow, & Schwartz, 2004), such as in identifying receptive field properties of retinal ganglion cells (Hida & Naka, 1982), lateral geniculate nucleus neurons (Reid & Alonso, 1995), and simple cells in the primary visual cortex (DeAngelis, Ohzawa, & Freeman, 1993a; DeAngelis, Ohzawa, & Freeman, 1993b; Jones & Palmer, 1987; Ohzawa, DeAngelis, & Freeman, 1990). In RC/STA, it is hypothesized that the output of the interested system is determined by a linear weighting of input stimuli followed by a nonlinear response function

Citation: Hayashi, R., Watanabe, O., Yokoyama, H., & Nishida, S. (2017). A new analytical method for characterizing nonlinear visual processes with stimuli of arbitrary distribution: Theory and applications. *Journal of Vision*, 17(6):14, 1–20, doi:10.1167/17.6.14.

doi: 10.1167/17.6.14

Received June 1, 2016; published June 21, 2017

ISSN 1534-7362 Copyright 2017 The Authors



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

and a noise response (for neural data this is typically Gaussian or Poisson, and for psychophysical data, typically Bernoulli, the so-called linear–nonlinear (L–N) model). The parameters of the linear filter in the first stage of the L–N model are estimated as proportional to the expectation of inputs conditioned by the system’s response. The technique has also been generalized to characterize a nonlinear system expressed as the Wiener/Volterra series expansion. The system consists of a linear combination of monomials in the components of the input vector (see Eq. 1), whose coefficients are estimated by a cross-correlation procedure (Franz & Schoelkopf, 2003; Orcioni, 2014; Schetzen, 1980). In this framework, RC and STA are methods for estimating the first-order term in the Wiener/Volterra representation of the system, and spike-triggered covariance analysis (STC, Schwartz, Chichilnisky, & Simoncelli, 2002; Simoncelli et al., 2004) provides an estimate of the components of the second-order term. The RC technique has also been extended to psychophysical experiments, defined as noise image classification, for deriving linear (Ahumada, 1996; Beard & Ahumada, 1998) and nonlinear (Neri, 2004) properties of sensory filters. These conventional methods, however, rely on the assumption that the ensembles of input stimuli satisfy a specific statistical distribution (RC/STA: spherical symmetry [Chichilnisky, 2001]; STC: Gaussian with a zero mean [Paninski, 2003]). Therefore, the use of these techniques has been valuable only at the early stage of sensory processing, when using statistically restricted stimuli such as randomly changing luminance patterns. On the other hand, there is an increasing demand for experiments that study neuronal responses to natural stimuli (David, Vinje, & Gallant, 2004; Vinje & Gallant, 2000). As natural images have specific statistical properties and sample only a subspace of all possible images explored during stimulation with white noise (Ruderman, 1994), it is preferable to use a method for identifying complex neural systems without any statistical constraint on the input stimulus.

Recent studies have shown that analyses using a generalized linear model (GLM) framework have advantages over STA and STC methods in system identification. In GLM (Nelder & Wedderburn, 1972), a weighted linear combination of inputs is related to the response via a link function (a function homologous to the nonlinear response function in a L–N model). The model parameters are then estimated by maximum-likelihood estimation using an iteratively reweighted least-squares method. It has been demonstrated that GLM achieves accurate parameter estimation with fewer data samples than required by STA, and that estimation using GLM with appropriate modifications is more robust to

internal noise (Knoblauch & Maloney, 2008). A GLM framework is also readily extendable to the estimation of higher-order nonlinear parameters (Neri, 2004). A drawback of the GLM method, however, is the calculation cost of its iterative optimization algorithm. The cost becomes very high when estimating nonlinear parameters in a large-scale system using large numbers of data samples. Moreover, the GLM method requires optimization of the estimation based on the entire data set; thus, it does not provide the means for sequential updates and its scalability for data size is limited.

Here, we introduce an alternative method, Watanabe’s method, based on moment-generating functions (Weisstein, 2017) to characterize a nonlinear system without iterative optimization. In the next section, the formulation of our method is described in detail and its advantages over conventional methods demonstrated when estimating the parameters of a known visual processing model. In a comparison of the proposed method with GLM, kernel estimation was shown to provide similar accuracy in many cases. In addition to numerical simulations, a psychophysical experiment was conducted to validate the proposed method with actual data, as it was not obvious whether reasonable and interpretable results could be obtained when estimating a large number of parameters using data recorded from real-world constrained experiments. In the experiment, we asked human observers to report perceived motion direction when a variety of moving sinusoidal gratings were simultaneously presented. We investigated how signals from different spatiotemporal frequency channels are integrated for unified motion perception. Although a previous study (Hayashi, Sugita, Nishida, & Kawano, 2009) suggested that nonlinear interactions between different frequency channels account for the observed behavioral data, the properties of higher-order interaction are still poorly understood. In the last half of this paper, we show the estimated first- and second-order kernels of visual motion process and discuss the validity of our method as well as its limitations.

## Method

Our new estimation method assumes three preconditions. First, we consider the situation where the output  $y$  of the system of interest takes only binary states (either  $y_0$  or  $y_1$ ), which is the case when we investigate the receptive field property of neurons from firing responses, or examine human observers’ psychophysical decision processes in two-alternative

forced-choice (2AFC) tasks. Second, the associated nonlinear system is time invariant and expressed mathematically as a Volterra series expansion of input  $\mathbf{x} = (x_1, x_2, \dots, x_N)^T$ , as follows:

$$F(\mathbf{x}) = F_0 + \sum_i^N F_{1,i} x_i + \sum_{i \leq j}^N F_{2,ij} x_i x_j + \dots \quad (1)$$

Third, we hypothesized that neural responses or a human observer's responses are stochastically determined by a certain noise distribution. Although this inner noise could take many forms, we assumed that the response follows the fluctuation of a normal distribution, the variance of which is 1/2 for simplicity. This means that the system's response is determined by the output of  $F(\mathbf{x})$  followed by a cumulative Gaussian, the sigmoid-type response function, as follows:

$$P(y = y_1 | \mathbf{x}) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{F(\mathbf{x})} e^{-t^2} dt = \frac{1 + \text{erf}(F(\mathbf{x}))}{2}, \quad (2)$$

where  $\text{erf}(x)$  is Gauss' error function given by Eq. 3 as follows:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \quad (3)$$

The conditional probability distribution of response  $y$ , given input  $\mathbf{x}$  is therefore described as follows:

$$P(y | \mathbf{x}) = \delta(y - y_1) \cdot \frac{1 + \text{erf}(F(\mathbf{x}))}{2} + \delta(y - y_0) \cdot \frac{1 - \text{erf}(F(\mathbf{x}))}{2}, \quad (4)$$

where  $\delta(x)$  is Dirac's delta function.

Based on an identity of probability distribution involving moment-generating functions and the Maclaurin series expansion of response function, in this case cumulative Gaussian function, we derived that the coefficients (or kernel) of Volterra series expansion of  $F(\mathbf{x})$  can be estimated by solving simultaneous equations composed of moments, i.e., expectation of the products of inputs  $\mathbf{x}$  and or output  $y$  (the detail derivations are described in the Appendix). Namely, we obtain a relational expression between the vectors/matrix of expectations  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $M$ , and the coefficients of the moment-generating function  $\mathbf{g}$  as:

$$\mathbf{a} = \frac{y_1 + y_0}{2} \mathbf{b} + \frac{y_1 - y_0}{\sqrt{\pi}} M \cdot \mathbf{g}, \quad (5)$$

where,

$$\mathbf{a} = (E[y], E[yx_1], \dots, E[yx_N], E[yx_1^2], E[yx_1x_2], \dots, E[yx_N^2], \dots)^T$$

$$\mathbf{b} = (1, E[x_1], \dots, E[x_N], E[x_1^2], E[x_1x_2], \dots, E[x_N^2], \dots)^T$$

$$\mathbf{g} = (G_0, G_{1,1}, G_{1,2}, \dots, G_{1,N}, G_{2,11}, G_{2,12}, \dots, G_{2,NN}, \dots)^T$$

$$M = \begin{pmatrix} 1 & E[x_1] & \dots & E[x_N] & E[x_1^2] & E[x_1x_2] & \dots & E[x_N^2] & \dots \\ E[x_1] & E[x_1^2] & \dots & E[x_1x_N] & E[x_1^3] & E[x_1^2x_2] & \dots & E[x_1x_N^2] & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ E[x_N] & E[x_1x_N] & \dots & E[x_N^2] & E[x_1^2x_N] & E[x_1x_2x_N] & \dots & E[x_N^3] & \dots \\ E[x_1^2] & E[x_1^3] & \dots & E[x_1^2x_N] & E[x_1^3] & E[x_1^2x_2] & \dots & E[x_1^2x_N^2] & \dots \\ E[x_1x_2] & E[x_1^2x_2] & \dots & E[x_1x_2x_N] & E[x_1^3x_2] & E[x_1^2x_2^2] & \dots & E[x_1x_2x_N^2] & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ E[x_N^2] & E[x_1x_N^2] & \dots & E[x_N^3] & E[x_1^2x_N^2] & E[x_1x_2x_N^2] & \dots & E[x_N^4] & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \end{pmatrix}$$

and the coefficients  $\mathbf{g}$  are described using the kernels of  $F(\mathbf{x})$  as follows:

$$\begin{aligned} G_0 &= c_0 \\ G_{1,i} &= c_1 F_{i,1} \\ G_{2,ij} &= \begin{cases} c_1 F_{2,ij} + 2c_2 F_{1,i} F_{1,j}, & \text{if } i \neq j \\ c_1 F_{2,ii} + c_2 F_{1,i}^2, & \text{if } i = j \end{cases} \end{aligned}$$

⋮

$$c_r = \sum_{n=0}^{\infty} \frac{(-1)^n}{n!(2n+1)} \binom{2n+1}{r} F_0^{2n+1-r}$$

Whereas RC and STA are methods exploiting the fact that  $E[yx_i]$  is proportional to the first-order kernel  $F_{1,i}$  when the input distribution is spherically symmetrical with a mean of zero, our method identifies the kernels, including higher-order ones, without relying on any assumptions about input distribution.

The internal noise variance at the stage of binary decision, i.e., the slope of the cumulative Gaussian function, affects only the estimation of the entire scale of the kernels. Thus, we can estimate the shape of kernels accurately even if the choice of internal noise variance does not match the actual value of a system under consideration, which is usually unknown. We assume the noise variance is 1/2 as the Maclaurin series expression of the cumulative Gaussian function is simple. In theory, the estimated kernels derived from the solution of Eq. 5 correspond with the ground truth, both in scale and shape, when the variance of fluctuation at the stage of binary decision is indeed 1/2.

For further simplicity, we set response  $y$  at either  $-1$  or  $1$  without loss of generality and consider the case's ability to estimate the kernels of Volterra series up to second-order terms from data observed by repeated experiments.

If we set the observed responses from  $m$  trials as

$$\mathbf{y} = (y_1, y_2, \dots, y_m)^T$$

and make matrix  $X$  consist of the input values of all stimulus dimensions and their binomial products from  $m$  trials as

$$X = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ \vdots & \vdots & \dots & \vdots \\ x_{N,1} & x_{N,2} & \dots & x_{N,m} \\ x_{1,1}^2 & x_{1,2}^2 & \dots & x_{1,m}^2 \\ x_{1,1} \cdot x_{2,1} & x_{1,2} \cdot x_{2,2} & \dots & x_{1,m} \cdot x_{2,m} \\ \vdots & \vdots & \dots & \vdots \\ x_{N,1}^2 & x_{N,2}^2 & \dots & x_{N,m}^2 \end{pmatrix}$$

then, vector  $\mathbf{a}$  and matrix  $M$  are obtained by the following approximation:

$$\mathbf{a} \approx \frac{1}{m} X \cdot \mathbf{y} \quad (6)$$

$$M \approx \frac{1}{m} X \cdot X^T \quad (7)$$

Therefore, the coefficients of the moment-generating function in Eq. 5 are given by the following equation:

$$\mathbf{g} = \frac{\sqrt{\pi}}{2} M^{-1} \cdot \mathbf{a}, \quad (8)$$

and the coefficients/kernels of Volterra series can be derived as

$$\begin{cases} \hat{F}_0 & = \operatorname{erf}^{-1}\left(\frac{2}{\sqrt{\pi}} G_0\right) \\ \hat{F}_{1,i} & = e^{\hat{F}_0^2} G_{1,i} \\ \hat{F}_{2,ij} & = \begin{cases} e^{\hat{F}_0^2} G_{2,ij} + 2\hat{F}_0 \hat{F}_{1,i} \hat{F}_{1,j}, & \text{if } i \neq j \\ e^{\hat{F}_0^2} G_{2,ij} + \hat{F}_0 \hat{F}_{1,i}^2, & \text{if } i = j \end{cases} \end{cases} \quad (9)$$

Note that the formulation for our method does not assume any statistical constraint on the stimulus ensemble, making it theoretically possible to characterize a nonlinear system using any type of stimuli, including natural images, the distribution of which is non-Gaussian. Not only our method, but also maximum likelihood estimators, such as GLM and a method using mutual information (Sharpee, Rust, & Bialek, 2003), are known to be input distribution-free estimations. However, these other methods use an iterative optimization algorithm to estimate the kernels. MATLAB (MathWorks, Inc., Natick, MA) codes that demonstrate how much faster our method works than GLM to generate similar results are available in the Supplementary Materials accompanying this article [to calculate the kernels using data of 144 dimensions and 32,000 samples, GLM took approximately 5,220 s whereas our method took 19 s on a desktop computer (Windows 10, Microsoft Corp., Redmond, WA; Corei7-6950X CPU 3.00 GHz, 64 GB RAM, 64-bit operating system, Intel Corp., Santa Clara, CA)].

