Journal of Physiology - Paris 104 (2010) 128–136

Contents lists available at ScienceDirect

Journal of Physiology - Paris

journal homepage: www.elsevier.com/locate/jphysparis

Modelling spike trains and extracting response latency with Bayesian binning

Dominik Endres [a,*], Johannes Schindelin [b], Peter Földiák [c], Mike W Oram [c]

[a] Section for Theoretical Sensomotorics, Department of Cognitive Neurology, University Clinic Tübingen and Hertie Institute for Clinical Brain Science and Center for Integrative Neuroscience, Frondsbergstrasse 23, 72070 Tübingen, Germany
[b] Max-Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany
[c] School of Psychology, University of St. Andrews, KY16 9JP, UK

ARTICLE INFO

Keywords:
Spike train analysis
Bayesian methods
Response latency
PSTH
SDF

ABSTRACT

The peristimulus time histogram (PSTH) and the spike density function (SDF) are commonly used in the analysis of neurophysiological data. The PSTH is usually obtained by binning spike trains, the SDF being a (Gaussian) kernel smoothed version of the PSTH. While selection of the bin width or kernel size is often relatively arbitrary there have been recent attempts to remedy this situation (Shimazaki and Shinomoto, 2007c,b,a). We further develop an exact Bayesian generative model approach to estimating PSTHs (Endres et al., 2008) and demonstate its superiority to competing methods using data from early (LGN) and late (STSa) visual areas. We also highlight the advantages of our scheme's automatic complexity control and generation of error bars. Additionally, our approach allows extraction of excitatory and inhibitory response latency from spike trains in a principled way, both on repeated and single trial data. We show that the method can be applied to data with high background firing rates and inhibitory responses (LGN) as well as to data with low firing rate and excitatory responses (STSa). Furthermore, we demonstrate on simulated data that our latency extraction method works for a range of signal-to-noise ratios and background firing rates. While further studies are needed to examine the sensitivity of our method to, for example, gradual changes in firing rate and adaptation, the current results suggest that Bayesian binning is a powerful method for the estimation of firing rate and the extraction response latency from neuronal spike trains.

1. Introduction

Plotting a peristimulus time histogram (PSTH), or a spike density function (SDF), from spiketrains evoked by and aligned to the onset of a stimulus or motor action is often one of the first steps in the analysis of neurophysiological data. The PSTH and SDF provide visualisation of characteristics of the neural response, such as instantaneous firing rates (or firing probabilities), latencies and response offsets. While there have been more principled approaches to the problem of determining the appropriate temporal resolution (Shimazaki and Shinomoto, 2007c,b,a) the PSTH and SDF are frequently constructed in an unsystematic manner (e.g. the choice of time bin size is driven by result expectations – what looks good – as much as by the data) even though they implicitly represent a model of the neuron's response as a function of time.

In Endres et al. (2008), we developed an exact Bayesian, generative model approach to estimating PSTHs. Our model encodes a spike generator described by an inhomogeneous Bernoulli process with piecewise constant (in time) firing probabilities. Relevant marginal distributions, e.g. the posterior distribution of the number of bins, can be evaluated from the full posterior distribution over the model parameters efficiently, i.e. in polynomial time. Furthermore, by extension of previous dynamic programming schemes (Endres and Földiák, 2005) the expected values, such as the predictive firing rate and its standard error, are computable with at most cubic effort.

In the following, after stating the complete model specification, we extend the performance comparisons in Endres et al. (2008) and illustrate the usefulness of our method. Next, we demonstrate how to use this model for principled feature extraction from spike trains. The features which we are interested in are latencies and firing rates, since previous studies (Oram et al., 2002) indicate that much of the stimulus-related information carried by neurons is contained in these relatively coarse measures. We give a 'minimal' definition of latency and show how the latency posterior distribution and the firing rate posterior density can be evaluated for data from visual areas LGN, STS and the motor cortex.

Note that we do in no way claim that a PSTH is a complete generative description of spiking neurons. We are merely concerned with inferring that part of the generative process which can be described by a PSTH in a Bayes-optimal way.

[*] Corresponding author.
E-mail addresses: dominik.endres@klinikum.uni-tuebingen.de, dme2@st-andrews.ac.uk (D. Endres), johannes.schindelin@gmx.de (J. Schindelin), Peter.Foldiak@st-andrews.ac.uk (P. Földiák), mwo@st-andrews.ac.uk (M.W Oram).

0928-4257/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jphysparis.2009.11.015

2. The model

2.1. Traditional approaches

There are, broadly speaking, two traditional approaches to estimating firing probabilities or firing rates from neurophysiological data: binning procedures and smoothing procedures, producing PSTHs and SDFs respectively (Richmond and Optican, 1987). Both are regularisation procedures, attempting to deal with data scarcity and noise by making various (frequently implicit) assumptions. Binning presupposes that the firing probabilities are constant within each bin, whereas smoothing presupposes that high-frequency fluctuations are mostly noise. These assumptions should, however, be evaluated by comparing the predictive performances of different types of models on real neurophysiological data, see Section 4.3, rather than being presupposed.

