Kernel bandwidth optimization in spike rate estimation

Hideaki Shimazaki · Shigeru Shinomoto

Abstract Kernel smoother and a time-histogram are
classical tools for estimating an instantaneous rate of
spike occurrences. We recently established a method
for selecting the bin width of the time-histogram, based
on the principle of minimizing the mean integrated
square error (MISE) between the estimated rate and
unknown underlying rate. Here we apply the same
optimization principle to the kernel density estimation
in selecting the width or "bandwidth" of the kernel,
and further extend the algorithm to allow a variable
bandwidth, in conformity with data. The variable ker-
nel has the potential to accurately grasp non-stationary
phenomena, such as abrupt changes in the firing rate,
which we often encounter in neuroscience. In order to
avoid possible overfitting that may take place due to
excessive freedom, we introduced a stiffness constant
for bandwidth variability. Our method automatically
adjusts the stiffness constant, thereby adapting to the
entire set of spike data. It is revealed that the classical
kernel smoother may exhibit goodness-of-fit compara-
ble to, or even better than, that of modern sophisticated
rate estimation methods, provided that the bandwidth
is selected properly for a given set of spike data, accord-
ing to the optimization methods presented here.

H. Shimazaki (✉)
Grün Unit, RIKEN Brain Science Institute,
Saitama 351-0198, Japan
e-mail: shimazaki@brain.riken.jp

S. Shinomoto
Department of Physics, Kyoto University,
Kyoto 606-8502, Japan
e-mail: shinomoto@scphys.kyoto-u.ac.jp

1 Introduction

Neurophysiologists often investigate responses of a
single neuron to a stimulus presented to an animal
by using the discharge rate of action potentials, or
spikes (Adrian 1928; Gerstein and Kiang 1960; Abeles
1982). One classical method for estimating spike rate is
the kernel density estimation (Parzen 1962; Rosenblatt
1956; Sanderson 1980; Richmond et al. 1990; Nawrot
et al. 1999). In this method, a spike sequence is convo-
luted with a kernel function, such as a Gauss density
function, to obtain a smooth estimate of the firing
rate. The estimated rate is sometimes referred to as a
spike density function. This nonparametric method is
left with a free parameter for kernel bandwidth that
determines the goodness-of-fit of the density estimate
to the unknown rate underlying data. Although theo-
ries have been suggested for selecting the bandwidth,
cross-validating with the data (Rudemo 1982; Bowman
1984; Silverman 1986; Scott and Terrell 1987; Scott
1992; Jones et al. 1996; Loader 1999a, b), individual
researchers have mostly chosen bandwidth arbitrarily.
This is partly because the theories have not spread
to the neurophysiological society, and partly due to
inappropriate basic assumptions of the theories them-
selves. Most optimization methods assume a stationary
rate fluctuation, while the neuronal firing rate often
exhibits abrupt changes, to which neurophysiologists, in
particular, pay attention. A fixed bandwidth, optimized

using a stationary assumption, is too wide to extract the details of sharp activation, while in the silent period, the fixed bandwidth would be too narrow and may cause spurious undulation in the estimated rate. It is therefore desirable to allow a variable bandwidth, in conformity with data.

The idea of optimizing bandwidth at every instant was proposed by Loftsgaarden and Quesenberry (1965). However, in contrast to the progress in methods that vary bandwidths at sample points only (Abramson 1982; Breiman et al. 1977; Sain and Scott 1996; Sain 2002; Brewer 2004), the local optimization of bandwidth at every instant turned out to be difficult because of its excessive freedom (Scott 1992; Devroye and Lugosi 2000; Sain and Scott 2002). In earlier studies, Hall & Schucany used the cross-validation of Rudemo and Bowman, within local intervals (Hall and Schucany 1989), yet the interval length was left free. Fan et al. applied cross-validation to locally optimized bandwidth (Fan et al. 1996), and yet the smoothness of the variable bandwidth was chosen manually.

In this study, we first revisit the fixed kernel method, and derive a simple formula to select the bandwidth of the kernel density estimation, similar to the previous method for selecting the bin width of a peristimulus time histogram (See Shimazaki and Shinomoto 2007). Next, we introduce the variable bandwidth into the kernel method and derive an algorithm for determining the bandwidth locally in time. The method automatically adjusts the flexibility, or the stiffness, of the variable bandwidth. The performance of our fixed and variable kernel methods are compared with established density estimation methods, in terms of the goodness-of-fit to underlying rates that vary either continuously or discontinuously. We also apply our kernel methods to the biological data, and examine their ability by cross-validating with data.

Though our methods are based on the classical kernel method, their performances are comparable to various sophisticated rate estimation methods. Because of the classics, they are rather convenient for users: the methods simply suggest bandwidth for the standard kernel density estimation.

2 Methods

2.1 Kernel smoothing

In neurophysiological experiments, neuronal response is examined by repeatedly applying identical stimuli.

The recorded spike trains are aligned at the onset of stimuli, and superimposed to form a *raw density*, as

$$x_t = \frac{1}{n} \sum_{i=1}^{N} \delta(t - t_i), \tag{1}$$

where n is the number of repeated trials. Here, each spike is regarded as a point event that occurs at an instant of time t_i ($i = 1, 2, \cdots, N$) and is represented by the Dirac delta function $\delta(t)$. The kernel density estimate is obtained by convoluting a kernel $k(s)$ to the raw density x_t,

$$\hat{\lambda}_t = \int x_{t-s} k(s) \, ds. \tag{2}$$

Throughout this study, the integral $\int$ that does not specify bounds refers to $\int_{-\infty}^{\infty}$. The kernel function satisfies the normalization condition, $\int k(s) \, ds = 1$, a zero first moment, $\int s k(s) \, ds = 0$, and has a finite bandwidth, $w^2 = \int s^2 k(s) \, ds < \infty$. A frequently used kernel is the Gauss density function,

$$k_w(s) = \frac{1}{\sqrt{2\pi} w} \exp\left(-\frac{s^2}{2w^2}\right), \tag{3}$$

where the bandwidth w is specified as a subscript. In the body of this study, we develop optimization methods that apply generally to any kernel function, and derive a specific algorithm for the Gauss density function in the Appendix.

2.2 Mean integrated squared error optimization principle

Assuming that spikes are sampled from a stochastic process, we consider optimizing the estimate $\hat{\lambda}_t$ to be closest to the unknown underlying rate λ_t. Among several plausible optimizing principles, such as the Kullback-Leibler divergence or the Hellinger distance, we adopt, here, the mean integrated squared error (MISE) for measuring the goodness-of-fit of an estimate to the unknown underlying rate, as

$$\text{MISE} = \int_a^b E(\hat{\lambda}_t - \lambda_t)^2 \, dt, \tag{4}$$

where E refers to the expectation with respect to the spike generation process under a given inhomogeneous rate λ_t. It follows, by definition, that $E x_t = \lambda_t$.