Another drawback of GLM is that the kernel estimation is optimized based on the entire data set and is not updated sequentially. This means that time-consuming GLM

calculations must be done from scratch whenever additional data samples are obtained. However, because our method relies on the calculation of the inverse matrix of various moments, our method can update the kernel estimation sequentially based on the Sherman–Morrison formula (Weisstein, 2017), a formula used to correct an already known inverse matrix using additional data/matrix. The computation becomes relatively cheaper when updating/correcting the previously estimated kernels based on the Sherman–Morrison formula than when calculating kernels from scratch. We describe the formulation on how to extend our method to enable sequential updates in the Appendix (MATLAB codes for this extension are also available in the Supplementary Materials). Such an extension will be useful when attempting to change the input stimulus in an adaptive way online during experiments or when the data size is too large to handle and requires smaller fractions for kernel estimation.

To summarize, our method has the following advantages over STA/STC techniques: (1) Our analysis does not require stimuli to be spherically symmetrical or Gaussian with a zero mean. (2) Our method can estimate nonlinear kernels. The advantage of our method over GLM is that (3) the nonlinear kernels can be derived analytically from data without iterative calculations. (4) Our method also has better scalability than GLM, as it can use sequential updating of the inverse matrix. It should be noted that we formulated our method based on the assumption that the internal noise distribution is Gaussian with variance of 1/2, but the noise does not always fit to Gaussian and its variance is usually unknown. (Possible expansion of our method using other noise model is discussed later). Because GLM also has to specify the link function (the type and slope of response function, which is usually unknown), neither our method nor GLM can estimate the scale of kernels. To the contrary, STA and STC do not require specification of the noise distribution in a system and retaining an advantage in this point over our method. However, STA/STC are not able to estimate the scale of kernels because of their derivation.

In the next section, we demonstrate that our method provides better parameter estimation of a toy model of a neuron in the primary visual area (V1) than STA and STC when the input distribution is not symmetrical and its mean is not zero. Our method is then compared with GLM in detail, using a system of randomly generated Volterra kernels.

## Numerical simulation

### Comparison with STA/STC: Identification of a neural model

Two types of cells are found in the V1 area (Hubel & Wiesel, 1962). One is a simple cell whose receptive field

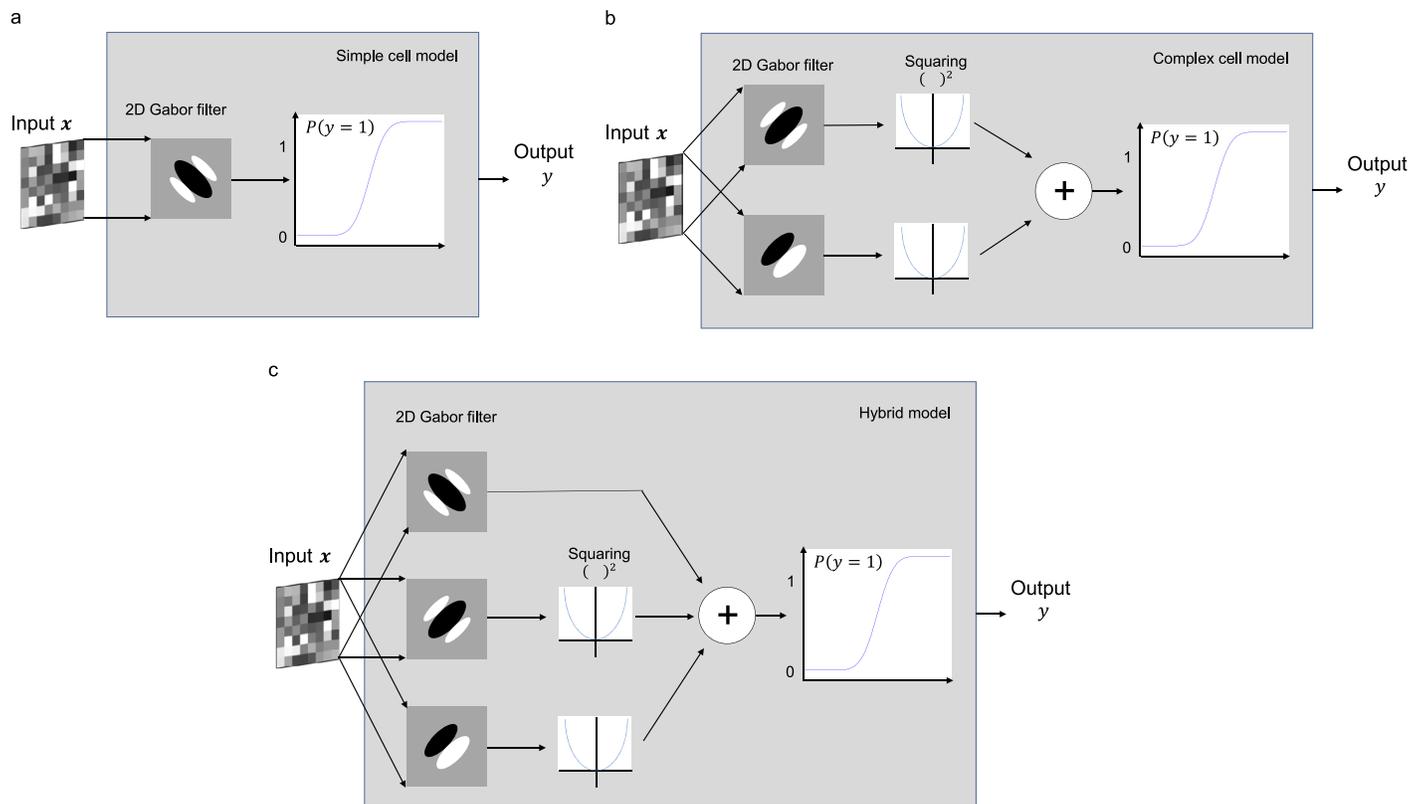


Figure 1. (a) 2D Gabor filter model (simple cell model): The response of a simple cell is hypothesized to be determined by the summation of an input image weighted with the 2D Gabor pattern, followed by the sigmoid-type response function. (b) Energy model (complex cell model): The response of the complex cell is hypothesized to be determined by the summation of the squared outputs of two Gabor filters whose phases are orthogonal with each other, followed by a sigmoid-type response function. (c) Hybrid model of simple cell and complex cell used for numerical simulations.

consists of distinct excitatory and inhibitory subregions, and which responds primarily to the gratings and edges of a particular orientation. The other type of cell is a complex cell that responds to the oriented gratings and edges regardless of stimulus positions. It is known that the two-dimensional (2D) Gabor filter model, i.e., the weighted summation of input image with a 2D Gabor function, followed by a nonlinear sigmoid-type function (Figure 1a), provides a good fit with the response properties of simple cells. A popular model for complex cells is that they act like energy mechanisms as proposed by Adelson and Bergen (1985): The response of the modeled complex cell is determined by the summation of the squared outputs of two Gabor filters whose phases are orthogonal with each other, followed by a sigmoid-type function [see Figure 1b and the review of V1 cell models (Heeger, 1991)].

For the first numerical simulation, we identified the parameters of a hybrid model of a simple cell and a complex cell (Figure 1c) from its responses to noise stimuli. The first-order kernels of this model correspond with the Gabor filter weights of the simple cell, and second-order kernels correspond to the energy model of the complex cell. As our inputs, we set 2D images of  $8 \times 8$

pixels, i.e., 64 dimensions. We set  $y = 1$  when the hybrid neural model fires in response to the image input and  $y = -1$  when the model is silent. The internal noise variance of this model was set as  $1/2$ . We used Gaussian noise inputs, the variances of which were not equal across pixels and had a steady offset to make the stimulus distribution asymmetrical and non-zero centered. The responses of the neural model were simulated repeatedly for 250,000 trials to estimate the kernels.

The receptive field of the simple cell was set as a Gabor function tuned to  $135^\circ$  (Figure 2a). Figure 2b–d show the estimated first-order kernels using STA, whitened STA, and our method, respectively. STA failed to reconstruct the fine shape of the simple cell's receptive field due to the non-zero centered and nonspherical stimulus input. Figure 2c is the result of the STA method after whitening stimulus statistics using a decorrelation technique (David et al., 2004) at postprocessing (see the formula of whitened STA in Appendix). The whitened STA still failed to estimate the first-order kernel accurately under the condition where the center of the distribution of the stimuli was deviated from zero. Figure 2d shows the estimated first-order kernels using our method. Our method provided an excellent estimation, despite the

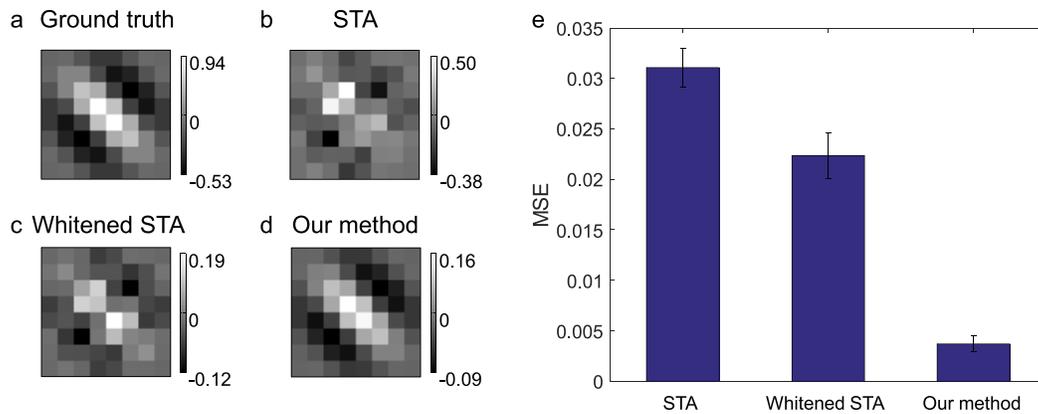


Figure 2. First-order kernel estimation of a simple cell model. (a) Ground truth of the 2D Gabor filter of a simple cell model. (b) The estimated first-order kernel using RC/STA from the model's responses. (c) The estimated first-order kernel using whitened STA. (d) The estimated first-order kernel using our method from the model's responses. (e) Mean squared error of the first-order kernel between the ground truth and the estimates. We repeated STA, whitened STA, and our method using 20 different noncentered stimulus distributions and plotted the means of MSE for three estimations. The error bars are standard error.

non-zero centered and nonspherical input distribution (correlation between estimated kernel and ground truth was significantly high ( $R^2 = 0.964$ ,  $p < 0.01$ ). The correlation between our method and ground truth was also significantly different from that between STA and ground truth, and between whitened STA and ground truth ( $p < 0.01$ , Meng's Z-test). The estimated kernel values by our method deviated slightly from the ground truth due to the truncation error of the second-order approximation in the calculation although the noise variance of the model was matched. Figure 2e shows the

mean squared errors (MSEs) of the estimations of STA, whitened STA, and our method after adjusting the scale of kernels to minimize each MSE. Our method shows significantly lower MSE over STA and whitened STA ( $t$  test of 20 estimations using different noncentered stimulus distribution,  $p < 0.01$ ).