For an intuitive understanding of the relative merits and drawbacks of traditional PSTH and SDF procedures, see Fig. 1. The left panels show data recorded from an STSa neuron (low background, excitatory response). The right panels show data from an LGN neuron (high background, inhibitory response). The raw data is shown in rastergram form in the top row. Generation of a PSTH with fixed bin duration, optimised for the data by the method described in (Shimazaki and Shinomoto, 2007c,b), is shown in the 2nd row. While a bin PSTH could in principle model sharp transients, the location of the bin boundaries are determined by the constant binwidth. Therefore, the precise onset of the transient is often not captured well. In addition, constant bin duration also forces many bins into time intervals where the spiketrains appears relatively constant, e.g. in [200 ms, 400 ms] of the STSa neuronal response. The SDF, obtained by smoothing these spiketrains with a Gaussian kernel of 10 ms width, is shown in the third row. Noise in the spiketrains is reduced to some degree (e.g. in the interval [200 ms, 400 ms], left column). However, the sharp transient at response onset (indicated by the dashed vertical line in each column), becomes blurred. Thus, smoothing means that potentially relevant timing information will be lost. We also note that point estimates of instantaneous firing rate are frequently extracted from the PSTH and SDF. Given the limited size of data sets obtained from neurophysiological experiments, reliable point estimates are hard to acquire, and measures of posterior uncertainty and variability should be a part of the estimation procedure.

2.2. Bayesian binning

We propose a compromise, allowing us to put the bin boundaries at only those time points where the changes in firing rate actually happen. In essence we keep the bins to allow for rapid changes in the instantaneous firing rate, but allow for varying bin durations to smooth high frequency noise. As a consequence, time intervals in which the firing rate changes little is modelled by one (or a few) bins, thereby reducing the risk of overfitting noise. Uncertainties and variabilities will be computed in an exact Bayesian fashion. The expectations (e.g. expected firing rates) thus generated will therefore have a more continuous appearance, yielding results that are visually similar to a smoothing technique.

We first give a formal definition of our model. We model a PSTH on $[t_{\min},\ t_{\max}]$, discretised into T contiguous intervals of duration $\Delta t = (t_{\max} - t_{\min})/T$ (see Fig. 2). We select a discretisation fine enough so that no more than one spike occurs in a Δt interval for any given spike train (we simply choose a Δt shorter than the absolute refractory period of the neuron under investigation, here $\Delta t = 1$ ms). Spike train i can then be represented by a binary vector $\vec{z}^i$ of dimensionality T. The PSTH is modelled using $M + 1$ contigu-

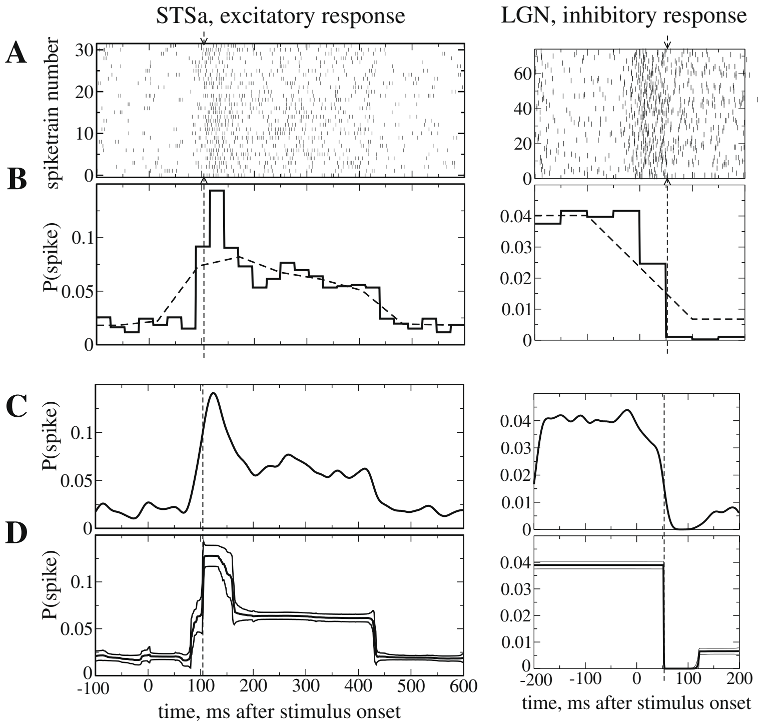

Fig. 1. Predicting a PSTH/SDF with three different methods. *Row A*: Rastergrams of excitatory responses recorded from a STSa neuron (left) and inhibitory responses (right) recorded from a LGN neuron. Each row represents a single stimulus presentation, each tick mark represents the time of a spike relative to stimulus onset (time = 0). *Row B*: bar PSTH (solid lines, optimal binsize using Shimazaki and Shinomoto (2007c)), and line PSTH (dashed lines, optimal binsize using Shimazaki and Shinomoto (2007b)). *Row C*: SDF obtained by smoothing the spike trains with a 10 ms Gaussian kernel. *Row D*: PSTH from Bayesian binning (Endres et al., 2008). The thick line represents the predictive firing rate, the thin lines show the predictive firing rate ±1 standard deviation. Models with $4 \leqslant M \leqslant 12$ were included on a risk level of $\alpha = 0.1$ (see Eq. (13)) for the STSa data, and $2 \leqslant M \leqslant 4$ for the LGN data. The vertical dashed line indicates the mode of the latency posterior (see Section 4.4 and Fig. 5).