In deriving optimization methods, we assume the Poisson nature, so that spikes are randomly sampled at a given rate λ_t. Spikes recorded from a single neuron correlate in each sequence (Shinomoto et al. 2003, 2005, 2009). In the limit of a large number of spike trains,

Algorithm 1 A method for selecting a fixed kernel bandwidth

i	Superimpose n spike sequences. Obtain a series of spike times $\{t_i\}_{i=1}^N$, where N is the total number of spikes.
ii	Compute the cost function for the bandwidth w of the kernel function $k_w(t)$: $$C_n(w) = \frac{1}{n^2} \sum_{i,j} \psi_w(t_i, t_j) - \frac{2}{n^2} \sum_{i \neq j} k_w(t_i - t_j),$$ where $\psi_w(t_i, t_j) = \int_a^b k_w(t - t_i) k_w(t - t_j)\, dt$†. Here $[a, b]$ is an interval of interest.
iii	Find w^* that minimizes $C_n(w)$.

† See the Appendix for a formula of the Gauss kernel function

however, mixed spikes are statistically independent and the superimposed sequence can be approximated as a single inhomogeneous Poisson point process (Cox 1962; Snyder 1975; Daley and Vere-Jones 1988; Kass et al. 2005).

2.3 Selection of the fixed bandwidth

Given a kernel function such as Eq. (3), the density function Eq. (2) is uniquely determined for a raw density Eq. (1) of spikes obtained from an experiment. A bandwidth w of the kernel may alter the density estimate, and it can accordingly affect the goodness-of-fit of the density function $\hat{\lambda}_t$ to the unknown underlying rate λ_t. In this subsection, we consider applying a kernel of a fixed bandwidth w, and develop a method for selecting w that minimizes the MISE, Eq. (4).

The integrand of the MISE is decomposed into three parts: $E\hat{\lambda}_t^2 - 2\lambda_t E\hat{\lambda}_t + \lambda_t^2$. Since the last component does not depend on the choice of a kernel, we subtract it from the MISE, then define a cost function as a function of the bandwidth w:

$$C_n(w) = \text{MISE} - \int_a^b \lambda_t^2\, dt$$
$$= \int_a^b E\hat{\lambda}_t^2\, dt - 2 \int_a^b \lambda_t E\hat{\lambda}_t\, dt. \qquad (5)$$

Rudemo and Bowman suggested the leave-one-out cross-validation to estimate the second term of Eq. (5) (Rudemo 1982; Bowman 1984). Here, we directly estimate the second term with the Poisson assumption (See also Shimazaki and Shinomoto 2007).

By noting that $\lambda_t = Ex_t$, the integrand of the second term in Eq. (5) is given as

$$Ex_t E\hat{\lambda}_t = E[x_t \hat{\lambda}_t] - E[(x_t - Ex_t)(\hat{\lambda}_t - E\hat{\lambda}_t)], \qquad (6)$$

from a general decomposition of covariance of two random variables. Using Eq. (2), the covariance (the second term of Eq. (6)) is obtained as

$$E[(x_t - Ex_t)(\hat{\lambda}_t - E\hat{\lambda}_t)]$$
$$= \int k_w(t - s)\, E[(x_t - Ex_t)(x_s - Ex_s)]\, ds$$
$$= \int k_w(t - s) \left[\delta(t - s) \frac{1}{n} Ex_s \right] ds$$
$$= \frac{1}{n} k_w(0)\, Ex_t. \qquad (7)$$

Here, to obtain the second equality, we used the assumption of the Poisson point process (independent spikes).

Using Eqs. (6) and (7), Eq. (5) becomes

$$C_n(w) = \int_a^b E\hat{\lambda}_t^2\, dt$$
$$- 2 \int_a^b \left\{ E[x_t \hat{\lambda}_t] - \frac{1}{n} k_w(0)\, Ex_t \right\} dt. \qquad (8)$$

Equation (8) is composed of observable variables only. Hence, from sample sequences, the cost function is estimated as

$$\hat{C}_n(w) = \int_a^b \hat{\lambda}_t^2\, dt - 2 \int_a^b \left\{ x_t \hat{\lambda}_t - \frac{1}{n} k_w(0) x_t \right\} dt. \qquad (9)$$

In terms of a kernel function, the cost function is written as

$$\hat{C}_n(w) = \frac{1}{n^2} \sum_{i,j} \psi_w(t_i, t_j)$$
$$- \frac{2}{n^2} \left\{ \sum_{i,j} k_w(t_i - t_j) - k_w(0) N \right\}$$
$$= \frac{1}{n^2} \sum_{i,j} \psi_w(t_i, t_j) - \frac{2}{n^2} \sum_{i \neq j} k_w(t_i - t_j), \qquad (10)$$

where

$$\psi_w(t_i, t_j) = \int_a^b k_w(t - t_i) k_w(t - t_j)\, dt. \qquad (11)$$

The minimizer of the cost function, Eq. (10), is an estimate of the optimal bandwidth, which is denoted by w^*. The method for selecting a fixed kernel bandwidth is summarized in Algorithm 1. A particular algorithm

Algorithm 2 A method for selecting a variable kernel bandwidth

I		Superimpose n spike sequences. Obtain a series of spike times $\{t_i\}_{i=1}^N$, where N is the total number of spikes.

II i Compute the local cost function at time t:

$$\hat{C}_n^t(w, W) = \frac{1}{n^2} \sum_{i,j} \psi_{w,W}^t(t_i, t_j)$$

$$-\frac{2}{n^2} \sum_{i \neq j} k_w(t_i - t_j)\, \rho_W^{t_i-t},$$

where $\psi_{w,W}^t(t_i, t_j) = \int k_w(u - t_i)$
$\times k_w(u - t_j)\, \rho_W^{u-t} du$†.
Find w^* that minimizes $\hat{C}_n^t(w, W)$ while changing w.

  ii Repeat i while changing W.
Find W^* that satisfies $W = \gamma^{-1} w^*$. Let $W_t^\gamma = W^*$ and $\bar{w}_t^\gamma = \gamma W^*$.

  iii Repeat i-ii while changing t to obtain W_t^γ and $\bar{w}_t^\gamma$.
Smooth the fixed bandwidth as

$$w_t^\gamma = \int \rho_{W_s^\gamma}^{t-s} \bar{w}_s^\gamma\, ds \Big/ \int \rho_{W_s^\gamma}^{t-s}\, ds.$$

III Compute the cost function for the variable bandwidth w_t^γ:

$$\hat{C}_n(\gamma) = \int_a^b \hat{\lambda}_t^2\, dt - \frac{2}{n^2} \sum_{i \neq j} k_{w_{t_i}^\gamma}(t_i - t_j),$$

where $\hat{\lambda}_t = \sum_i k_{w_{t_i}^\gamma}(t - t_i)$.