We also identified the second-order kernels; i.e., the parameters of a complex cell, from the responses of the same system. The receptive field of the two constituent filters of the complex cell is tuned to  $45^\circ$  orientation with cosine and sine phase modulation in this

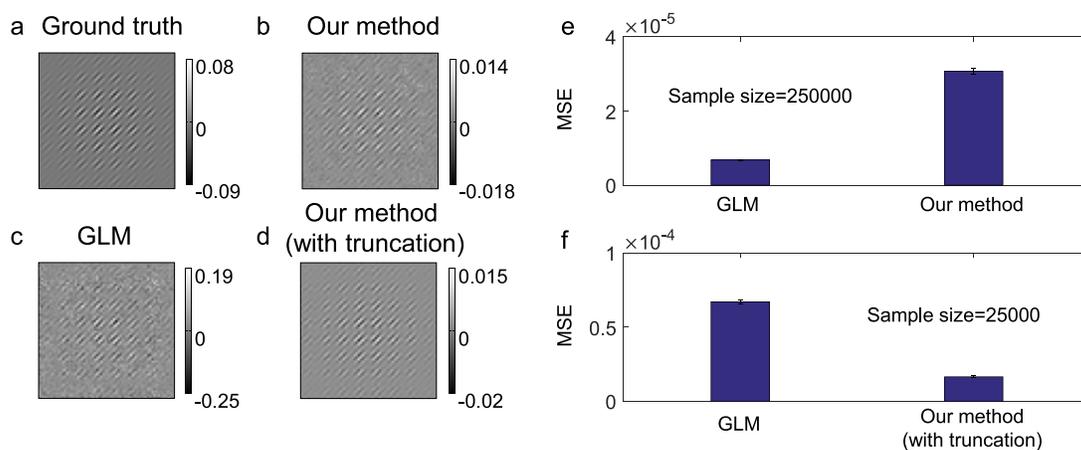


Figure 3. Second-order kernel estimation of a complex cell model. (a) Ground truth of the second-order kernel of a complex cell model, i.e., the squared summation of a quadrature pair of Gabor filters oriented to  $45^\circ$ . (b) The estimated second-order kernel using our method (sample size = 250,000). Although second-order kernels are not defined in the domain where the input component  $j > i$ , in our method, we duplicated their values as  $F_{2,ij} = F_{2,ji}$  in this figure for visibility. (c) The estimated second-order kernel using the GLM (Probit analysis, sample size = 25,000). (d) The estimated second-order kernel using our method truncating small eigenvalues (except for the top two) to zero at postprocessing (sample size = 25,000). (e) Mean squared error of the second-order kernel between the ground truth and the estimates. We repeated GLM and our method using 20 different noncentered stimulus distributions and plotted the means of MSE. The error bars are standard error. Sample size was 250,000. (f) Mean squared errors of the second-order kernel estimation using GLM and our method with eigenvalue-truncation. Sample size was 25,000.

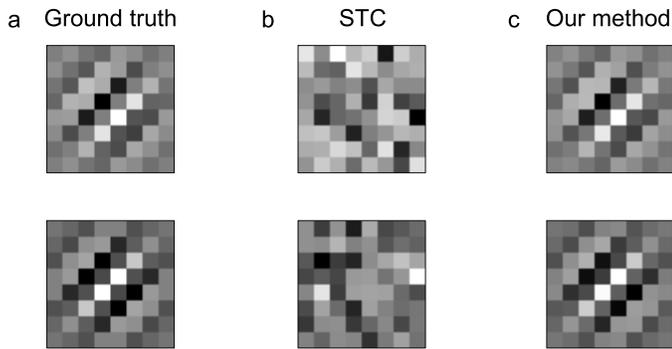


Figure 4. Linear parameter estimation of a complex cell model using eigenvalue decomposition analysis. (a) A pair of 2D Gabor filters in the complex cell model. (b) The first (upper row) and second (bottom row) eigenvectors of the whitened STC matrix. Note that eigenvectors have an arbitrary property in terms of the constant factor; thus, the polarity of the estimated filter may differ from the original one. (c) The first (upper row) and second (bottom row) eigenvectors of the second-order kernel matrix estimated by our method.

simulation (Figure 1c). Figure 3a shows the ground truth of second-order kernels of the complex cell model, i.e., the squared summation of the two constituent Gabor filters (also see Figure 4a). Figure 3b shows the estimated second-order kernels using our method. Our method was able to estimate the second-order kernel quite accurately, despite using asymmetrical input. The correlation between the estimated kernel and ground truth was significantly high ( $R^2 > 0.859$ ,  $p < 0.01$ ). The estimated kernel values by our method deviated slightly from the ground truth due to the truncation error of the second-order approximation in the calculation.

In cases where the system consists of a linear filter followed by a nonlinear stage, we can estimate the linear-filter parameters from the second-order kernel matrix (and STC matrix) by determining along which stimulus axes the variance of the spike-triggered stimulus ensemble significantly differs from that of the raw stimulus ensemble (Simoncelli et al., 2004). Specifically, the eigenvectors of the second-order kernel matrix represent the principal axes of the spike-triggered stimulus ensemble, and the corresponding eigenvalues represent the variances along each of these axes. The eigenvectors for positive eigenvalues indicate that the corresponding stimulus dimension is excitatory, and the eigenvectors for negative eigenvalues indicate that the corresponding stimulus dimension is inhibitory (see the later section on eigenvalue decomposition analysis for detail). Figure 4a shows the ground truth of the 2D Gabor filters of the complex cell model. Figure 4b and 4c are the estimated Gabor filters from the eigenvectors of the top two eigenvalues. Again, conventional STC (with whitening in postpro-

cessing) was not able to provide a good estimation of the shape of the Gabor filter due to the asymmetrical stimulus input (Figure 4b). However, our method was able to accurately estimate the second-order kernel (Figure 4c), thus outperforming the conventional methods in accurate estimation of nonlinear parameters.

### Comparison with GLM: Identification of a model of randomly generated Volterra kernels

To compare our method with GLM from various aspects, we tested it to identify a nonlinear image processing system, with known randomly generated Volterra kernels. In the following, we chose Probit analysis as a GLM to compare with our method; Probit analysis is widely used for estimating the kernels of binary-response systems, in which the link function is cumulative Gaussian as in our method. We examined a system processing 2D images of  $8 \times 8$  pixels; i.e., 64 dimensions. The zero-, first-, and second-order kernels were generated randomly from a Gaussian distribution  $[N(0, 1)]$ ; 20 different systems were estimated for statistical comparisons. The input was a random noise image, the intensity distribution of which followed a Gaussian distribution  $[N(0,1)]$ .

Figure 5a shows the correlation of first-order kernels between the ground truth and estimation as a function of trial numbers, i.e., the number of data samples, used to estimate one system; the averages of 20 estimates for 20 randomly generated systems are shown. We used the correlation coefficients as metrics for the accuracy of estimation because the exact scale of the kernels cannot be determined as shown in the previous section. The kernels estimated by our method became close to the ground truth when sample size was enough large (correlation was significantly high as  $R^2 = 0.829$ , when sample size was 16,000). The performance is comparable with that of GLM, although the calculation cost is much lower with our method. The results also show that our method outperforms STA for a wide range of kernel estimation. (Postprocess whitening did not affect the results of STA since we used Gaussian stimulus input in this simulation). Figure 5b is the plot of the correlation of the second-order kernels between the ground truth and estimation. Although GLM shows better estimates than our method in terms of accuracy, the difference is small; additionally, our method provides accurate kernel estimation (whose correlation is significantly high as  $R^2 = 0.911$ , when sample size was 16,000).

The accuracy of the second-order kernel estimation by our method depends on the rank of the second-order kernel matrix. Figure 6a–c demonstrate how estimation performance changed as a function of rank (we fixed

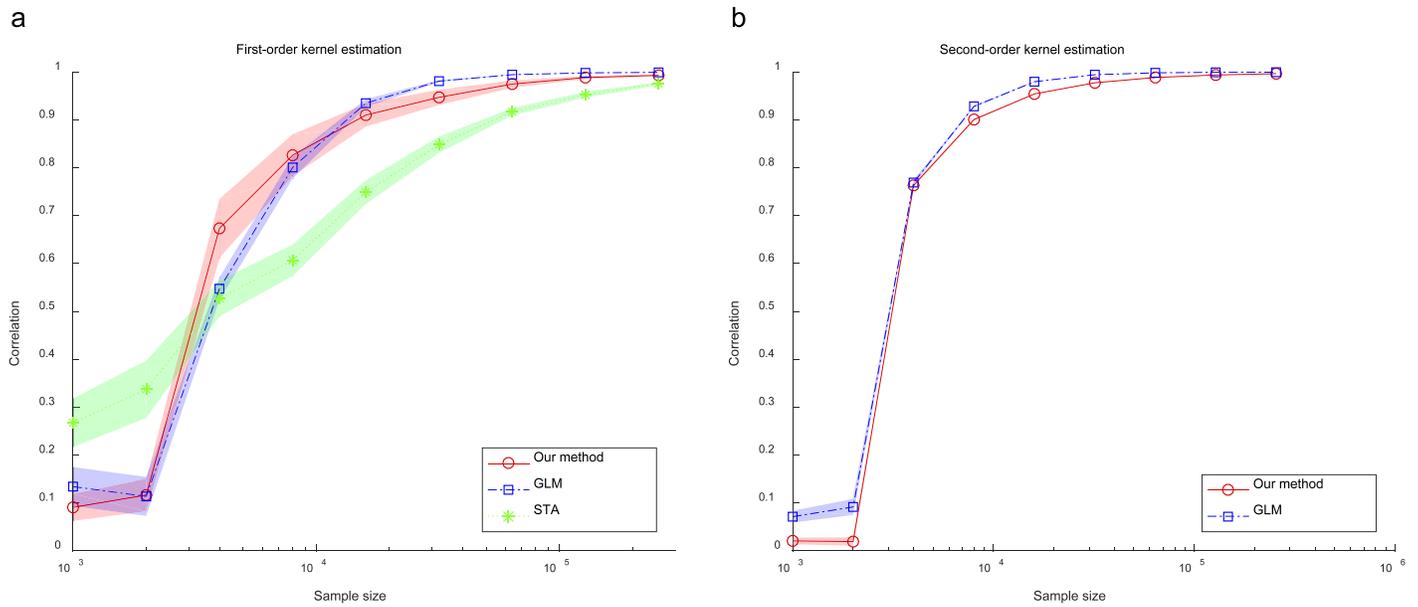


Figure 5. Correlation of the kernels between the ground truth and estimates is plotted as a function of the number of data samples. The solid lines are the mean of 20 repetitions of simulations. The shaded regions around each curve mark 95% confidence intervals of the values about the mean. Red, blue, and green lines are the results of the estimation by our method, the GLM, and STA, respectively. (a) Results of the first-order kernel estimation. (b) Results of the second-order kernel estimation.