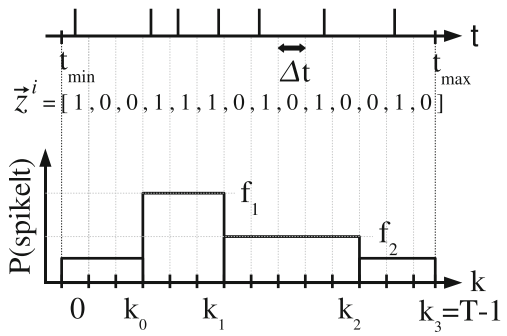

Fig. 2. *Top*: A spike train, recorded between times $t_{\min}$ and $t_{\max}$ is represented by a binary vector $\vec{z}^i$. *Bottom*: The time span between $t_{\min}$ and $t_{\max}$ is discretised into T intervals of duration $\Delta t = (t_{\max} - t_{\min})/T$, such that interval k lasts from $k \times \Delta t + t_{\min}$ to $(k+1) \times \Delta t + t_{\min}$. Δt is chosen such that at most one spike is observed per Δt interval for any given spike train. Then, we model the firing probabilities $P(\text{spike}|t)$ by $M + 1 = 4$ contiguous, non-overlapping bins (M is the number of bin boundaries inside the time span $[t_{\min}, t_{\max}]$), having inclusive upper boundaries k_m and $P(\text{spike}|t \in (t_{\min} + \Delta t(k_{m-1}+1), t_{\min} + \Delta t(k_m + 1)]) = f_m$. For details, see text.

ous, non-overlapping bins having inclusive upper boundaries k_m. The firing probability $P(\text{spike}|t \in (t_{\min} + \Delta t(k_{m-1}+1), \; t_{\min} + \Delta t(k_m + 1)]) = f_m$ is constant within each bin. The relationship between the firing probabilities f_m and the instantaneous firing rates is given by

$$\text{firing rate} = \frac{f_m}{\Delta t} \tag{1}$$

M is the number of bin boundaries within $[t_{\min}, t_{\max}]$. The probability of a spike train $\vec{z}^i$ of independent spikes/gaps is then

$$P(\vec{z}^i|\{f_m\},\{k_m\},M) = \prod_{m=0}^{M} f_m^{s(\vec{z}^i,m)}(1-f_m)^{g(\vec{z}^i,m)} \tag{2}$$

where $s(\vec{z}^i, m)$ is the number of spikes and $g(\vec{z}^i, m)$ is the number of non-spikes, or gaps in spiketrain $\vec{z}^i$ in bin m, i.e. between intervals $k_{m-1}+1$ and k_m (both inclusive). This implies $s(\vec{z}^i,m) + g(\vec{z}^i,m) = T$. In other words, we model spiketrains using an inhomogeneous Bernoulli process with piecewise constant probabilities. For completeness, we define $k_{-1} = -1$ and $k_M = T - 1$. Note that there is no binomial factor associated with the contribution of each bin, as there would be if the spike order within a given bin was irrelevant. However, we do *not* want to ignore the spike timing information, but rather, we build a simplified generative model of the spike train over time. Therefore, the probability of a (multi)set of spiketrains $\{\vec{z}^i\} = \{z_1, \ldots, z_N\}$, assuming independent generation, is

$$\begin{aligned} P(\{\vec{z}^i\}|\{f_m\},\{k_m\},M) &= \prod_{i=1}^{N}\prod_{m=0}^{M} f_m^{s(\vec{z}^i,m)}(1-f_m)^{g(\vec{z}^i,m)} \\ &= \prod_{m=0}^{M} f_m^{s(\{\vec{z}^i\},m)}(1-f_m)^{g(\{\vec{z}^i\},m)} \end{aligned} \tag{3}$$

where $s(\{\vec{z}^i\},m) = \sum_{i=1}^{N} s(\vec{z}^i,m)$ and $g(\{\vec{z}^i\},m) = \sum_{i=1}^{N} g(\vec{z}^i,m)$.

2.3. The priors

We will make a non-informative prior assumption for $p(\{f_m\},\{k_m\})$, namely

$$p(\{f_m\},\{k_m\}|M) = p(\{f_m\}|M)P(\{k_m\}|M) \tag{4}$$

i.e. we have no a priori preferences for the firing rates based on the bin boundary positions (we assume that the bin boundary positions are independent of the firing rates). Note that the prior of the f_m, being continuous model parameters, is a density. Given the form of Eq. (2) and the constraint $f_m \in [0, 1]$, it is natural to choose a conjugate prior

$$p(\{f_m\}|M) = \prod_{m=0}^{M} \mathrm{B}(f_m; \sigma_m, \gamma_m) \tag{5}$$

The Beta density is defined in the usual way (see e.g. Berger, 1985):

$$\mathrm{B}(p;\sigma,\gamma) = \frac{\Gamma(\sigma+\gamma)}{\Gamma(\sigma)\Gamma(\gamma)} p^{\sigma-1}(1-p)^{\gamma-1} \tag{6}$$

There are only finitely many configurations of the k_m. Assuming we have no preferences for any of them, the prior for the bin boundaries becomes

$$P(\{k_m\}|M) = \frac{1}{\binom{T-1}{M}} \tag{7}$$

where the denominator is just the number of possibilities in which M ordered bin boundaries can be distributed across $T - 1$ places (bin boundary M always occupies position $T - 1$, see Fig. 2 , hence there are only $T - 1$ positions left).