IV Repeat II-III to find γ^* that minimizes $\hat{C}_n(\gamma)$. Apply II with γ^* to obtain the variable bandwidth $w_t^{\gamma^*}$.

† See the Appendix for a formula of the Gauss function

developed for the Gauss density function is given in the Appendix.

2.4 Selection of the variable bandwidth

The method described in Section 2.3 aims to select a single bandwidth that optimizes the goodness-of-fit of the rate estimate for an entire observation interval $[a, b]$. For a non-stationary case, in which the degree of rate fluctuation greatly varies in time, the rate estimation may be improved by using a kernel function whose bandwidth is adaptively selected in conformity

with data. The spike rate estimated with the variable bandwidth w_t is given by

$$\hat{\lambda}_t = \int x_{t-s} k_{w_t}(s)\, ds. \tag{12}$$

Here we select the variable bandwidth w_t as a fixed bandwidth optimized in a local interval. In this approach, the interval length for the local optimization regulates the shape of the function w_t, therefore, it subsequently determines the goodness-of-fit of the estimated rate to the underlying rate. We provide a method for obtaining the variable bandwidth w_t that minimizes the MISE by optimizing the local interval length.

To select an interval length for local optimization, we introduce the local MISE criterion at time t as

$$local\text{MISE} = \int E\left(\hat{\lambda}_u - \lambda_u\right)^2 \rho_W^{u-t} du, \tag{13}$$

where $\hat{\lambda}_u = \int x_{u-s} k_w(s)\, ds$ is an estimated rate with a fixed bandwidth w. Here, a weight function ρ_W^{u-t} localizes the integration of the squared error in a characteristic interval W centered at time t. An example of the weight function is once again the Gauss density function. See the Appendix for the specific algorithm for the Gauss weight function. As in Eq. (5), we introduce the local cost function at time t by subtracting the term irrelevant for the choice of w as

$$C_n^t(w, W) = local\text{MISE} - \int \lambda_u^2 \rho_W^{u-t} du. \tag{14}$$

The optimal fixed bandwidth w^* is obtained as a minimizer of the estimated cost function:

$$\hat{C}_n^t(w, W) = \frac{1}{n^2} \sum_{i,j} \psi_{w,W}^t(t_i, t_j)$$

$$-\frac{2}{n^2} \sum_{i \neq j} k_w(t_i - t_j)\, \rho_W^{t_i-t}, \tag{15}$$

where

$$\psi_{w,W}^t(t_i, t_j) = \int k_w(u - t_i)\, k_w(u - t_j)\, \rho_W^{u-t} du. \tag{16}$$

The derivation follows the same steps as in the previous section. Depending on the interval length W, the optimal bandwidth w^* varies. We suggest selecting an interval length that scales with the optimal bandwidth as $\gamma^{-1} w^*$. The parameter γ regulates the interval length for local optimization: With small $\gamma (\ll 1)$, the fixed bandwidth is optimized within a long interval; With large $\gamma (\sim 1)$, the fixed bandwidth is optimized within a short interval. The interval length and fixed bandwidth, selected at time t, are denoted as W_t^γ and $\bar{w}_t^\gamma$.

The locally optimized bandwidth $\bar{w}_t^\gamma$ is repeatedly obtained for different $t(\in [a, b])$. Because the intervals

overlap, we adopt the Nadaraya-Watson kernel regression (Nadaraya 1964; Watson 1964) of $\bar{w}_t^\gamma$ as a local bandwidth at time t:

$$w_t^\gamma = \int \rho_{W_s^\gamma}^{t-s} \bar{w}_s^\gamma \, ds \Big/ \int \rho_{W_s^\gamma}^{t-s} \, ds. \qquad (17)$$

The variable bandwidth w_t^γ obtained from the same data, but with different γ, exhibits different degrees of smoothness: With small $\gamma (\ll 1)$, the variable bandwidth fluctuates slightly; With large $\gamma (\sim 1)$, the variable bandwidth fluctuates significantly. The parameter γ is thus a smoothing parameter for the variable bandwidth. Similar to the fixed bandwidth, the goodness-of-fit of the variable bandwidth can be estimated from the data. The cost function for the variable bandwidth selected with γ is obtained as

$$\hat{C}_n(\gamma) = \int_a^b \hat{\lambda}_t^2 \, dt - \frac{2}{n^2} \sum_{i \neq j} k_{w_{t_i}^\gamma} (t_i - t_j), \qquad (18)$$

where $\hat{\lambda}_t = \int x_{t-s} k_{w_t^\gamma}(s) \, ds$ is an estimated rate, with the variable bandwidth w_t^γ. The integral is calculated numerically. With the stiffness constant γ^* that minimizes Eq. (18), local optimization is performed in an ideal

interval length. The method for optimizing the variable kernel bandwidth is summarized in Algorithm 2.

3 Results

3.1 Comparison of the fixed and variable kernel methods

By using spikes sampled from an inhomogeneous Poisson point process, we examined the efficiency of the kernel methods in estimating the underlying rate. We also used a sequence obtained by superimposing ten non-Poissonian (gamma) sequences (Shimokawa and Shinomoto 2009), but there was practically no significant difference in the rate estimation from the Poissonian sequence.

Figure 1 displays the result of the fixed kernel method based on the Gauss density function. The kernel bandwidth selected by Algorithm 1 applies a reasonable filtering to the set of spike sequences. Figure 1(d) shows that a cost function, Eq. (10), estimated from the spike data is similar to the original MISE, Eq. (4), which was computed using the knowledge of the underlying