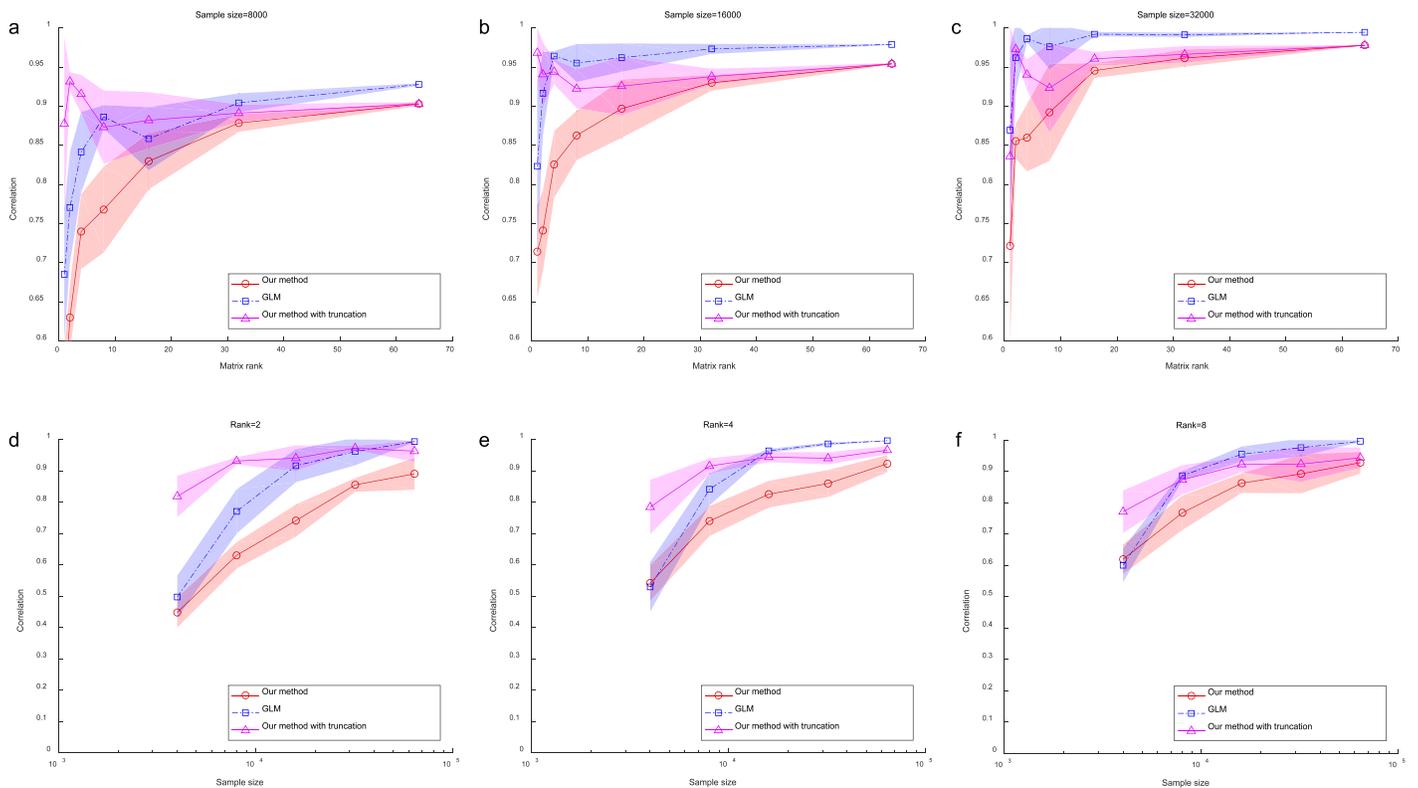


Figure 6. Upper row: Correlation of the second-order kernels between the ground truth and estimates plotted as a function of the rank of the second-order kernel matrix. Red, blue, and magenta lines are the results of the estimation by our method, GLM, and our method truncating small eigenvalues, respectively. Results using data of (a) 8,000 samples, (b) 16,000 samples, and (c) 32,000 samples. Lower row: Correlation of the second-order kernels between the ground truth and estimates plotted as a function of the number of data samples. Results when the rank of the second-order kernel matrix is (d) 8, (e) 4, and (f) 2.

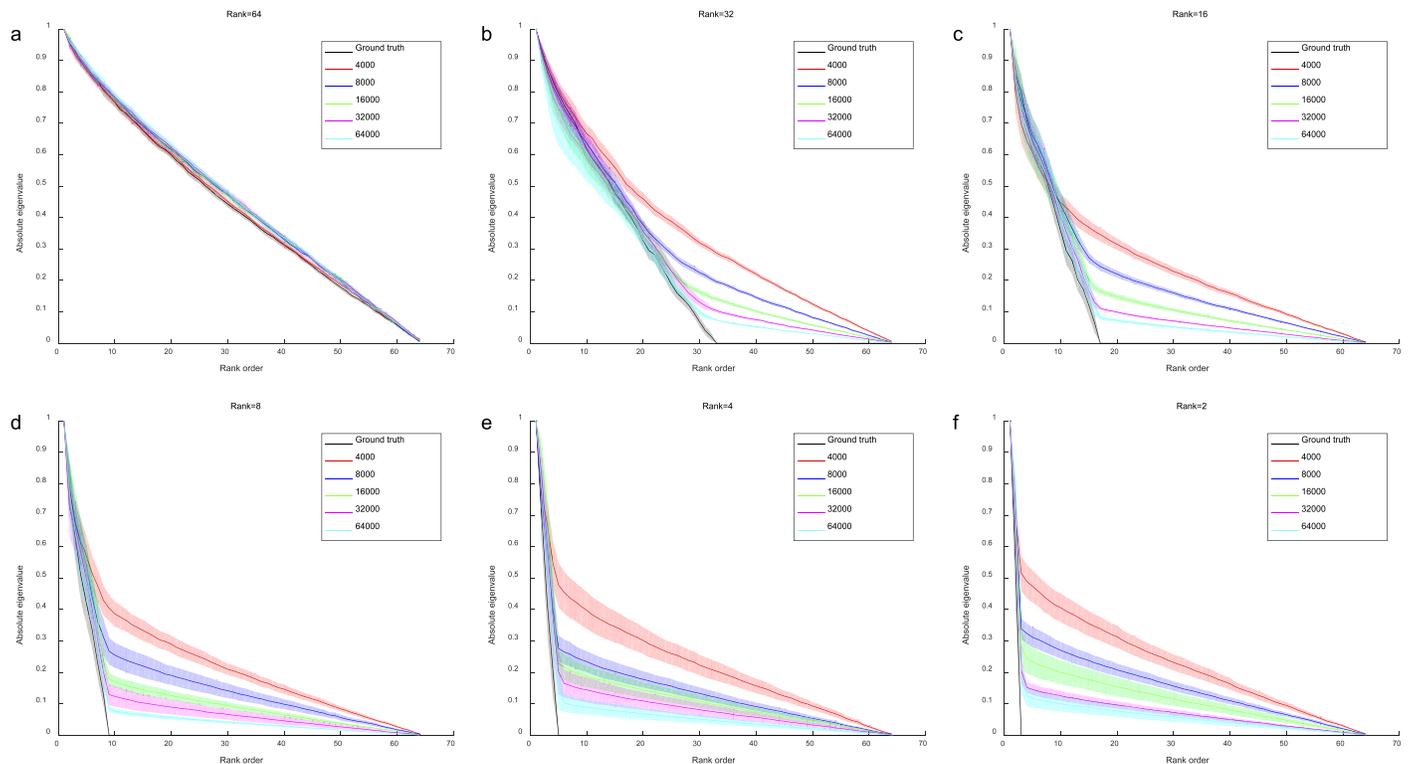


Figure 7. Sorted eigenvalues (absolute value) of the estimated second-order kernel matrix. The colored lines are the results of estimations using data of 4,000, 8,000, 16,000, 32,000, and 64,000 samples. (a–f) Rank of the second-order kernel matrix: 64 (full rank), 32, 16, 8, 4, and 2, respectively.

the number of trials at 8,000, 16,000, and 32,000, respectively). Estimation with our method (red lines) became worse as the rank of second-order kernel matrix became lower, whereas GLM (blue lines) did not show such a large degradation with changes in rank. Figure 7 shows the eigenvalues (absolute values) of the estimated second-order kernel matrix when the number of trials changed from 4,000 to 64,000 (Figure 7a–f depict the sorted eigenvalues when the rank was 64 (full rank), 32, 16, 8, 4, and 2, respectively). The results indicate that the estimated eigenvalues that were supposed to be zero did not achieve zero values by our method, even as the number of trials increased. This is due to a limitation of our method that truncates the calculation of moments up to the second-order term, whereas GLM does not rely on such an approximation. However, because small eigenvalues are relatively unimportant and true eigenvalues related to the system parameter are distinguishable from erroneous values in many cases, as seen in Figure 7, we can practically solve the deficits of our method by truncating small eigenvalues to zero. The magenta line in Figure 6 depicts the performance of our method when truncating small eigenvalues at postprocessing; the results were comparable with those provided by the GLM, regardless of the rank of the second-order kernel matrix. Figure 6d–f shows the change in estimation perfor-

mance as a function of the sample size, when the rank is 2, 4, and 8, respectively. Although GLM provides significantly better estimates than ours in term of accuracy, if the sample size is sufficiently large, then the calculation with GLM using many data samples takes a very long time. A modified version of our method gives very accurate estimations when the data sample size is large and gives even better estimates than GLM when the sample size is limited.

In the first simulation using the hybrid neural model, we used data from 250,000 trials to obtain reasonable estimates of the second-order kernels (Figure 3b). Figure 3e shows the MSEs of estimations of GLM and our method after adjusting the scale of kernels to minimize each MSE. GLM showed significantly better estimation accuracy than our method when we applied our method straight forwardly. The estimation was improved by truncating the eigenvalues except for largest two, despite using fewer data samples (25,000 trials: i.e., 10 times fewer samples, Figure 3d). The correlation between estimated kernel and ground truth was very high ( $R^2 = 0.924$ , which was statistically indistinguishable from the estimation without truncating using 10-fold larger data sets ( $p < 0.01$ , Meng's Z-test). Figure 3f shows that our method with eigenvalue truncation provided significantly better estimation than GLM when sample size was 25,000 ( $t$  test,  $p < 0.001$ ).

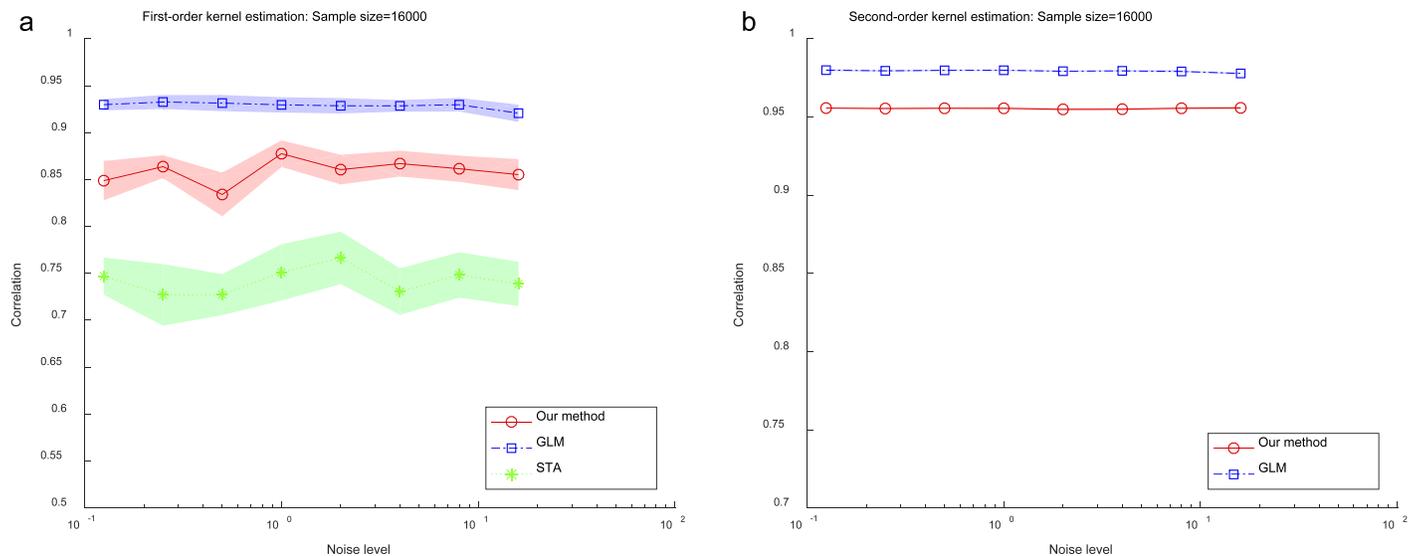


Figure 8. Correlation of the kernels between the ground truth and the estimates, plotted as a function of the inner noise variance at the stage of binary decision. Red, blue, and green lines are the results of the estimation by our method, GLM, and STA, respectively: (a) first-order kernel estimation and (b) second-order kernel estimation.

### Effect of inner noise variance

In theory, the estimate of our method does not rely on whether the actual value of the noise variance at the stage of binary decision is  $1/2$  or not if the scale of the kernels is not considered. To test this, we simulated how the estimation performance changed depending on the inner noise variance of a system. As can be seen in Figure 8, the estimation accuracy of our method did not change significantly over a wide range of noise variance.

## Psychophysical experiments

To validate our method with actual data, we conducted the following psychophysical experiments. We analyzed data for 144 dimensions and estimated the second-order kernels in which the number of parameters exceeded 2,000. In such a case of relatively large-scale parameter estimation, GLM calculations become very time-consuming.

### Observers

Author OW and two naïve observers with normal or corrected-to-normal vision participated in our experiments. Experiments were conducted in accordance with the principles embodied in the Declaration of Helsinki under the approval of the ethics committee of the Muroran Institute of Technology.

### Apparatus

Visual stimuli were presented on a gamma-corrected CRT monitor (Dell P992, Dell, Inc., Round Rock, TX) using OpenGL in the C programming environment. Observers sat 1 m away from the monitor in a dark room, and their head movements were restricted using chin rests. The monitor resolution was set to  $1,024 \times 768$  pixels ( $50 \text{ pixel}/^\circ$ ), and the refresh rate was set to 120 Hz.

### Visual stimuli and task

At the beginning of each trial, a fixation point (bull's-eye marker, outer diameter =  $0.18^\circ$ ) was displayed in the center of a blank screen (uniform gray,  $42.24 \text{ cd}/\text{m}^2$ ). Then, 500 ms after the button was pressed, the fixation point disappeared, leaving only the blank screen for 200 ms. Moving sinusoidal gratings were then displayed on the monitor for 200 ms. After a 200-ms blank period, a fixation point reappeared on the monitor prior to the button-press response that triggered the next trial. Observers were instructed to judge whether they perceived the gratings to be moving upward or downward, as a whole, and indicate their observations by pressing a 2AFC button.