2.4. Computing the evidence $P(\{\vec{z}^i\}|M)$ and other posterior expectations

To calculate quantities of interest for a given number of bins $(M + 1)$, e.g. predicted firing probabilities and their variances or expected bin boundary positions, we need to compute averages over the posterior

$$p(\{f_m\},\{k_m\}|M,\{\vec{z}^i\}) = \frac{p(\{\vec{z}^i\},\{f_m\},\{k_m\}|M)}{P(\{\vec{z}^i\}|M)} \tag{8}$$

This requires the evaluation of the evidence, or marginal likelihood of a model with M bins:

$$P(\{\vec{z}^i\}|M) = \sum_{k_{M-1}=M-1}^{T-2} \cdots \sum_{k_0=0}^{k_1-1} P(\{\vec{z}^i\}|\{k_m\},M)P(\{k_m\}|M) \tag{9}$$

where the summation boundaries are chosen such that the bins are non-overlapping and contiguous and

$$P(\{\vec{z}^i\}|\{k_m\},M) = \int_0^1 df_0 \ldots \int_0^1 df_M P(\{\vec{z}^i\}|\{f_m\},\{k_m\},M)p(\{f_m\}|M) \tag{10}$$

At first glance, computing the sums in Eq. (9) seems computationally intensive. M sums over $O(T)$ many summands suggest a computational complexity of $O(T^M)$, which is impractical. To appreciate why, consider the following example: In a typical neurophysiological experiment, we might want to estimate the PSTH in a $T = 700$ ms time window with $\Delta t = 1$ ms. If we tried to model this distribution by $M + 1 = 11$ bins, we would have to check $\binom{699}{10}$ configurations, i.e. the number of possibilities to distribute 10 ordered bin boundaries across 699 places. This is $> 10^{21}$. Even if we checked 10 configurations per microsecond, we would take more than 20 million years to finish. However, as demonstrated in Endres et al. (2008), the computational complexity can be reduced to $O(MT^2)$ using dynamic programming. In the above example, the time to compute the evidence reduces to ≈0.5 s, which is fast enough to be useful.

Other posterior expectations, can be evaluated in the same fashion. For example, given the model parameters $\{k_m\}$, $\{f_m\}$ and M, the predictive firing probability at time index t can formally be written as

$$P(\text{spike}|t, \{f_m\}, \{k_m\}, M) = \sum_{m=0}^{M} f_m \mathscr{T}(t \in \{k_{m-1}+1, k_m\}) \tag{11}$$

where the indicator function $\mathscr{T}(x) = 1$ iff x is true and 0 otherwise. Thus, the sum has exactly one nonzero contribution from that bin which contains t. Multiplying Eq. (11) with Eq. (8) and marginalising $\{f_m\}$ and $\{k_m\}$ yields the predictive firing rate at t given M and the data $\{\vec{z}^i\}$.

3. Model selection vs. model averaging

To choose the best M given $\{\vec{z}^i\}$, or better, a probable range of Ms, we need to determine the model posterior

$$P(M|\{\vec{z}^i\}) = \frac{P(\{\vec{z}^i\}|M)P(M)}{\sum_m P(\{\vec{z}^i\}|m)P(m)} \tag{12}$$

where $P(M)$ is the prior over M, which we assume to be uniform. The sum in the denominator runs over all values of m which we choose to include, at most $m \leqslant T - 1$.

Once $P(M|\{\vec{z}^i\})$ is evaluated, we could select the most probable M'. However, selecting a single M means 'contriving' information, namely that all of the posterior probability is concentrated at M'. It is more appropriate to average any predictions over all possible M, even if evaluating such an average has a computational cost (cost is of $O(T^3)$, since $M \leqslant T - 1$). If the structure of the data allow, it is possible and indeed useful given a large enough T, to reduce this cost. We reduce the computational cost by finding a range of M such that the risk of excluding a model, even though it provides a good description of the data, is low. In analogy to the significance levels of orthodox statistics, we shall call this risk α. If the posterior of M is unimodal (which it has been in most observed cases, see Fig. 3, for an example), we can then choose the smallest interval of Ms around the maximum of $P(M|\{\vec{z}^i\})$ such that

$$P(M_{\min} \leqslant M \leqslant M_{\max}|\{\vec{z}^i\}) \geqslant 1 - \alpha \tag{13}$$

and carry out the averages over this range of M after renormalising the model posterior.

4. Examples and comparison to other methods

4.1. Data acquisition

The experimental protocols used to record data from STSa neurons have been described before (van Rossum et al., 2008). Briefly, extra-cellular single-unit recordings were made using standard techniques from the upper and lower banks of the anterior part of the superior temporal sulcus (STSa) and the inferior temporal cortex (IT) of two monkeys (*Macaca mulatta*) performing a visual fixation task. To measure the effect of contrast on the response, grey-scale versions of preferred and non-preferred stimuli were presented for 333 ms followed by an 333 ms inter-stimulus interval. Stimuli at different Michelson contrast levels were presented in random order. All contrast manipulations were performed after correcting for the measured gamma function of the display monitor. The anterior–posterior extent of the recorded cells was from 7 mm to 9 mm anterior of the interaural plane consistent with previous studies showing visual responses to static images in this region (Bruce et al., 1981; Perrett et al., 1982; Baylis et al., 1987; Oram and Perrett, 1992). The recorded cells were located in the upper bank (TAa, TPO), lower bank (TEa, TEm) and fundus (PGa, IPa) of STS and in the anterior areas of TE (AIT of (Tanaka et al., 1991)). These areas are rostral to FST and we collectively call them the anterior STS (STSa), see Barraclough et al. (2005) for further discussion. The recorded firing patterns were turned into distinct samples, each of which contained the spikes from 300 ms before to 600 ms after the stimulus onset with a temporal resolution of 1 ms.