Fig. 1 Fixed kernel density estimation. (**a**) The underlying spike rate λ_t of the Poisson point process. (**b**) 20 spike sequences sampled from the underlying rate, and (**c**) Kernel rate estimates made with three types of bandwidth: too small, optimal, and too large. The *gray area* indicates the underlying spike rate. (**d**) The cost function for bandwidth w. *Solid line* is the estimated cost function, Eq. (10), computed from the spike data; The *dashed line* is the exact cost function, Eq. (5), directly computed by using the known underlying rate
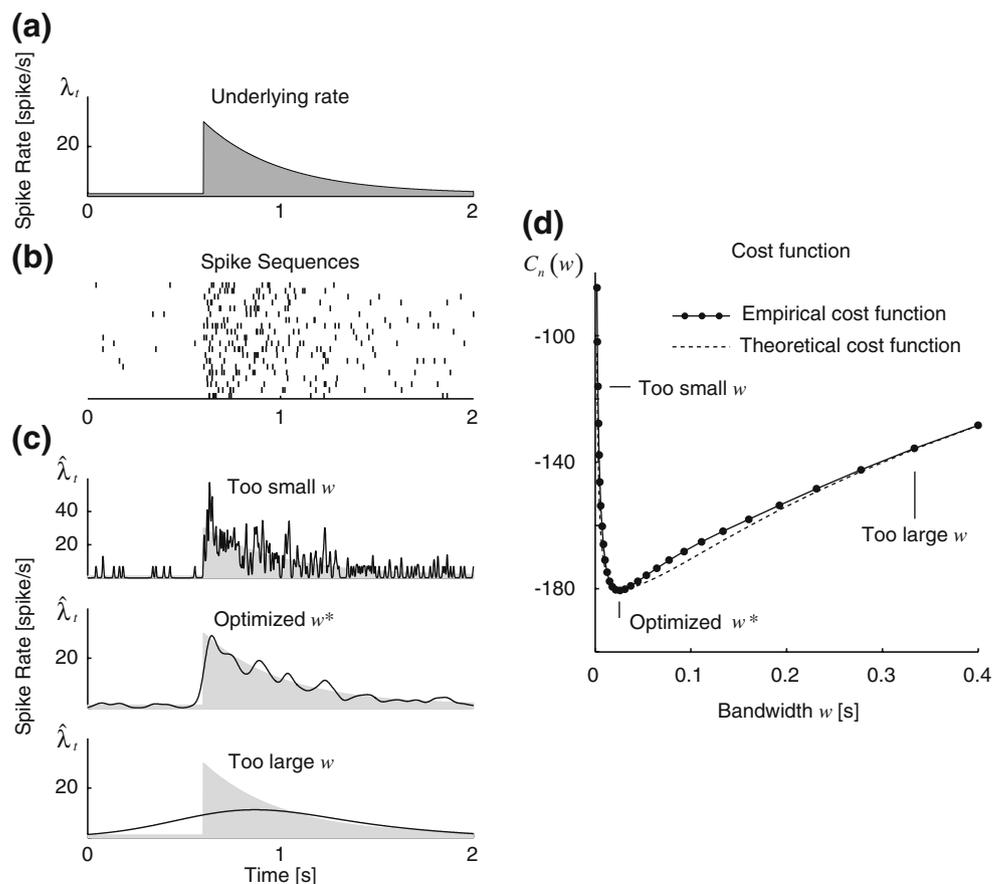
rate. This demonstrates that MISE optimization can, in practice, be carried out by our method, even without knowing the underlying rate.

Figure 2(a) demonstrates how the rate estimation is altered by replacing the fixed kernel method with the variable kernel method (Algorithm 2), for identical spike data (Fig. 1(b)). The Gauss weight function is used to obtain a smooth variable bandwidth. The manner in which the optimized bandwidth varies in the time axis is shown in Fig. 2(b): the bandwidth is short in a moment of sharp activation, and is long in the period of smooth rate modulation. Eventually, the sharp activation is grasped minutely and slow modulation is expressed without spurious fluctuations. The stiffness constant γ for the bandwidth variation is selected by minimizing the cost function, as shown in Fig. 2(c).

3.2 Comparison with established density estimation methods

We wish to examine the fitting performance of the fixed and variable kernel methods in comparison with established density estimation methods, by paying attention to their aptitudes for either continuous or dis-

continuous rate processes. Figure 3(a) shows the results for sinusoidal and sawtooth rate processes, as samples of continuous and discontinuous processes, respectively. We also examined triangular and rectangular rate processes as different samples of continuous and discontinuous processes, but the results were similar. The goodness-of-fit of the density estimate to the underlying rate is evaluated in terms of integrated squared error (ISE) between them.

The established density estimation methods examined for comparison are the histogram (Shimazaki and Shinomoto 2007), Abramson's adaptive kernel (Abramson 1982), Locfit (Loader 1999b), and Bayesian adaptive regression splines (BARS) (DiMatteo et al. 2001; Kass et al. 2003) methods, whose details are summarized below.

A histogram method, which is often called a peristimulus time histogram (PSTH) in neurophysiological literature, is the most basic method for estimating the spike rate. To optimize the histogram, we used a method proposed for selecting the bin width based on the MISE principle (Shimazaki and Shinomoto 2007).

Abramson's adaptive kernel method (Abramson 1982) uses the sample point kernel estimate $\hat{\lambda}_t = \sum_i k_{w_{t_i}}(t - t_i)$, in which the bandwidths are adapted
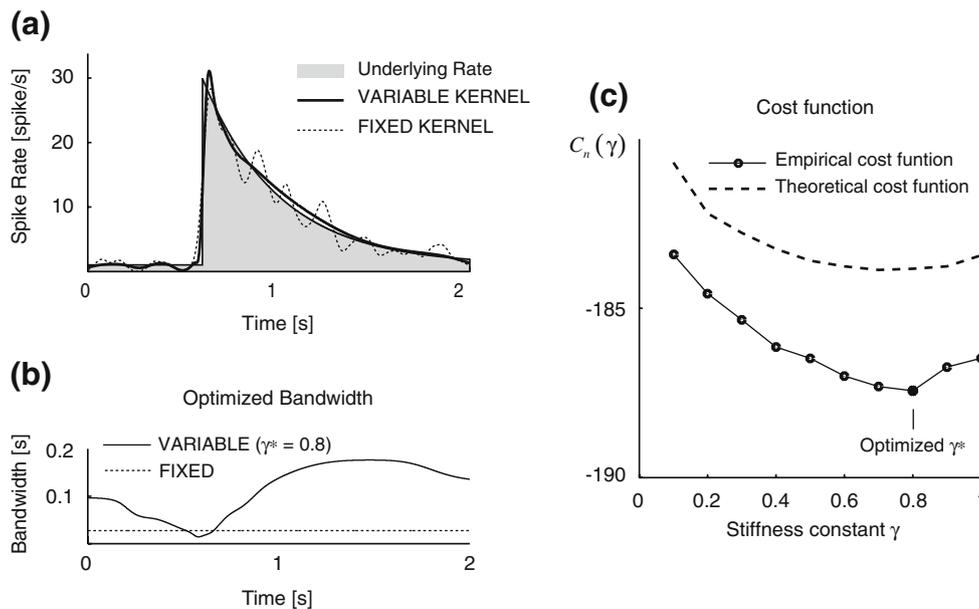
Fig. 2 Variable kernel density estimation. (**a**) Kernel rate estimates. The *solid* and *dashed lines* are rate estimates made by the variable and fixed kernel methods for the spike data of Fig. 1(b). The *gray area* is the underlying rate. (**b**) Optimized bandwidths. The *solid line* is the variable bandwidth determined with the optimized stiffness constant $\gamma* = 0.8$, selected by Algorithm 2;