We preset 144 different moving sinusoidal gratings that moved either upward or downward, each with a spatial frequency consisting of one of eight values (spanning the range  $0.13$ – $1.47 \text{ c}/^\circ$  in 0.5-octave steps) and a temporal frequency consisting of one of nine values (spanning the range:  $1.5$ – $24.0 \text{ Hz}$  in 0.5-octave steps). For simplicity, we denoted upward-moving

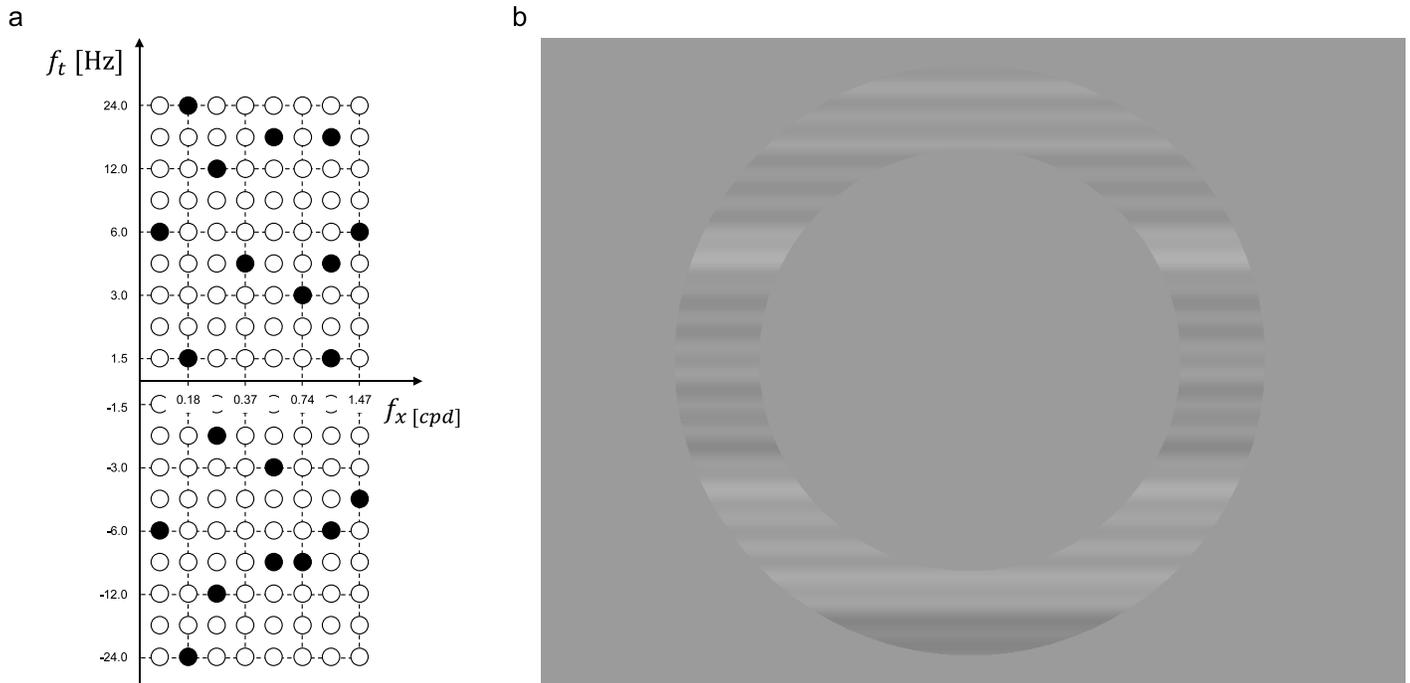


Figure 9. (a) The schema of generating visual stimulus in the psychophysical experiments. We sampled the spatiotemporal frequency space using a log scale step: 144 points = 8 points of the spatial frequency ( $f_s$ ) x 18 points of the temporal frequency ( $f_t$ ). Note that positive temporal frequency represents upward motion, whereas negative temporal frequency represents downward motion. We randomly selected 20 samples from the 144 possible points during each trial and generated the moving gratings so that they had corresponding frequencies. (b) An example of visual stimulus used in the psychophysical experiments. Twenty gratings with randomly chosen contrast and phase were summed and displayed at each trial through an annulus window. The contrast and phase of each grating were also randomly selected in the range of [0%–5%] and  $[-\pi\pi]$ , respectively.

gratings as having positive temporal frequencies and downward-moving gratings as having negative temporal frequencies (8 spatial frequencies  $\times$  18 temporal frequencies). We then randomly sampled 20 gratings from the 144 candidate gratings in each trial, and superimposed them to display as stimulus (see the schema of generating stimulus in Figure 9a). The contrast and phase of each grating were set in a uniformly random manner in the range of [0%–5%] and  $[-\pi\pi]$ , respectively. To prevent our observers from actively tracking a particular grating, gratings were displayed at the peripheral visual field as an annulus (eccentricity ranging from  $5.0^\circ$  to  $7.0^\circ$ ), as depicted in Figure 9b. Each observer completed 12,500 trials (125 trials/session, 4 sessions/day, for a total of 100 sessions over 25 days). Data from three observers (37,500 trials) were pooled for the analysis. No strong response bias was observed [45.6% upward report (46.5%, 42.2%, and 48.2% for three observers, respectively)].

### Analysis

We regarded “upward” reports as  $y = 1$ , and “downward” reports as  $y = -1$ . The contrast values of 144 gratings correspond to input vector  $x$  of 144

dimensions, varying from 0 to  $2.11 \text{ [cd/m}^2\text{]}$ . We denote the nonlinear model of observers’ assessments of motion direction using frequency values instead of using the index number of moving gratings in order to make the first- and second-order kernels interpretable in the frequency space as follows:

$$F(x) = l + \sum_{f_s, f_t} k(f_s, f_t) \cdot x(f_s, f_t) + \sum_{f_s < f_s', f_t < f_t'} h(f_s, f_t, f_s', f_t') \cdot x(f_s, f_t) \cdot x(f_s', f_t'), \tag{10}$$

where  $f_s$  and  $f_s'$  are the spatial frequency, and  $f_t$  and  $f_t'$  are the temporal frequency of the gratings,  $x(f_s, f_t)$  is the contrast value of the grating whose spatiotemporal frequencies are  $f_s$  and  $f_t$ . We identified the zero-order ( $l$ ), first-order ( $k(f_s, f_t)$ ), and second-order ( $h(f_s, f_t, f_s', f_t')$ ) kernels using Eq. SSS2 in Appendix. Regularization was applied, as described in the Appendix, because in a real experiment, the ground truth of the kernels is unknown. The regularization parameter  $\lambda$  was set to 0.01, which provided the best prediction performance ( $R = 0.79$ ) in a 10-fold crossvalidation test. The results shown in the latter session, however, are basically the same, regardless of regularization.

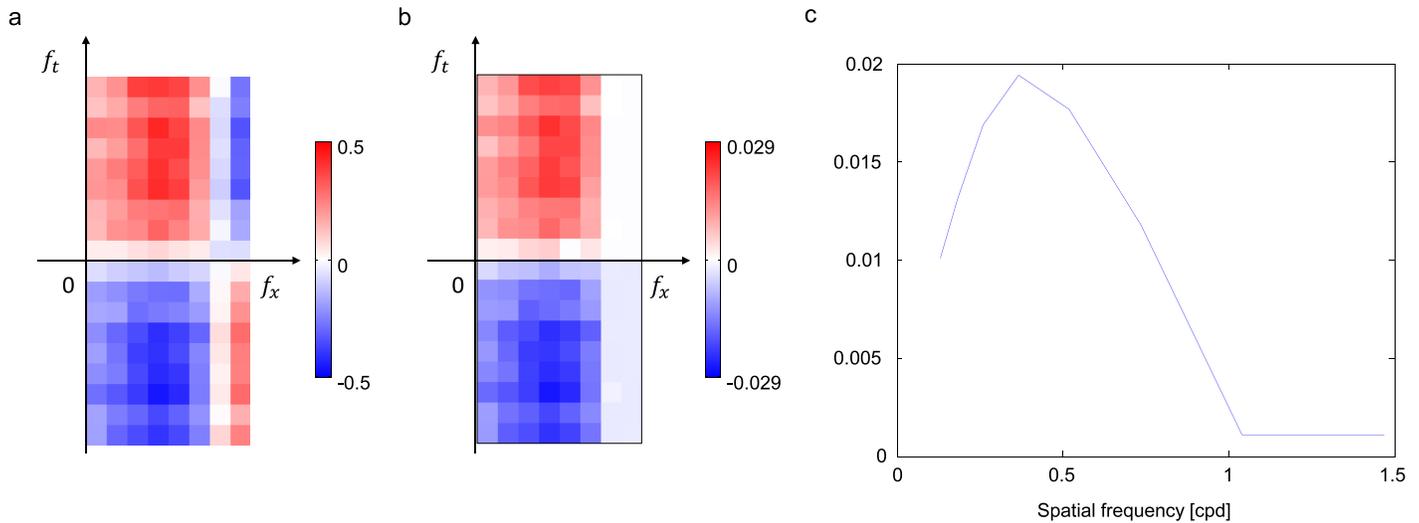


Figure 10. (a) The estimated first-order kernel plotted in the frequency space. The upper half of the space corresponds to the kernel for upward moving gratings, and vice versa. The color of each pixel is scaled depending on the size of the kernel, as shown in the scale bar. The color red indicates that the corresponding moving grating contributes to the assessment of upward motion, and the color blue indicates that the corresponding moving grating contributes to the assessment of downward motion. (b) The estimated first-order kernel with non-negative constraint derived from prior knowledge about the first-order kernel of motion perception. (c) The first-order kernel values as a function of spatial frequency (Mean across temporal frequencies).

## Experimental results

### First-order kernels

Figure 10a shows the estimated first-order kernel plotted in the frequency space. The color of each pixel represents the coefficient value of the first-order kernel of the corresponding grating in motion. The color red indicates that the motion of the corresponding grating contributes to the assessment of upward motion, and the color blue indicates that the corresponding grating contributes to the assessment of downward motion. Our results revealed that moving gratings with low spatial frequency have excitatory effects on motion direction assessment, whereas gratings with high spatial frequency have inhibitory effects (i.e., a high spatial frequency grating that is moving upward facilitates a downward report and vice versa). The first-order kernel is mainly modulated as a function of spatial frequency, and the modulation by temporal frequency is relatively broad.

### Second-order kernels

Figure 11a shows the estimated second-order kernel plot. Each panel in Figure 11a corresponds with the second-order kernel plotted in the frequency space of  $(f_s, f_t)$  when  $(f_s, f_t)$  is fixed and is arranged depending on the  $(f_s, f_t)$  values. The pixels surrounded by dotted lines indicate the value of the second-order kernel when

$(f_s, f_t) = (f'_s, f'_t)$  (i.e., the effect of the squared contrast of a single grating on observers' assessment of motion).