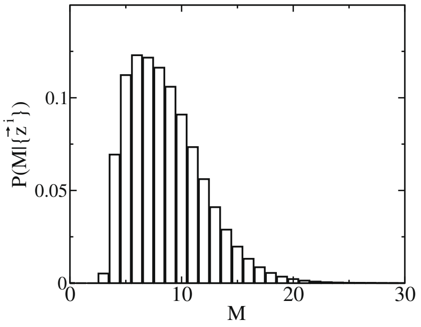

Fig. 3. Model posterior $P(M|\{\vec{z}^i\})$ (see Eq. (12)) computed from the data shown for the STSa neuron in Fig. 1. The shape is fairly typical for model posteriors computed from the neural data used in this paper: a sharp rise at a moderately low M followed by a maximum (here at $M = 6$) and an approximately exponential decay. Even though a maximum M of 699 would have been possible, $P(M > 23|\{\vec{z}^i\}) < 0.001$. Thus, we can accelerate the averaging process for quantities of interest (e.g. the predictive firing rate) by choosing a moderately small maximum M.

Recordings from LGN (see Oram et al., 1999) were made using standard techniques from a rhesus monkey performing a fixation task. Spike data from single neurons were collected with 1 ms resolution. Up to 64 different images were used as stimuli for LGN recordings: bars at four orientations and dots at four sizes, each at up to 8 to different contrast levels. Each stimulus was presented for 300 ms centred on the receptive field. The stimuli covered the excitatory receptive field and extended into the near surround. Reward was delivered after every 1–4 stimulus presentations if the monkey maintained fixation within 0.5 degrees. LGN parvo-cellular neurons were recorded with receptive field centres varying between 3 and 20 degree eccentricities in the lower contralateral hemifield.

4.2. Inferring PSTHs

An example of the PSTH generated by our method is given in Fig. 1. The 32 spiketrains recorded from one neuron in area STSa to a stimulus are shown in rastergram form in the top left. Spikes times are relative to the stimulus onset. For the STSa data we discretised the interval from -100 ms pre-stimulus to 600 ms post-stimulus into $\Delta t = 1$ ms time intervals and computed the model posterior (Eq. (12)) (see Fig. 4, right). The prior parameters were equal for all bins and set to $\sigma_m = 1$ and $\gamma_m = 32$, corresponding to a firing probability of $\approx$30 spikes/s in each 1 ms time interval, typical for the STSa neurons in this study.[1] An analogous approach was used to model the response of one LGN neuron (right column of Fig. 1), illustrating that the approach applies to inhibitory as well as excitatory responses.

Models with $4 \leqslant M \leqslant 12$ (expected bin sizes between $\approx$23 ms and 148 ms) were included on an $\alpha = 0.1$ risk level Eq. (13) for the STSa data, and $2 \leqslant M \leqslant 4$ for the LGN data in the subsequent calculation of the predictive firing rate (i.e. the *expected* firing rate, hence the continuous appearance) and standard deviation (Fig. 1, row D). For comparison, Fig. 1, row B, shows a bar and a line PSTH computed with the recently developed alternative methods described in Shimazaki and Shinomoto (2007c,b). Roughly speaking, these methods try to optimise a compromise between minimal within-bin variance and maximal between-bin variance. In this

[1] Alternatively, one could search for the σ_m, γ_m which maximise of $P(\{\vec{z}^i\}|\sigma_m, \gamma_m) = \sum_M P(\{\vec{z}^i\}|M)P(M|\sigma_m, \gamma_m)$, where $P(\{\vec{z}^i\}|M)$ is given by Eq. (9). Using a uniform $P(M|\sigma_m, \gamma_m)$, we found $\sigma_m \approx 2.3$ and $\gamma_m \approx 37$ for the STSa data in Fig. 1.

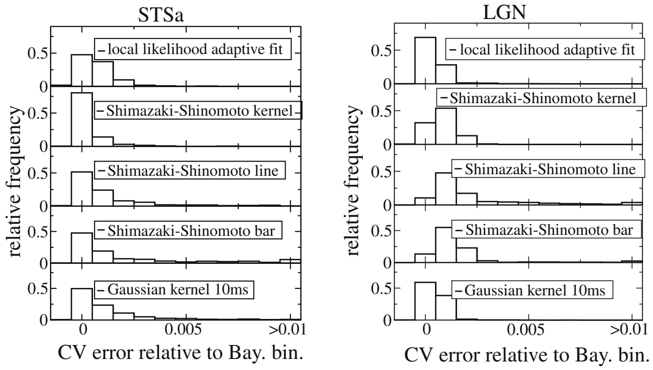

Fig. 4. Comparison of Bayesian binning with competing methods by 5-fold crossvalidation of data sets from STSa (left) and LGN (right). The CV error is the negative expected log-probability of the test data. The histograms show relative frequencies of CV error differences between our Bayesian binning approach and a local likelihood adaptive fit (Loader, 1997, 1999) (top), Shimazaki's and Shinomoto's kernel, line and bar methods ((Shimazaki and Shinomoto, 2007a,b), rows 2–4 respectively) and smoothing with a Gaussian kernel of 10 ms width (bottom).

example, the bar PSTH consists of 26 bins. Row C of Fig. 1 depicts a SDF obtained by smoothing the spiketrains with a 10 ms wide Gaussian kernel.