the *dashed line* is the fixed bandwidth selected by Algorithm 1. (**b**) The cost function for bandwidth stiffness constant. The *solid line* is the cost function for the bandwidth stiffness constant γ, Eq. (18), estimated from the spike data; the *dashed line* is the cost function computed from the known underlying rate
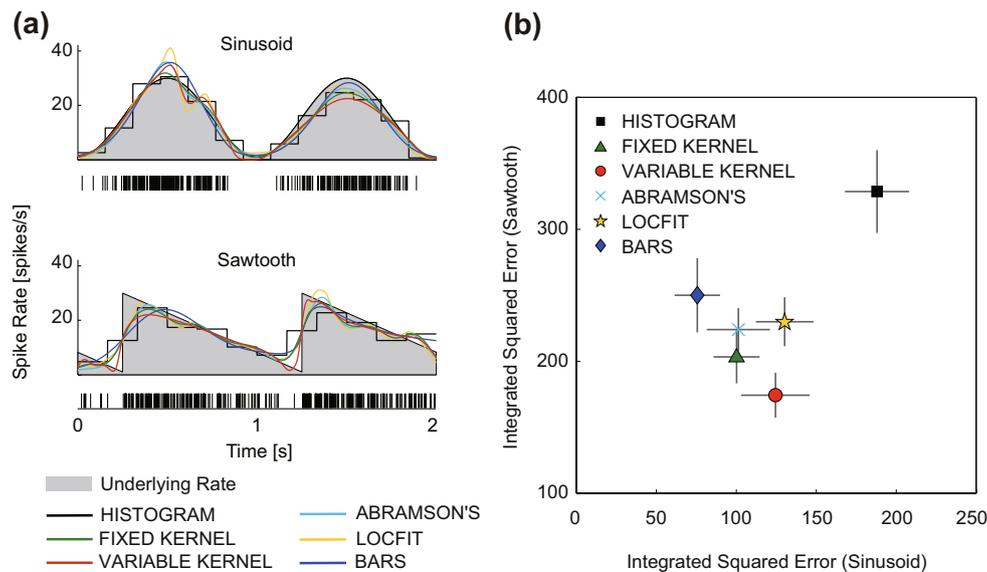
Fig. 3 Fitting performances of the six rate estimation methods, histogram, fixed kernel, variable kernel, Abramson's adaptive kernel, Locfit, and Bayesian adaptive regression splines (BARS). (**a**) Two rate profiles (2 [s]) used in generating spikes (*gray area*), and the estimated rates using six different methods. The raster plot in each panel is sample spike data ($n = 10$, superimposed). (**b**) Comparison of the six rate estimation methods in their goodness-of-fit, based on the integrated squared error (ISE) between the underlying and estimated rate. The abscissa and the ordinate are the ISEs of each method applied to sinusoidal and sawtooth underlying rates (10 [s]). The mean and standard deviation of performance evaluated using 20 data sets are plotted for each method

at the sample points. Scaling the bandwidths as $w_{t_i} = w (g/\hat{\lambda}_{t_i})^{1/2}$ was suggested, where w is a pilot bandwidth, $g = (\prod_i \hat{\lambda}_{t_i})^{1/N}$, and $\hat{\lambda}_t$ is a fixed kernel estimate with w. Abramson's method is a two-stage method, in which the pilot bandwidth needs to be selected beforehand. Here, the pilot bandwidth is selected using the fixed kernel optimization method developed in this study.

The Locfit algorithm developed by Loader (1999b) fits a polynomial to a log-density function under the principle of maximizing a locally defined likelihood. We examined the automatic choice of the adaptive bandwidth of the local likelihood, and found that the default fixed method yielded a significantly better fit. We used a nearest neighbor based bandwidth method, with a parameter covering 20% of the data.

The BARS (DiMatteo et al. 2001; Kass et al. 2003) is a spline-based adaptive regression method on an exponential family response model, including a Poisson count distribution. The rate estimated with the BARS is the expected splines computed from the posterior distribution on the knot number and locations with a Markov chain Monte Carlo method. The BARS is, thus, capable of smoothing a noisy histogram without missing abrupt changes. To create an initial histogram, we used 4 [ms] bin width, which is small enough to examine rapid changes in the firing rate.

Figure 3(a) displays the density profiles of the six different methods estimated from an identical set of spike trains ($n = 10$) that are numerically sampled from a sinusoidal or sawtooth underlying rate (2 [s]). Figure 3(b) summarizes the goodness-of-fit of the six methods to the sinusoidal and sawtooth rates (10 [s]) by averaging over 20 realizations of samples.

For the sinusoidal rate function, representing continuously varying rate processes, the BARS is most efficient in terms of ISE performance. For the sawtooth rate function, representing discontinuous non-stationary rate processes, the variable kernel estimation developed here is the most efficient in grasping abrupt rate changes. The histogram method is always inferior to the other five methods in terms of ISE performance, due to the jagged nature of the piecewise constant function.

3.3 Application to experimental data

We examine, here, the fixed and variable kernel methods in their applicability to real biological data. In particular, the kernel methods are applied to the spike data of an MT neuron responding to a random dot stimulus (Britten et al. 2004). The rates estimated from $n = 1$, 10, and 30 experimental trials are shown in

Fig. 4. Fine details of rate modulation are revealed as we increase the sample size (Bair and Koch 1996). The fixed kernel method tends to choose narrower bandwidths, while the variable kernel method tends to choose wider bandwidths in the periods in which spikes are not abundant.

The performance of the rate estimation methods is cross-validated. The bandwidth, denoted as w_t for both fixed and variable, is obtained with a training data set of n trials. The error is evaluated by computing the cost function, Eq. (18), in a cross-validatory manner:

$$\hat{C}_n\left(w_t\right) = \int_a^b \hat{\lambda}_t^2 \, dt - \frac{2}{n^2} \sum_{i \neq j} k_{w_{t'_i}}\left(t'_i - t'_j\right), \tag{19}$$
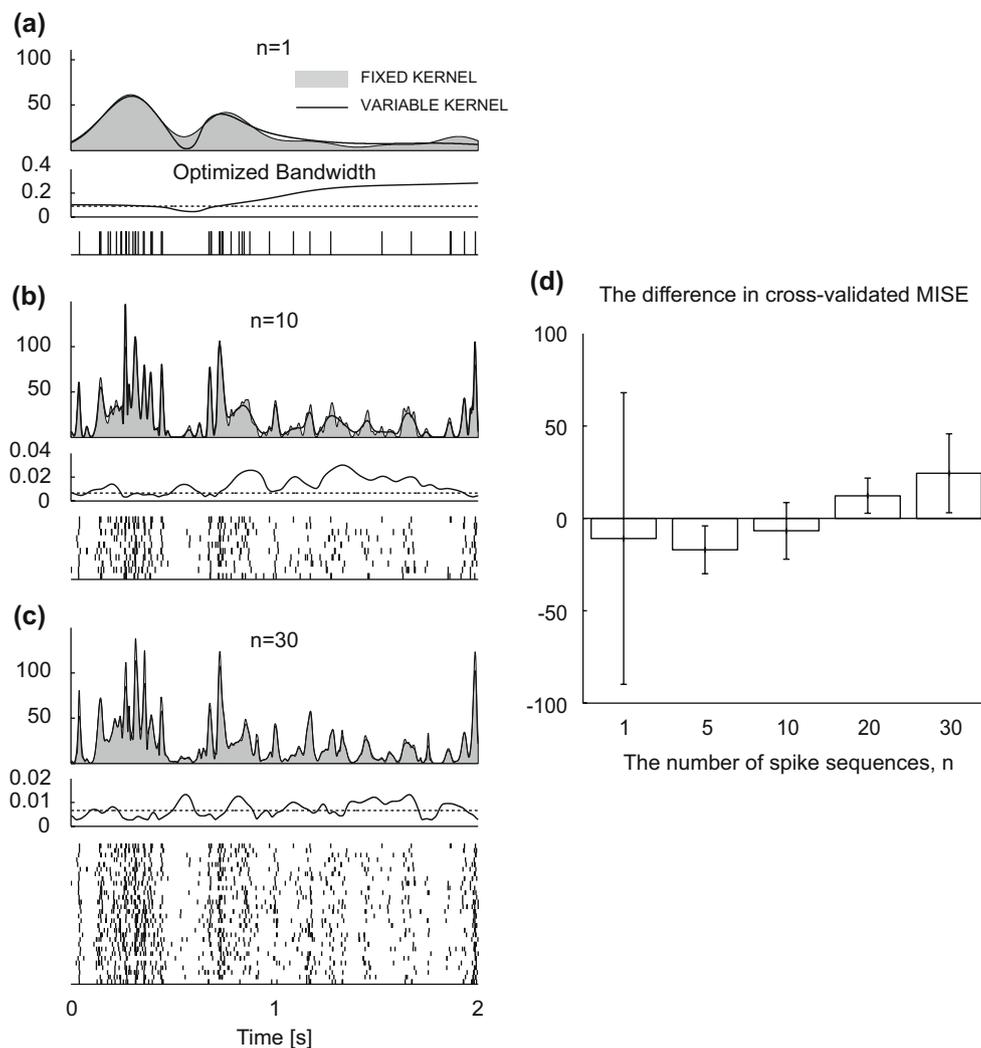
Fig. 4 Application of the fixed and variable kernel methods to spike data of an MT neuron (j024 with coherence 51.2% in nsa2004.1 (Britten et al. 2004)). (**a**–**c**): Analyses of $n = 1$, 10, and 30 spike trains; (*top*) Spike rates [spikes/s] estimated with the fixed and variable kernel methods are represented by the *gray area* and *solid line*, respectively; (*middle*) optimized fixed and variable bandwidths [s] are represented by *dashed* and *solid lines*, respectively; (*bottom*) A raster plot of the spike data used in the estimation. (**d**) Comparison of the two kernel methods. *Bars* represent the difference between the cross-validated cost functions, Eq. (19), of the fixed and variable kernel methods (fix less variable). The positive difference indicates superior fitting performance of the variable kernel method. The cross-validated cost function is obtained as follows. Whole spike sequences ($n_{\text{total}} = 60$) are divided into n_{total}/n blocks, each composed of $n(= 1, 5, 20, 30)$ spike sequences. A bandwidth was selected using spike data of a training block. The cross-validated cost functions, Eq. (19), for the selected bandwidth are computed using the $n_{\text{total}}/n - 1$ leftover test blocks, and their average is computed. The cost function is repeatedly obtained n_{total}/n-times by changing the training block. The mean and standard deviation, computed from n_{total}/n samples, are displayed

where the test spike times $\{t'_i\}$ are obtained from n spike sequences in the leftovers, and $\hat{\lambda}_t = \frac{1}{n} \sum_i k_{w_t}(t - t'_i)$. Figure 4(d) shows the performance improvements by the variable bandwidth over the fixed bandwidth, as evaluated by Eq. (19). The fixed and variable kernel methods perform better for smaller and larger sizes of data, respectively. In addition, we compared the fixed kernel method and the BARS by cross-validating the log-likelihood of a Poisson process with the rate estimated using the two methods. The difference in the log-likelihoods was not statistically significant for small samples ($n = 1, 5$ and 10), while the fixed kernel method fitted better to the spike data with larger samples ($n = 20$ and 30).

4 Discussion

In this study, we developed methods for selecting the kernel bandwidth in the spike rate estimation based on the MISE minimization principle. In addition to the principle of optimizing a fixed bandwidth, we further considered selecting the bandwidth locally in time, assuming a non-stationary rate modulation.

We tested the efficiency of our methods using spike sequences numerically sampled from a given rate (Figs. 1 and 2). Various density estimators constructed on different optimization principles were compared in their goodness-of-fit to the underlying rate (Fig. 3). There is in fact no oracle that selects one among various optimization principles, such as MISE minimization or likelihood maximization. Practically, reasonable principles render similar detectability for rate modulation; the kernel methods based on MISE were roughly comparable to the Locfit based on likelihood maximization in their performances. The difference of the performances is not due to the choice of principles, but rather due to techniques; kernel and histogram methods lead to completely different results under the same MISE minimization principle (Fig. 3(b)). Among the smooth rate estimators, the BARS was good at representing continuously varying rate, while the variable kernel method was good at grasping abrupt changes in the rate process (Fig. 3(b)).

We also examined the performance of our methods in application to neuronal spike sequences by cross-validating with the data (Fig. 4). The result demonstrated that the fixed kernel method performed well in small samples. We refer to Cunningham et al. (2008) for a result on the superior fitting performance of a fixed kernel to small samples in comparison with the Locfit and BARS, as well as the Gaussian process smoother (Cunningham et al. 2008; Smith and Brown 2003; Koyama and Shinomoto 2005). The adaptive methods, however, have the potential to outperform the fixed method with larger samples derived from a non-stationary rate profile (See also Endres et al. 2008 for comparisons of their adaptive histogram with the fixed histogram and kernel method). The result in Fig. 4 confirmed the utility of our variable kernel method for larger samples of neuronal spikes.