### Kernel estimation with constraint of prior knowledge

Although our method does not impose any prerequisite regarding the stimulus distribution, this does not necessarily mean that our method is always able to estimate the kernels close to ground truth using any stimulus. Whatever the method used, if the inputs are not sampled appropriately to capture all aspects of the system's behavior, then the estimated model would be merely functional and valid only over the range of the stimuli sampled. The first-order kernel estimated in the previous section indicates that signals with low spatial frequency facilitate motion assessment consistent with their direction, whereas signals with high spatial frequency encourage motion assessment opposite to the signals' direction. Therefore, when viewing only a single grating at a time, the estimated model predicts that a grating with high spatial frequency will always be perceived to move in opposition to its original direction of motion, which does not actually happen for the tested frequency range. It is likely that the inhibitory effect of high spatial frequencies is estimated as a first-order effect because in our experiments, high and low spatial frequency gratings are displayed together in most cases, and second-order interaction between high and low spatial frequencies steadily affects the observer's assessment of movement direction. It is not

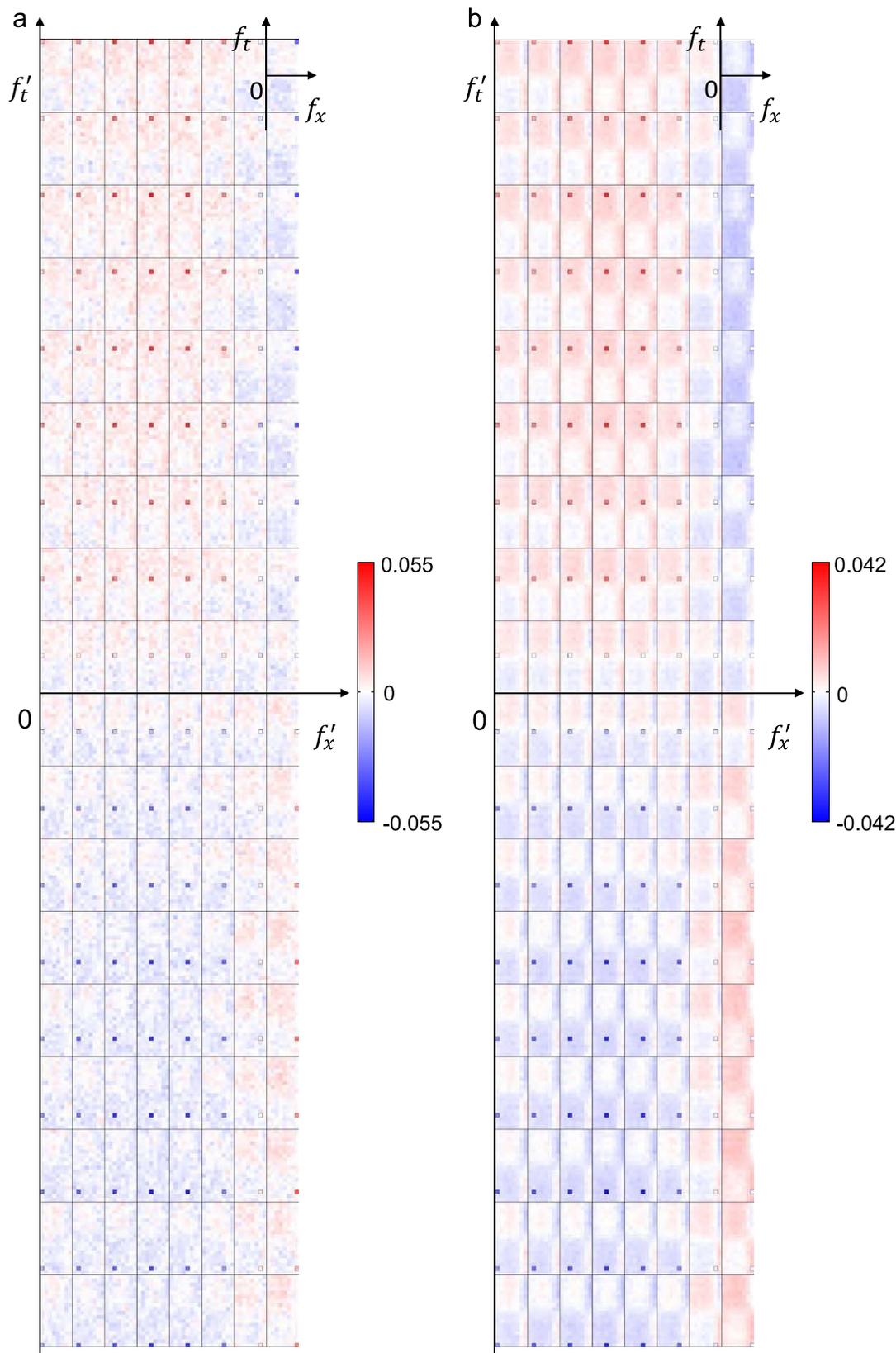


Figure 11. (a) The estimated second-order kernel plot. The figure consists of  $8 \times 18$  panels, segmented by black lines. Each panel depicts the second-order kernel plotted in the frequency space of  $(f_s, f_t)$  when  $(f'_s, f'_t)$  is fixed. Panels are arranged at the location depending on the  $(f'_s, f'_t)$ . Although second-order kernels are not defined in the domain where  $f'_s > f_s$  or  $f'_t > f_t$  in our method, we duplicate  $h(f_s, f_t, f'_s, f'_t) = h(f'_s, f'_t, f_s, f_t)$  for visibility. The pixels surrounded by the dotted lines indicate the value of the second-order kernel when  $(f_s, f_t) = (f'_s, f'_t)$ , i.e., the effect of the squared contrast of a single grating on direction assessment. The color of each pixel is scaled depending on the size of the kernel, as shown in the scale bar. (b) The estimated second-order kernel with the non-negative constraint derived from prior knowledge about the first-order kernel of motion perception.

feasible, however, to characterize the nonlinear model of motion processes by sampling all possible stimulus spaces, such as by randomly changing the number of sampled gratings at each trial, due to the explosion of search space and the number of completed trials necessary.

Instead of testing all possible stimuli in our experiment, we can reasonably assume that the first-order kernels and the squared terms of upward-moving gratings are positive, while those of downward-moving gratings are negative, and use such prior knowledge as a constraint with which to calculate the kernels (i.e., solve the constrained linear least-square problem of Eq. 8 or Eq. SSS2; Coleman & Li, 1996). Although estimating the kernels under this constraint requires an iterative optimization algorithm, the size of the data set to be optimized is already reduced in our framework. Thus, implementing a non-negative constraint in our method still has advantages in saving calculation time versus GLM. STA/STC does not provide a framework to estimate the kernels with such a constraint.

Figure 10b shows the first-order kernels estimated with this constraint of prior knowledge (the best prediction performance was  $R = 0.79$  in a 10-fold crossvalidation test when regularization parameter  $\lambda = 0.1$ ). The estimated values are positive for upward-moving gratings (positive temporal frequency) and vice versa, as expected. The bandwidth (half width at half maximum) of spatial frequency tuning of the first-order kernel is about 1 octave (Figure 10c). Figure 11b shows the second-order kernels estimated with the same constraint. We can see the clear coupling between the low spatial frequency gratings moving in one direction and the high spatial frequency gratings moving in the opposite direction, as well as negative coupling with high spatial frequency gratings moving in the same direction.

### Eigenvalue decomposition analysis of the second-order kernel matrix

When we denote the second-order kernels in matrix form as  $H$ , the second order contribution of input  $\mathbf{x}$  to the assessment of direction can be expressed as

$$\mathbf{x}^T \cdot H \cdot \mathbf{x} \quad (11)$$

If Eq. 11 becomes positive, then it contributes to the assessment of upward motion, and vice versa.

If matrix  $H$  has an eigenvector  $\mathbf{v}$  of an eigenvalue  $\lambda'$ , then they satisfy the following equation:

$$\mathbf{v}^T \cdot H \cdot \mathbf{v} = \lambda' \quad (12)$$

Therefore, the combination of the moving gratings whose contrasts are proportional to the coefficients of an eigenvector contributes to motion direction assessment consistent with the sign of an eigenvalue to

the extent of its absolute value. Figure 12a depicts the sorted eigenvalues (absolute value) of the second-order kernel matrix estimated by our method with the non-negative constraint. Only two eigenvalues take large values in the matrix. Figure 12b shows the six eigenvalues and their eigenvectors plotted in such a way as to correspond with the frequency space. The upper row indicates the top three positive eigenvalues and their eigenvectors, and the lower row indicates the top three negative eigenvalues and their eigenvectors. Note that the eigenvectors have an arbitrary property in terms of a constant factor. Thus, the sign of each pixel in the eigenvector plot is not directly related with motion direction assessment. The results show that, when low spatial frequency gratings moving upward and high spatial frequency gratings moving downward are observed together, they increase reports of upward movement, and the high spatial frequency gratings in upward motion are negatively associated with these effects. The opposite is true for the eigenvector of the smallest negative eigenvalue and the assessment of downward motion. There is no meaningful pattern in the rest of the eigenvectors of relatively small eigenvalues. The results (Figure 12a) also indicate that meaningful eigenvalues are distinguishable from other noisy values in the actual data, supporting that the modification of our method that truncates small eigenvalues is practically useful.

## Discussion

### Watanabe method

In this paper, we implemented a new kernel estimation method based on the assumption that the internal noise of a system follows Gaussian distribution and the response function can be described as a cumulative Gaussian. Although this is a common assumption used for modeling studies, human psychophysical studies suggest that the actual distribution of intrinsic perceptual noise has higher kurtosis than Gaussian (Neri, 2013). One possible modification of our method is to use the Cauchy distribution/arctangent function as an intrinsic noise model/response function and derive estimation formula based on its Maclaurin series expression. It will be interesting to compare the effects of different response functions on psychophysical data analysis.

Previous psychophysical studies have shown that the internal noise variance is closer to 2 (Neri, 2010), and is even non-stationary in a given experiment (Goris, Movshon, & Simoncelli, 2014). Thus, it is conceivable that the assumption regarding the internal noise

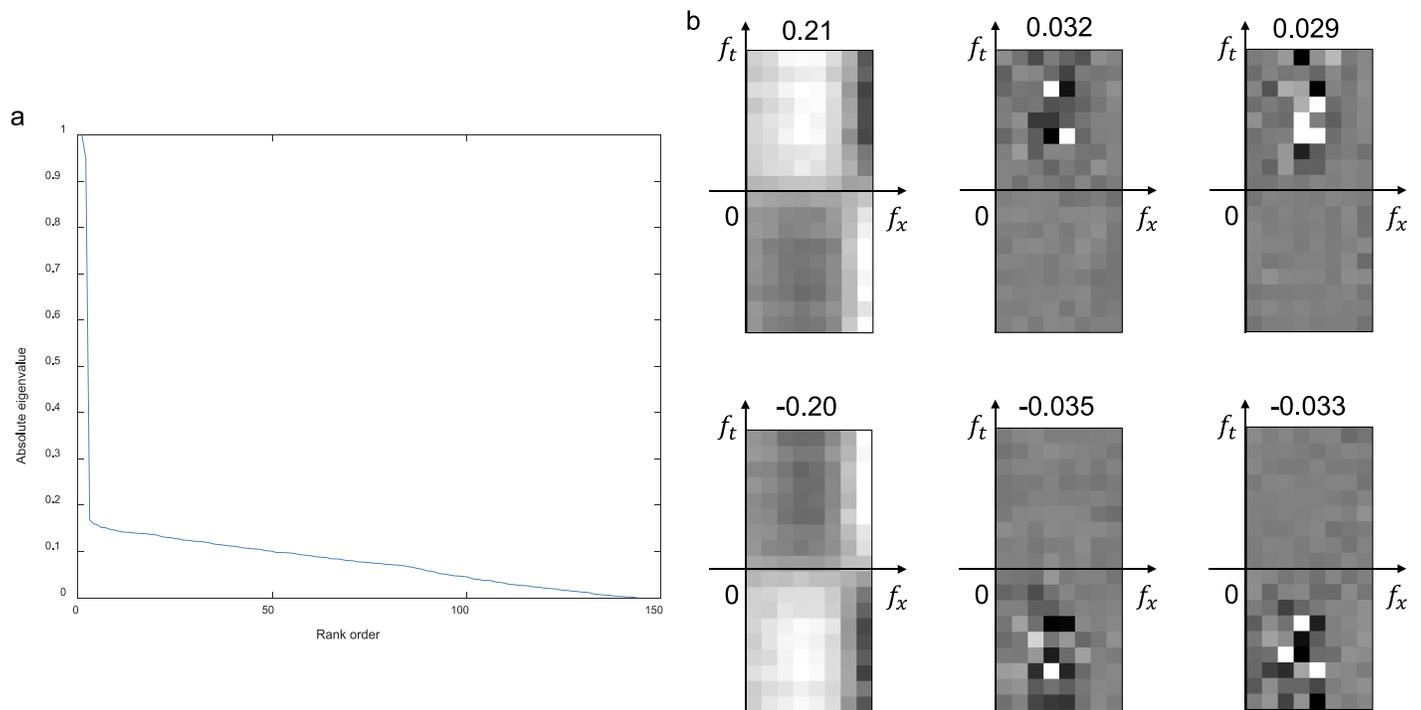


Figure 12. (a) Plot of eigenvalues (absolute value) of the second-order kernel matrix. (b) The results of eigenvalue decomposition analysis of the second-order kernel matrix. The figures show the three largest and three smallest eigenvalues and the corresponding eigenvectors arranged in the frequency space. Figures in the top row are eigenvector plots for positive eigenvalues, showing the combinations of gratings that facilitate the upward response, whereas the figures in the bottom row are those for negative eigenvalues that facilitate the downward response. The top left plot for the largest positive eigenvalue indicates that low spatial frequency gratings moving upward and high spatial frequency gratings moving downward work together to facilitate an assessment of upward motion, whereas high spatial frequency gratings moving upward are negatively related with such effects. The bottom left plot for the smallest negative eigenvalue is a horizontally flipped version of the pattern in the top left plot.

variance of  $1/2$  in our method will not hold true for many systems, which may lead to inaccurate kernel estimates. However, in our method, the internal noise variance affects only the entire scale, but not the shape of the estimated kernels, in theory. This is validated by the simulation showing that estimation by our method is accurate over a wide range of inner noise variance.