All four tested methods produce results which are largely consistent with the spiketrains. However, Bayesian binning is better suited than Gaussian smoothing to model steep changes, such as the transient response at ≈100 ms in the STSa response and ≈50 ms in the LGN response. While the binning methods from Shimazaki and Shinomoto (2007c,b) share this property, visual inspection of the rastergrams Fig. 1 suggested that they suffer from two drawbacks: firstly, the evenly spaced bin boundaries means that predicted transients in the PSTH may not match that in the data. Secondly, binning methods in which the bin duration is the only temporal parameter in the model are forced to put many bins even in intervals where the response seems relatively constant. In contrast, Bayesian binning can put bin boundaries anywhere in the time span of interest. Thus, Bayesian binning can model transients accurately and model the sample period with fewer bins – the model posterior for the STSa neuron has its maximum at $M = 6$ (7 bins), whereas the bar PSTH consists of 26 bins – thereby allowing for greater smoothing in periods where the instantaneous rate is (relatively) stable whilst simultaneously capturing transitions in firing rate. The impact of variable bin boundaries and bin widths is even more evident for the data from the LGN neuron where essentially, only three bins are needed.

4.3. Performance comparison

For a quantitative comparison between Bayesian binning and other methods, we split the data into distinct sets, each of which contained the responses of a cell to a different stimulus. This procedure yielded 336 neuron/stimulus combinations from 20 STSa neurons and 1316 neuron/stimulus combinations from 19 LGN neurons with at least 20 spiketrains per combination. We then performed 5-fold crossvalidation, the crossvalidation (CV) error given by the negative logarithm of the data (spike or gap) in the test sets:

$$\text{CV error} = -\langle \log(P(\text{spike}|t)) \rangle \qquad (14)$$

The CV error indicates how well the PSTHs or SDFs generated from the sample data predict the test data. We average the CV error over the five estimates to obtain a single estimate for each of the neuron/stimulus combinations. In Endres et al. (2008), we already demonstrated that Bayesian binning outperforms SDFs obtained by Gaussian smoothing, and the bin and line histogram methods from Shimazaki and Shinomoto (2007c,b). Here, we also test Bayesian binning against the kernel smoothing method described in Shimazaki and Shinomoto (2007a) and a local likelihood adaptive fit (Loader, 1999). We calculated the difference in CV error for each neuron/stimulus combination between Bayesian binning and the alternative method. A positive value indicates that Bayesian binning predicts the test data more accurately than the alternative method. Fig. 4, shows the relative frequencies of CV error differences between the other methods and our approach. In the large majority of cases we are at least as good, but frequently better than the competitors, indicating the general utility of our approach. The average CV error differences, summarised in Table 1, support this claim. Note that while the pattern of the CV error differences varies between STSa and LGN data sets (e.g. Shimazaki and Shinomoto bar method is relatively accurate on average for the LGN data but not STSa data), Bayesian binning outperforms all methods for both the LGN and STSa data.

4.4. Minimal definitions of excitatory and inhibitory response latency

Another frequently used feature for the description of a neuron's response is response latency. However, a precise definition of response latency seems less agreed. If one picked neurophysiologists at random and asked them what exactly latency was, one would possibly receive rather different answers. One statistical based approach, e.g. used in Oram and Perrett (1992), Oram and Perret (1996) and Földiák et al. (2004), is to smooth the stimulus-aligned spiketrains, then determine a baseline level from a section that is believed not to contain a response to the stimulus under investigation (e.g. the so called pre-stimulus period). Next, find the first time index after which the SDF is above baseline + 2.58×(standard deviation of the baseline period) for at least 25 consecutive time indexes. Others have used variants of this definition with somewhat changed parameters. An alternative definition is given in Luczak et al. (2007) who used the mean spike time as a latency measure.

In general, there is some consensus that (excitatory) 'latency is where the signal starts'. The issue, however, is how to determine this point in time. Signal vs. no signal can usually be translated into firing rate above or below a threshold, which we will call the *signal level* (see Fig. 5). We therefore define (excitatory) latency as that point in time prior to which there was no signal, and after which there is a signal for at least some duration. This is the 'minimal' latency definition which we will employ in the following. Conversely, inhibitory latency is that point in time prior to which there was a signal, and after which there is no signal for at least some duration.

Table 1
Mean log prediction error results from 5-fold crossvalidation on 336 STSa datasets and 1316 LGN datasets. A positive value means that our method predicts the data better than the competitors. S.-S. bar: bar PSTH (Shimazaki and Shinomoto, 2007b), S.-S. line: line PSTH (Shimazaki and Shinomoto, 2007b), Gauss: SDF computed by smoothing with a 10 ms wide Gaussian kernel, L.l. fit: local likelihood adaptive fit (Loader, 1997), S.-S. ker. optimised kernel method from Shimazaki and Shinomoto (2007a), Bay. bin.: Bayesian binning.