We derived the optimization methods under the Poisson assumption, so that spikes are randomly drawn from a given rate. If one wishes to estimate spike rate of a single or a few sequences that contain strongly correlated spikes, it is desirable to utilize the information as to non-Poisson nature of a spike train (Cunningham et al. 2008). Note that a non-Poisson spike train may be dually interpreted, as being derived either irregularly from a constant rate, or regularly from a fluctuating rate (Koyama and Shinomoto 2005; Shinomoto and Koyama 2007). However, a sequence obtained by superimposing many spike trains is approximated as a Poisson process (Cox 1962; Snyder 1975; Daley and Vere-Jones 1988; Kass et al. 2005), for which dual interpretation does not occur. Thus the kernel methods developed in this paper are valid for the superimposed sequence, and serve as the peristimulus density estimator for spike trains aligned at the onset or offset of the stimulus.

Kernel smoother is a classical method for estimating the firing rate, as popular as the histogram method. We have shown in this paper that the classical kernel methods perform well in the goodness-of-fit to the underlying rate. They are not only superior to the histogram method, but also comparable to modern sophisticated methods, such as the Locfit and BARS. In particular, the variable kernel method outperformed competing methods in representing abrupt changes in the spike rate, which we often encounter in neuroscience. Given simplicity and familiarity, the kernel smoother can still be the most useful in analyzing the spike data, provided that the bandwidth is chosen appropriately as instructed in this paper.

Appendix: Cost functions of the Gauss kernel function

In this appendix, we derive definite MISE optimization algorithms we developed in the body of the paper with the particular Gauss density function, Eq. (3).

A.1 A cost function for a fixed bandwidth

The estimated cost function is obtained, as in Eq. (10):

$$n^2 \hat{C}_n(w) = \sum_{i,j} \psi_w(t_i, t_j) - 2 \sum_{i \neq j} k_w(t_i - t_j),$$

where, from Eq. (11),

$$\psi_w(t_i, t_j) = \int_a^b k_w(t-t_i) k_w(t-t_j)\, dt.$$

A symmetric kernel function, including the Gauss function, is invariant to exchange of t_i and t_j when computing $k_w(t_i - t_j)$. In addition, the correlation of the kernel function, Eq. (11), is symmetric with respect to t_i and t_j. Hence, we obtain the following relationships

$$\sum_{i \neq j} k_w(t_i - t_j) = 2 \sum_{i < j} k_w(t_i - t_j), \tag{20}$$

$$\sum_{i,j} \psi_w(t_i, t_j) = \sum_i \psi_w(t_i, t_i) + 2 \sum_{i < j} \psi_w(t_i, t_j). \tag{21}$$

By plugging Eqs. (20) and (21) into Eq. (10), the cost function is simplified as .

$$n^2 \hat{C}_n(w) = \sum_i \psi_w(t_i, t_i)$$
$$+ 2 \sum_{i < j} \left\{ \psi_w(t_i, t_j) - 2 k_w(t_i - t_j) \right\}. \tag{22}$$

For the Gauss kernel function, Eq. (3), with bandwidth w, Eq. (11) becomes

$$\psi_w(t_i, t_j) = \frac{1}{\sqrt{\pi} 4w} e^{-\frac{(t_i - t_j)^2}{4w^2}}$$
$$\times \left\{ \mathrm{erf}\left(\frac{2b - t_i - t_j}{2w} \right) - \mathrm{erf}\left(\frac{2a - t_i - t_j}{2w} \right) \right\}, \tag{23}$$

where $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt$. A simplified equation is obtained by evaluating the MISE in an unbounded domain: $a \to -\infty$ and $b \to +\infty$. Using $\mathrm{erf}(\pm\infty) = \pm 1$, we obtain

$$\psi_w(t_i, t_j) \approx \frac{1}{\sqrt{\pi} 2w} e^{-\frac{(t_i - t_j)^2}{4w^2}}. \tag{24}$$

Using Eq. (24) in Eq. (22), we obtain a formula for selecting the bandwidth of the Gauss kernel function, $2\sqrt{\pi} n^2 \hat{C}_n(w)$:

$$\frac{N}{w} + \frac{2}{w} \sum_{i < j} \left\{ e^{-\frac{(t_i - t_j)^2}{4w^2}} - 2\sqrt{2} e^{-\frac{(t_i - t_j)^2}{2w^2}} \right\}. \tag{25}$$

A.2 A local cost function for a variable bandwidth

The local cost function is obtained, as in Eq. (15):

$$n^2 \hat{C}_n^t(w, W) = \sum_{i,j} \psi_{w,W}^t(t_i, t_j) - 2 \sum_{i \neq j} k_w(t_i - t_j) \rho_W^{t_i - t},$$

where, from Eq. (16),

$$\psi_{w,W}^t(t_i, t_j) = \int k_w(u - t_i) k_w(u - t_j) \rho_W^{u-t}\, du.$$

For the summations in the local cost function, Eq. (15), we have equalities:

$$\sum_{i \neq j} k_w(t_i - t_j) \rho_W^{t_i - t} = \sum_{i < j} k_w(t_i - t_j) \left\{ \rho_W^{t_i - t} + \rho_W^{t_j - t} \right\}, \tag{26}$$

$$\sum_{i,j} \psi_{w,W}^t(t_i, t_j) = \sum_i \psi_{w,W}^t(t_i, t_i) + 2 \sum_{i < j} \psi_{w,W}^t(t_i, t_j). \tag{27}$$

The first equation holds for a symmetric kernel. The second equation is derived because Eq. (16) is invariant to an exchange of t_i and t_j. Using Eqs. (26) and (27), Eq. (15) can be computed as

$$n^2 \hat{C}_n^t(w, W) = \sum_i \psi_{w,W}^t(t_i, t_j)$$
$$+ 2 \sum_{i < j} \left[\psi_{w,W}^t(t_i, t_j) - k_w(t_i - t_j) \left\{ \rho_W^{t_i - t} + \rho_W^{t_j - t} \right\} \right]. \tag{28}$$

For the Gauss kernel function and the Gauss weight function with bandwidth w and W respectively, Eq. (16) is calculated as

$$\psi_{w,W}^t\left(t_i, t_j\right) = \frac{1}{2\pi w^2} \frac{1}{\sqrt{2\pi} W}$$

$$\times \int \exp\left[-\frac{(u - t_i)^2 + (u - t_j)^2}{2w^2} - \frac{(u - t)^2}{2W^2}\right] du. \quad (29)$$

By completing the square with respect to u, the exponent in the above equation is written as

$$\frac{w^2 + 2W^2}{2w^2 W^2}\left\{u + \frac{\tau W^2 + (t_i - t) w^2}{w^2 + 2W^2}\right\}^2$$

$$- \frac{\left[(t - t_i)^2 + (t - t_j)^2\right] w^2 + (t_i - t_j)^2 W^2}{2w^2\left(w^2 + 2W^2\right)}. \quad (30)$$

Using the formula $\int e^{-Au^2} du = \sqrt{\frac{\pi}{A}}$, Eq. (16) is obtained as

$$\psi_{w,W}^t\left(t_i, t_j\right) = \frac{1}{2\pi w \sqrt{w^2 + 2W^2}}$$

$$\times \exp\left[-\frac{\left[(t - t_i)^2 + (t - t_j)^2\right] w^2 + (t_i - t_j)^2 W^2}{2w^2\left(w^2 + 2W^2\right)}\right]. \quad (31)$$

References

Abeles, M. (1982). Quantification, smoothing, and confidence-limits for single-units histograms. *Journal of Neuroscience Methods, 5*(4), 317–325.