In actual experiments, observers' response biases usually change over time, whereas our method estimates only the average bias as a zero-order kernel. It will be interesting to extend our method considering such bias changes as independent parameters, as shown in Neri (2004), in future work.

It is conceivable that our method is a family of GLMs in the sense that our method covers any nonlinear system and uses a cumulative Gaussian as a link function. The difference is that our method truncates the moment calculation up to a certain order (in this paper, second-order approximation) to obtain analytical estimates of nonlinear kernels, whereas GLM does not rely on such an approximation, at the cost of the iterative optimization calculation. Thus, as a tradeoff for the lower calculation cost, our method showed a severe reduction in estimation accuracy when

the rank of the second-order kernel matrix is low. This is because as the rank becomes lower, the number of informative inputs consistent with the eigenvectors of the system becomes fewer and the approximation error increases. We showed in our simulation that truncating small eigenvalues of the second-order kernel matrix at post-processing is a practical solution to this issue.

GLM usually provides better kernel estimation than our method in terms of accuracy if sufficiently large data samples are available. Our method, however, can provide similar and sufficiently accurate results with a much faster calculation time. Our method will be useful especially in cases when visualizing kernels quickly or handling very large data sets to evaluate nonlinear interactions in very large-scale systems.

In contrast to GLM, our method has better scalability, based on the formula of updating the previously estimated inverse matrix with additional data. Such extension will be useful when changing input stimuli in an adaptive way online during experiments or when the data size is too large to handle and smaller fractions of the data are required for kernel estimation.

Besides GLM, recent studies proposed other maximum likelihood estimators for estimating neural system

(Mcfarland, Cui, & Butts, 2013; Paninski, 2004; Park, Archer, Priebe, & Pillow, 2013; Truccolo, Eden, Fellows, Donoghue, & Brown, 2004) by implementing a specific property of neural processing into model. Although these maximum likelihood estimators have advantages in terms of accuracy for the target neural system and in calculation speed, they still rely on optimization algorithm using iterative calculation. Our method targets not only to neural process but also any nonlinear system that can be described as Volterra series expansion and output binary response in general.

## Psychophysical experiments

The kernels estimated by STA and GLM showed similar results to those obtained by our method. However, STA can only estimate first-order kernels, and calculations with GLM were time-consuming compared with our approach.

The second-order kernel of visual motion processing estimated by our method with natural constraint indicates that signals with low spatial frequency encourage an assessment of motion that is consistent with their direction, whereas signals with high spatial frequency encourage an assessment of motion opposite their actual direction. This finding is consistent with the results of previous studies. The direction perception of a low spatial frequency grating is facilitated by a grating of three-octaves-higher spatial frequency moving in the opposite direction and inhibited by its motion in the same direction (Serrano-Pedraza, Goddard, & Derrington, 2007). It has been also shown that low spatial frequencies dominate ongoing motion perception, and high spatial frequencies are negatively related to the dominant low spatial frequencies (Derrington & Henning, 1989; Hayashi et al., 2010; Nishida, Yanagi, & Sato, 1995). Although we identified this model of motion direction assessment from the observers' responses to 20 different moving gratings presented simultaneously, the same model (with a slight modification in setting, a zero-order kernel = 0) can duplicate several findings in previous research when viewing only one, two, or three gratings. Contrary to previous studies, our study does not determine high and low frequencies in relation to other frequencies, but rather is fixed in the tested frequency range due to the limited stimulus sampling in our model.

It is also noteworthy that the first-order kernel varies mainly depending on spatial frequency and is relatively insensitive to temporal frequency. Previous studies of the spatiotemporal frequency mechanism in human vision revealed that spatial frequency processing is served by a moderately large number (at least six or seven distinct classes; Watson & Robson, 1981; Wilson, McFarlane, & Phillips, 1983) of mechanisms or

detectors, each with relatively limited spatial bandwidths (about  $\pm 1$  octave (Blackmore & Campbell, 1969; De Valois, 1977; Watson & Robson, 1981). Temporal frequency, on the other hand, is considered to be processed by only two or three mechanisms, each with much broader temporal bandwidths (Hess & Plant, 1985; Hess & Snowden, 1992; Watson & Robson, 1981). Our findings, namely that the first-order effect for motion direction assessment is modulated by a spatial frequency mechanism with the bandwidth of  $\pm 1$  octave separation and a temporal frequency mechanism with a much broader bandwidth, show good consistency with these previous studies.

Together with the performance of our method being as high as GLM in numerical simulations, these findings support the validity of our method for the quantitative characterization of nonlinear processes underlying neural responses and psychophysical decisions.

## Summary

We proposed a novel analytical method to estimate higher-order kernels of a nonlinear system whose outputs take only binary states, such as the case of neural responses and human observers' reports in 2AFC tasks. Our method has several advantages over conventional RC/STA and STC techniques. (1) Our method does not rely on assumptions about a specific stimulus distribution. (2) Our method can estimate nonlinear kernels, except their scale. (3) Our method provides better estimations over a wide range of parameters. The advantage over GLM is that (4) the nonlinear kernels are analytically derived from data without an iterative calculation and similar results to GLM can be obtained much more quickly. (5) Our method also has better scalability than GLM, using sequential updating of an inverse matrix. Although the estimation accuracy of our method is reduced severely when the rank of the second-order kernel matrix is low, we can practically obtain an estimation as close to the ground truth as GLM in many cases, by truncating small eigenvalues to zero.

To validate our method, we conducted psychophysical experiments with visual motion to demonstrate the application of our method to real data. The results showed that assessment of motion direction is dominated by moving gratings of low spatial frequency in terms of first-order effect, while the gratings of high spatial frequency affect motion assessment in the opposite direction in the second-order term, consistent with previous studies. Spatiotemporal frequency tuning of the estimated model is consistent with the previous findings regarding the property of spatial and temporal frequency channels. Moreover, the second-order interaction between low and high spatial frequency channels

explains the motion capture or motion contrast effect reported in previous studies. Both numerical simulation and psychophysical experiments support the validity of our method to estimate the nonlinear properties of neural and/or perceptual processes.

*Keywords:* system identification, non-linear kernel estimation, GLM, reverse correlation, moment-generating function, Volterra series, nonlinear systems

## Acknowledgments

Osamu Watanabe (1972–2015) conceived the basic idea of Watanabe’s method and published a preliminary report in the form of a conference proceeding in Japanese (Watanabe, 2011). He also conducted psychophysical experiments on motion perception while consulting with RH and SN. RH ran numerical simulation, analyzed and interpreted Watanabe’s computer data, and evaluated and wrote the manuscript in consultation with HY and SN. HY refined the method and implemented codes for numerical simulation. We thank the bereaved family of Osamu Watanabe for giving us permission to use his data and research records, which were indispensable in writing this paper. We also thank Dr. Shotaro Akaho for his advice and comments about the extension of Watanabe’s method with regularization. RH was supported by MEXT KAKENHI (Grant-in-Aid for Scientific research) Grant numbers JP26119533, JP25351005 and JSPS KAKENHI Grant number JP16K13117 and a NEDO Grant in Aid of Research. SN was supported by MEXT KAKENHI Grant numbers JP15H05915.

\*RH, OW, and HY contributed equally to this article. Commercial relationships: none.

Corresponding author: Ryusuke Hayashi.

Email: r-hayashi@aist.go.jp.

Address: Systems Neuroscience Group, Human Informatics Research Institute, National Institute of Advanced Industrial Science and Technology, Tsukuba, Ibaraki, Japan.

## References

- Adelson, E. H., & Bergen, J. R. (1985). Spatiotemporal energy models for the perception of motion. *Journal of Optical Society of America A*, 2(2), 284–299.
- Ahumada, A. J. (1996). Perceptual classification images from Vernier acuity masked by noise. *Perception*, 25, 18–18.
- Beard, B. L., & Ahumada, A. J. (1998). A technique to extract relevant image features for visual tasks. *Proceedings of SPIE*, 3299, 79–85.
- Blackmore, C., Campbell, F. W. (1969). On the existence of neurons in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology*, 203, 237–260.
- Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12, 199–213.
- Coleman, T. F., Li, Y. (1996). A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM Journal on Optimization*, 6(4), 1040–1058.
- David, S., Vinje, W. E., Gallant, J. L. (2004). Natural stimulus statistics alter the receptive field structure of V1 neurons. *The Journal of Neuroscience*, 24(31), 6991–7006.
- DeAngelis, G.C., Ohzawa, I., Freeman, R. D. (1993a). Spatiotemporal organization of simple cell receptive fields in the cat’s striate cortex. I. General characteristics and postnatal development. *Journal of Neurophysiology*, 69, 1091–1117.
- DeAngelis, G. C., Ohzawa, I., Freeman, R. D. (1993b). Spatiotemporal organization of simple cell receptive fields in the cat’s striate cortex. II. Linearity of temporal and spatial summation. *Journal of Neurophysiology*, 69, 1118–1135.
- Derrington, A. M., & Henning, D. G. (1989). Some observations on the masking effects of two-dimensional stimuli. *Vision Research*, 29, 241–246.
- De Valois, K. K. (1977). Spatial frequency adaptation can enhance contrast sensitivity. *Vision Research*, 17, 1057–1065.
- Franz, M. O., & Shölkopf, B. (2003). Implicit Wiener Series—Part I: Cross-Correlation vs. Regression in Reproducing Kernel Hilbert Spaces, *Max-Planck-Institute Technical Report No. TR-114*.
- Goris, R. L., Movshon, J. A., & Simoncelli, E. P. (2014). Partitioning neuronal variability. *Nature Neuroscience*, 17(6), 858–865.
- Hayashi, R., Sugita, Y., Nishida, S., & Kawano, K. (2010). How motion signals are integrated across frequencies: Study on motion perception and ocular following responses using multi-slit stimuli. *Journal of Neurophysiology*, 103, 230–243. [PubMed] [Article]
- Heeger, D. J. (1991). Nonlinear model of neural responses in cat visual cortex. In M. S. Landy & J. A. Movshon (Eds.), *Computational models of visual processing* (pp. 119–133). Cambridge, MA: Bradford Books, MIT Press.

- Hess, R. F., & Plant, G. T. (1985). Temporal frequency discrimination in human vision: Evidence for an additional mechanism in the low spatial and high temporal frequency region. *Vision Research*, *25*, 1493–1500.
- Hess, R. F. & Snowden, R. J. (1992). Temporal properties of human visual filters: Number, shape and spatial covariation. *Vision Research*, *32*, 47–59.
- Hida, E., & Naka, K. (1982). Spatio-temporal visual receptive fields as revealed by spatio-temporal random noise. *Zeitschrift für Naturforschung C*, *37*, 1048–1049.
- Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, *160*, 106–154.
- Jones, J. P., Palmer, L. A. (1987). The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of Neurophysiology*, *58*(6), 1187–1211.
- Knoblauch, K., & Maloney, L. T. (2008). Estimating classification images with generalized linear and additive models. *Journal of Vision*, *8*(16):10, 1–19, doi:10.1167/8.16.10. [PubMed] [Article]
- McFarland, J. M., Cui, Y., Butts, D. A. (2013). Inferring nonlinear neural computation based on physiologically plausible inputs. *PLoS Computational Biology*, *9*(7), e1003143, 1–18.
- Nelder, J., Wedderburn, R. (1972). Generalized linear models, *Journal of the Royal Statistical Society, Series A*, *135*(3), 370–384.
- Neri, P. (2004). Estimation of nonlinear psychophysical kernels. *Journal of Vision*, *4*(2):2, 82–91, doi:10.1167/4.2.2. [PubMed] [Article]
- Neri, P. (2010). How inherently noisy is human sensory processing? *Psychonomic Bulletin & Review*, *17*, 802–808.
- Neri, P. (2013). The statistical distribution of noisy transmission in human sensors. *Journal of Neural Engineering*, *10*(01), 016014.
- Nishida, S., Yanagi, J., & Sato, T. (1995). Motion assimilation and contrast in superimposed gratings: Effects of spatiotemporal frequencies. Presented at the ARVO annual meeting, Ft. Lauderdale, FL, 1995.
- Ohzawa, I., DeAngelis, G. C., & Freeman, R. D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science*, *246*(4972), 1037–104.
- Orcioni, S. (2014). Improving the approximation ability of Volterra series identified with a cross-correlation method. *Nonlinear Dynamics*, *78*, 2861–2869.
- Paninski, L. (2003). Convergence properties of some spike-triggered analysis techniques. *Network: Computation in Neural Systems*, *14*, 437–464.
- Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, *15*, 243–262.
- Park, I. M., Archer, E., Priebe, N., & Pillow, J. W. (2013). Spectral methods for neural characterization using generalized quadratic models. *Advances in Neural Information Processing Systems*, *26*, 2454–2462.
- Reid, R. C., & Alonso, J. M. (1995). Specificity of monosynaptic connections from thalamus to visual cortex. *Nature*, *378*(6554), 281–284.
- Ringach, D., & Shapley, R. (2004). Reverse correlation in neurophysiology. *Cognitive Science*, *28*, 147–166.
- Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, *5*, 517–548.
- Schetzen, M. (1980). *The Volterra and Wiener theories of nonlinear systems*. New York: John Wiley & Sons.
- Schwartz, O., Chichilnisky, E. J., & Simoncelli, E. P. (2002). Characterizing neural gain control using spike-triggered covariance. *Advances in Neural Information Processing Systems*, *14*, 269–276.
- Serrano-Pedraza, I., Goddard, P., Derrington, A. M. (2007). Evidence for reciprocal antagonism between motion sensors tuned to coarse and fine features. *Journal of Vision*, *7*(12):8, 1–14, doi:10.1167/7.12.8. [PubMed] [Article]
- Sharpee, T., Rust, N. C., Bialek, W. (2003). Maximally informative dimensions: Analyzing neural responses to natural signals. *Advances in Neural Information Processing Systems*, *15*, 277–284.
- Simoncelli, E. P., Paninski, L., Pillow, J., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. In M. Gazzaniga (Ed.), *The cognitive neuroscience* (3rd ed., pp. 327–338). Cambridge, MA: MIT Press.
- Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P. & Brown, E. N. (2004). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of Neurophysiology*, *93*(2), 1074–1089.
- Vinje, W. E., Gallant, J. L. (2000). Sparse coding and decorrelation in primary visual cortex during natural vision. *Science*, *287*, 1273–1276.
- Watanabe, O. (2011). Estimating psychophysical decision processes with arbitrarily distributed stimuli