Method	CV error difference to Bayesian binning	
	STSa	LGN
S.-S. bar	$(2.35 \pm 0.23) \times 10^{-3}$	$(1.686 \pm 0.074) \times 10^{-3}$
S.-S. line	$(1.22 \pm 0.10) \times 10^{-3}$	$(2.400 \pm 0.085) \times 10^{-3}$
Gauss	$(1.29 \pm 0.11) \times 10^{-3}$	$(4.73 \pm 0.14) \times 10^{-4}$
L.l. fit	$(7.34 \pm 0.48) \times 10^{-4}$	$(4.32 \pm 0.16) \times 10^{-4}$
S.-S. ker.	$(3.14 \pm 0.39) \times 10^{-4}$	$(8.21 \pm 0.26) \times 10^{-4}$
Bay. bin.	0	0

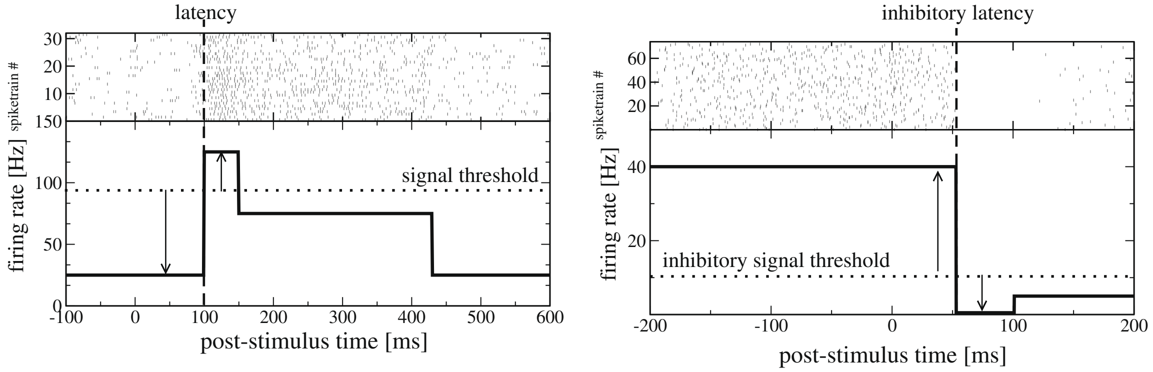

Fig. 5. Our minimal latency definitions. *Left*: (excitatory) Latency is that point in time before which the firing probability was consistently below the signal level (dotted horizontal line), and after which the firing probability is above the signal level for at least one bin. *Right*: Inhibitory latency is that point in time before which the firing probability was consistently above the inhibitory signal level (dotted horizontal line), and after which the firing probability is below the signal level for at least one bin. These definitions have two important implications: the latency is at a bin boundary, and there can be at most one latency (possibly none).

With bin boundaries $\{k_m\}$ and firing probabilities $\{f_m\}$, the latency must be at a bin boundary because firing probabilities are constant within each bin. Note also that our latency definition implies that there can be at most one (excitatory or inhibitory) latency. Furthermore, if the firing probabilities are below the signal level in every bin or if the firing rate in the first bin (f_0) is already above the signal level, then there will be no (excitatory) latency. Likewise, if the firing rate in the first bin is already below the signal level or if it is above the signal level everywhere, then there will be no inhibitory latency.

To obtain a latency posterior distribution we formally define the probability that the latency is at time index t given $\{k_m\}$, $\{f_m\}, M$ and the signal level S as

$$P(\text{excitatory latency at } t|\{k_m\},\{f_m\},M,S) = \begin{cases} 1 & \text{if } \exists\ k_j \in \{k_m\} : k_j + 1 = t \\ & \text{and } f_j \geqslant S \text{ and } \ \forall\ i < j : f_i < S \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

which can be exactly averaged over the posterior Eq. (8) by a dynamic programming algorithm similar to that used for the evidence evaluation. The general framework for computing expectations of functions of bin boundaries and firing probabilities is equivalent to that of Endres and Földiák (2005). Similarly, define

$$P(\text{inhibitory latency at } t|\{k_m\},\{f_m\},M,S) = \begin{cases} 1 & \text{if } \exists\ k_j \in \{k_m\} : k_j + 1 = t \\ & \text{and } f_j \leqslant S \text{ and } \ \forall\ i < j : f_i > S \\ 0 & \text{otherwise} \end{cases} \tag{16}$$

and average over the posterior Eq. (8) to obtain the inhibitory latency posterior.

Assuming the data span the response range of the neuron (i.e. the data contain responses to at least one effective stimulus), one can determine the signal level S as follows. For a given S, marginalise the latency posterior across the time interval of interest, thereby obtaining the probability P_S that a signal exists at that S. Repeat this procedure for different S until the maximal P_S is found. Results obtained by this procedure are shown in Fig. 6.

The latency posterior of an excitatory response from the STSa neuron/stimulus combination shown in Fig. 1 has two distinct modes (Fig. 6, left B). The first peak is at ≈83 ms, the second peak ≈104 ms after stimulus onset. The two peaks can be understood from the rastergrams (Fig. 6A): there appears to be an earlier response onset in some of the trials. The latency posterior of the