Abramson, I. (1982). On bandwidth variation in kernel estimates-a square root law. *The Annals of Statistics, 10*(4), 1217–1223.

Adrian, E. (1928). *The basis of sensation: The action of the sense organs*. New York: W.W. Norton.

Bair, W., & Koch, C. (1996). Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Computation, 8*(6), 1185–1202.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika, 71*(2), 353.

Breiman, L., Meisel, W., & Purcell, E. (1977). Variable kernel estimates of multivariate densities. *Technometrics, 19*, 135–144.

Brewer, M. J. (2004). A Bayesian model for local smoothing in kernel density estimation. *Statistics and Computing, 10*, 299–309.

Britten, K. H., Shadlen, M. N., Newsome, W. T., & Movshon, J. A. (2004). Responses of single neurons in macaque mt/v5 as a function of motion coherence in stochastic dot stimuli. The Neural Signal Archive. nsa2004.1. http://www.neuralsignal.org.

Cox, R. D. (1962). *Renewal theory*. London: Wiley.

Cunningham, J., Yu, B., Shenoy, K., Sahani, M., Platt, J., Koller, D., et al. (2008). Inferring neural firing rates from spike trains using Gaussian processes. *Advances in Neural Information Processing Systems, 20*, 329–336.

Daley, D., & Vere-Jones, D. (1988). *An introduction to the theory of point processes*. New York: Springer.

Devroye, L., & Lugosi, G. (2000). Variable kernel estimates: On the impossibility of tuning the parameters. In E. Giné, D. Mason, & J. A. Wellner (Eds.), *High dimensional probability II* (pp. 405–442). Boston: Birkhauser.

DiMatteo, I., Genovese, C. R., & Kass, R. E. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika, 88*(4), 1055–1071.

Endres, D., Oram, M., Schindelin, J., & Foldiak, P. (2008). Bayesian binning beats approximate alternatives: Estimating peristimulus time histograms. *Advances in Neural Information Processing Systems, 20*, 393–400.

Fan, J., Hall, P., Martin, M. A., & Patil, P. (1996). On local smoothing of nonparametric curve estimators. *Journal of the American Statistical Association, 91*, 258–266.

Gerstein, G. L., & Kiang, N. Y. S. (1960). An approach to the quantitative analysis of electrophysiological data from single neurons. *Biophysical Journal, 1*(1), 15–28.

Hall, P., & Schucany, W. R. (1989). A local cross-validation algorithm. *Statistics & Probability Letters, 8*(2), 109–117.

Jones, M., Marron, J., & Sheather, S. (1996). A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association, 91*(433), 401–407.

Kass, R. E., Ventura, V., & Brown, E. N. (2005). Statistical issues in the analysis of neuronal data. *Journal of Neurophysiology, 94*(1), 8–25.

Kass, R. E., Ventura, V., & Cai, C. (2003). Statistical smoothing of neuronal data. *Network-Computation in Neural Systems, 14*(1), 5–15.

Koyama, S., & Shinomoto, S. (2005). Empirical Bayes interpretations of random point events. *Journal of Physics A-Mathematical and General, 38*, 531–537.

Loader, C. (1999a). Bandwidth selection: Classical or plug-in? *The Annals of Statistics, 27*(2), 415–438.

Loader, C. (1999b). *Local regression and likelihood*. New York: Springer.

Loftsgaarden, D. O., & Quesenberry, C. P. (1965). A nonparametric estimate of a multivariate density function. *The Annals of Mathematical Statistics, 36*, 1049–1051.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications, 9*(1), 141–142.

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

Parzen, E. (1962). Estimation of a probability density-function and mode. *The Annals of Mathematical Statistics, 33*(3), 1065.

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

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density-function. *The Annals of Mathematical Statistics, 27*(3), 832–837.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics, 9*(2), 65–78.

Sain, S. R. (2002). Multivariate locally adaptive density estimation. *Computational Statistics & Data Analysis, 39*, 165–186.

Sain, S., & Scott, D. (1996). On locally adaptive density estimation. *Journal of the American Statistical Association, 91*(436), 1525–1534.

Sain, S., & Scott, D. (2002). Zero-bias locally adaptive density estimators. *Scandinavian Journal of Statistics, 29*(3), 441–460.

Sanderson, A. (1980). Adaptive filtering of neuronal spike train data. *IEEE Transactions on Biomedical Engineering, 27*, 271–274.

Scott, D. W. (1992). *Multivariate density estimation: Theory, practice, and visualization*. New York: Wiley-Interscience.

Scott, D. W., & Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association, 82*, 1131–1146.

Shimazaki, H., & Shinomoto, S. (2007). A method for selecting the bin size of a time histogram. *Neural Computation, 19*(6), 1503–1527.

Shimokawa, T., & Shinomoto, S. (2009). Estimating instantaneous irregularity of neuronal firing. *Neural Computation, 21*(7), 1931–1951.

Shinomoto, S., & Koyama, S. (2007). A solution to the controversy between rate and temporal coding. *Statistics in Medicine, 26*, 4032–4038.

Shinomoto, S., Kim, H., Shimokawa, T., Matsuno, N., Funahashi, S., Shima, K., et al. (2009). Relating neuronal firing patterns to functional differentiation of cerebral cortex. *PLoS Computational Biology, 5*, e1000433.

Shinomoto, S., Miyazaki, Y., Tamura, H., & Fujita, I. (2005). Regional and laminar differences in *in vivo* firing patterns of primate cortical neurons. *Journal of Neurophysiology, 94*(1), 567–575.

Shinomoto, S., Shima, K., & Tanji, J. (2003). Differences in spiking patterns among cortical neurons. *Neural Computation, 15*(12), 2823–2842.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman & Hall.

Smith, A. C., & Brown, E. N. (2003). Estimating a state-space model from point process observations. *Neural Computation, 15*(5), 965–991.

Snyder, D. (1975). *Random point processes*. New York: Wiley.

Watson, G. S. (1964). Smooth regression analysis. *Sankhya: The Indian Journal of Statistics, Series A, 26*(4), 359–372.