[in Japanese]. *Technical Report of IEICE*, 110(461), 319–324.

Watson, A. B., & Robson, J. G. (1981). Discrimination at threshold: Labeled detectors in human vision. *Vision Research*, 21, 1115–1122.

Weisstein, E. W. (2017, updated). Moment-generating function, from MathWorld—A Wolfram web resource. Retrieved from <http://mathworld.wolfram.com/Moment-GeneratingFunction.html>

Wilson, H. R., McFarlane, D. K., & Phillips, G. C. (1983). Spatial frequency tuning of orientation selective units estimated by oblique masking. *Vision Research*, 23, 873–882.

## Appendix

### Derivations of Watanabe method using the moment-generating function

The Maclaurin series expression of Gauss' error function  $\text{erf}(x)$  can be given by Eq. S1 below:

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n}{n!(2n+1)} x^{2n+1} \quad (\text{S1})$$

On the other hand, the moment-generating function is known as an alternative specification of probability distribution in statistics. The moment-generating function of  $P(y, x)$ , or the joint probability distribution of input  $\mathbf{x}$  and output  $y$ , is written as

$$\varphi(s_y, \mathbf{s}_x) = \int e^{s_y y + \mathbf{s}_x^T \mathbf{x}} P(y, \mathbf{x}) dy d\mathbf{x} \quad (\text{S2}),$$

and the moment-generating function of  $P(x)$ , or the probability distribution of input  $\mathbf{x}$ , is written as

$$\psi(\mathbf{s}_x) = \int e^{\mathbf{s}_x^T \mathbf{x}} P(\mathbf{x}) d\mathbf{x} \quad (\text{S3}),$$

where,  $s_y$  is a scalar variable and  $\mathbf{s}_x = (s_1, s_2, \dots, s_N)^T$ .

By substituting Eq. 4 into Eq. S2 using Bayes' theorem, we can express  $\varphi(s_y, \mathbf{s}_x)$  as

$$\varphi(s_y, \mathbf{s}_x) = \int \left\{ \frac{e^{s_y y_1} + e^{s_y y_0}}{2} + \frac{e^{s_y y_1} - e^{s_y y_0}}{2} \text{erf}(F(\mathbf{x})) \right\} e^{\mathbf{s}_x^T \mathbf{x}} P(\mathbf{x}) d\mathbf{x} \quad (\text{S4}).$$

When we substitute a Maclaurin series expression of Gauss' error function (Eq. S1), a Volterra expansion of  $F(\mathbf{x})$  (Eq. 1) and a moment-generating function of  $P(\mathbf{x})$  (Eq. S3) into Eq. S4, we obtain,

$$\begin{aligned} \varphi(s_y, \mathbf{s}_x) &= A\psi \\ &+ B \left[ G_0\psi + \sum_i G_{1,i} \frac{\partial\psi}{\partial s_i} + \sum_{i \leq j} G_{2,ij} \frac{\partial^2\psi}{\partial s_i \partial s_j} + \dots \right] \end{aligned} \quad (\text{S5}),$$

where

$$\begin{aligned} A &= \frac{e^{s_y y_1} + e^{s_y y_0}}{2}, & B &= \frac{e^{s_y y_1} - e^{s_y y_0}}{\sqrt{\pi}} \\ G_0 &= c_0 \\ G_{1,i} &= c_1 F_{i,1} \\ G_{2,ij} &= \begin{cases} c_1 F_{2,ij} + 2c_2 F_{1,i} F_{1,j}, & \text{if } i \neq j \\ c_1 F_{2,ii} + c_2 F_{1,i}^2, & \text{if } i = j \end{cases} \\ &\vdots \\ c_r &= \sum_{n=0}^{\infty} \frac{(-1)^n}{n!(2n+1)} \binom{2n+1}{r} F_0^{2n+1-r} \end{aligned}$$

The moment-generating function has an important property, namely that the moments, i.e., expectations of the products of input  $\mathbf{x}$  and/or output  $y$ , are given by its partial derivatives as follows

$$E[yx_{i_1}x_{i_2} \dots x_{i_n}] = \frac{\partial^{n+1}}{\partial s_y \partial s_{i_1} \dots \partial s_{i_n}} \varphi(s_y, \mathbf{s}_x) \Big|_{s_y=0, \mathbf{s}_x=0} \quad (\text{S6}),$$

$$E[x_{i_1}x_{i_2} \dots x_{i_n}] = \frac{\partial^n}{\partial s_{i_1} \dots \partial s_{i_n}} \psi(\mathbf{s}_x) \Big|_{\mathbf{s}_x=0} \quad (\text{S7}).$$

Therefore, from Eqs. S5, S6, and S7, we obtain a relational expression between the vectors/matrix of moments  $\mathbf{a}$ ,  $\mathbf{b}$ , and  $M$ , and the coefficients of the moment-generating function  $\mathbf{g}$  as shown in Eq. 5 in the main text:

$$\mathbf{a} = \frac{y_1 + y_0}{2} \mathbf{b} + \frac{y_1 - y_0}{\sqrt{\pi}} M \cdot \mathbf{g}, \quad (5)$$

where,

$$\mathbf{a} = (E[y], E[yx_1], \dots, E[yx_N], E[yx_1^2], E[yx_1x_2], \dots, E[yx_N^2], \dots)^T$$

$$\mathbf{b} = (1, E[x_1], \dots, E[x_N], E[x_1^2], E[x_1x_2], \dots, E[x_N^2], \dots)^T$$

$$\mathbf{g} = (G_0, G_{1,1}, G_{1,2}, \dots, G_{1,N}, G_{2,11}, G_{2,12}, \dots, G_{2,NN}, \dots)^T$$

$$M = \begin{pmatrix} 1 & E[x_1] & \dots & E[x_N] & E[x_1^2] & E[x_1x_2] & \dots & E[x_N^2] & \dots \\ E[x_1] & E[x_1^2] & \dots & E[x_1x_N] & E[x_1^3] & E[x_1^2x_2] & \dots & E[x_1x_N^2] & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\ E[x_N] & E[x_1x_N] & \dots & E[x_N^2] & E[x_1^2x_N] & E[x_1x_2x_N] & \dots & E[x_N^3] & \dots \\ E[x_1^2] & E[x_1^3] & \dots & E[x_1^2x_N] & E[x_1^4] & E[x_1^3x_2] & \dots & E[x_1^2x_N^2] & \dots \\ E[x_1x_2] & E[x_1^2x_2] & \dots & E[x_1x_2x_N] & E[x_1^3x_2] & E[x_1^2x_2^2] & \dots & E[x_1x_2x_N^2] & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \\ E[x_N^2] & E[x_1x_N^2] & \dots & E[x_N^3] & E[x_1^2x_N^2] & E[x_1x_2x_N^2] & \dots & E[x_N^4] & \dots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \ddots \end{pmatrix}$$

### Whitened STA and STC

Ordinary STA is calculated as follows:

$$STA = E[\mathbf{x}|y = 1] - E[\mathbf{x}]$$

When we denote covariance matrix of  $\mathbf{x}$  as

$$Cov(\mathbf{x}) = E[(\mathbf{x} - E[\mathbf{x}])(\mathbf{x} - E[\mathbf{x}])^T],$$

then whitened STA used in the main text and (David et al., 2004) can be calculated by the following equation.

$$STA_{whitened} = Cov(\mathbf{x})^{-1}STA$$

Ordinary STC is calculated as follows.

$$STC = Cov(\mathbf{x}|_{y=1})$$

When we define the square root of  $Cov(\mathbf{x})$  as  $\sqrt{Cov(\mathbf{x})}$ , then whitened STC can be calculated by the following equation.

$$STC_{whitened} = Cov(\sqrt{Cov(\mathbf{x})}^{-1}\mathbf{x}|_{y=1})$$

### Extension of Watanabe method using the Sherman–Morrison formula

To estimate the nonlinear system’s kernel, we calculate the matrix of moments  $M$  by the approximation using Eq. 7. If we set the input data matrix  $X$  obtained by time  $t$  as  $X_{1:t}$ , the estimated matrix  $M$  at time  $t$  as  $M_t$ , and additional input data at time  $t + 1$  as  $X_{t+1}$ , then we can update/correct the moment matrix  $M_{t+1}$  as follows:

$$\begin{aligned} M_{t+1} &= \frac{1}{n_{1:t}+n_{t+1}}(X_{1:t}^T X_{1:t} + X_{t+1}^T X_{t+1}) \\ &= \frac{1}{n_{1:t}+n_{t+1}}(n_{1:t}M_t + X_{t+1}^T X_{t+1}) \end{aligned} \quad (SS1),$$

where  $n_{1:t}$  is the number of trials/data samples by time  $t$  and  $n_{t+1}$  is the number of data samples obtained at time  $t + 1$ .

Using Sherman–Morrison formula, the inverse matrix of  $M_{t+1}$  is derived from the following equation:

$$\begin{aligned} M_{t+1}^{-1} &= \frac{n_{1:t} + n_{t+1}}{n_{1:t}} \\ &\left( M_t^{-1} - M_t^{-1} X_{t+1}^T (n_{1:t} I_{n_{t+1}} + X_{t+1} M_t^{-1} X_{t+1}^T)^{-1} X_{t+1} M_t^{-1} \right), \end{aligned} \quad (SS2)$$

Similarly,  $a_{1:t}$ , the estimated vector of moments by time  $t$  is updated as

$$a_{t+1} = \frac{1}{n_{1:t} + n_{t+1}}(n_{1:t}a_t + X_{t+1}^T \mathbf{y}_{t+1}) \quad (SS3),$$

where  $\mathbf{y}_{t+1}$  is the responses obtained at time  $t+1$ . Thus, we can update the estimated kernel by substituting Eq. SS2 and Eq. SS3 into Eq. 8.

The calculation using such a sequential update becomes cheaper than the calculation of the inverse matrix of  $M_{t+1}$  from scratch.

### Extension of Watanabe method using L2-norm regularization

We can introduce regularization into our method to prevent overfitting and or to solve ill-posed problems that arise when the number of coefficients exceed the number of observations. L2 regularization of our analysis is to estimate the coefficients  $\mathbf{g}$  that minimize the following equation:

$$\|\mathbf{a} - \frac{2}{\sqrt{\pi}} M \cdot \mathbf{g}\|^2 + \lambda \|\mathbf{g}\|^2 \quad (SSS1),$$

where  $\lambda$  is a hyperparameter adjusting the balance between the least squared error from Eq. 8 and the penalty of the size of coefficients to be optimized. By solving the partial derivative of Eq. SSS1 with respect to  $\mathbf{g}$ , regularized coefficients  $\hat{\mathbf{g}}$  are given by

$$\hat{\mathbf{g}} = \frac{\sqrt{\pi}}{2} (M \cdot M + \lambda I)^{-1} \cdot M \cdot \mathbf{a} \quad (SSS2).$$