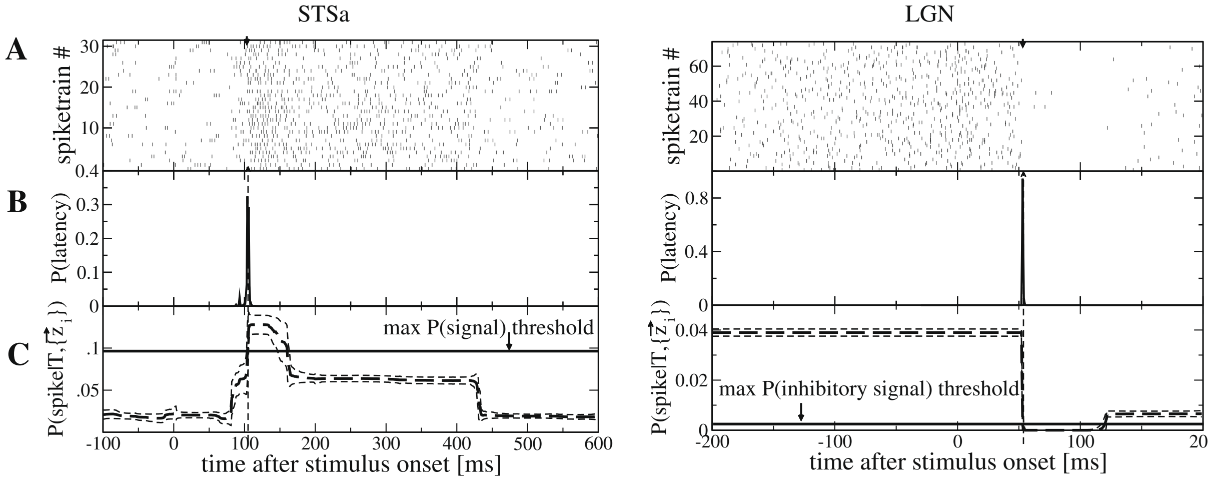

Fig. 6. *Left*: Excitatory response in STSa. *Right*: Inhibitory response in LGN. *Row A*: Each tick mark represents a spike, recorded from (left) the same STSa neuron and (right) the same LGN neurons as in Fig. 1. *Row B*: Latency posterior. Left: The main mode of P(latency) of the STSa neuron is at 104 ms after stimulus onset, indicated by the dashed vertical lines. There is also a smaller mode at 83 ms which is due to an earlier response onset in some trials. The single mode of *P*(latency) of the LGN neuron is at 43 ms. *Row C*: Expected instantaneous firing rates (thick dashed line) plus/minus one standard deviation (thin dashed lines). The signal level *S* for the latency posterior calculation was chosen so that the probability for the existence of a signal, P_S, was maximised. This signal level is indicated by the thick horizontal line.

inhibitory response from the same LGN neuron/stimulus combination in Fig. 1 is shown in Fig. 6, right B. Here a single distinct peak is evident, indicating that our Bayesian binning method is capable of detecting the latency of inhibitory as well as excitatory responses.

4.5. Effects of sample size, background firing rate and signal-to-noise ration on latency inference

An important aspect of inferring the PSTH and latency from neuronal data is the sensitivity of the method to the number of trials. Recordings frequently yield limited sample sizes and it is important that any analysis degrades gracefully as the amount of available data decreases. We investigated this degradation by drawing 30 partially overlapping sub-samples containing 1, 3, 10 or 30 trials from the datasets shown in Fig. 6. For each sub-sample, we either computed the squared difference of the expected latency (STSa) or the expected inhibitory latency (LGN) to the respective latency computed from the full dataset. The resulting root-mean-squared (RMS) deviations are shown in Fig. 7. While the quality of the latency estimates clearly decreases with the number of trials, they are closer than one would expect by chance for even a single trial (assuming uniformly distributed latency guesses in the interval [0 ms, 200 ms], one would expect a RMS of ≈57 ms). The slower convergence of the STSa latencies with increasing number of trials can be attributed to the same reason as the bi-modality of the latency posterior of the full dataset (see Fig. 6, left): this STSa neuron responded earlier in some of the trials. If a sub-sample contains mostly those early response trials, then the latency posterior will have a strong early mode, which will increase the RMS.

We also studied the effects of a changing signal-to-noise ratio (SNR) on the quality of the latency estimates. To vary the SNR, we simulated neural responses drawn from an inhomogeneous Bernoulli process with a latency of 80 ms, after which the neuron fired with a rate of 80 Hz for 50 ms. The baseline firing rate before the latency was varied between 5 Hz (high SNR) and 50 Hz (low SNR). The resulting latency posteriors for datasets containing 30 trials are shown in Fig. 8, top half. 'Per trial' posteriors are obtained by calculating a latency posterior from each trial and averaging them across all trials. Latency posteriors computed from all trials are concentrated in the vicinity of the generating latency up to a baseline of 50 Hz. When computing the latency posteriors per trial, the posterior standard deviations become too large to be useful for baselines above 30 Hz.

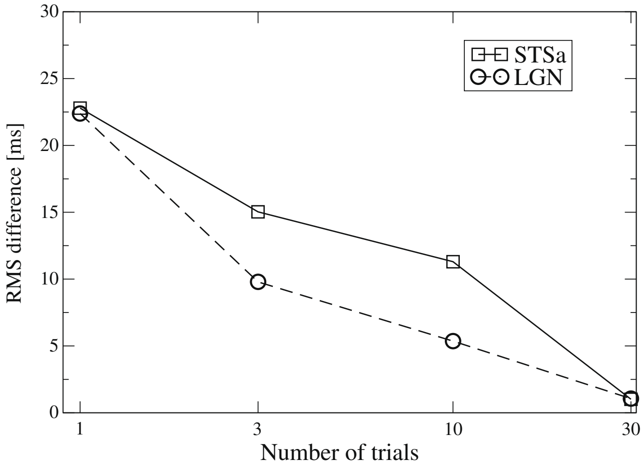

Fig. 7. Effect of sample size on the consistency of latency estimates. For a given number of trials, we drew 30 random, partially overlapping sub-samples from the datasets shown in Fig. 6 and computed the expected latencies (for STSa) and the expected inhibitory latencies (LGN). Root-mean-square (RMS) deviations are between the sub-sample expectations and the latency expectation calculated from all trials. For details, see text.

Fig. 8, bottom half, shows the effect of shifting the background rate and the response, i.e. the firing rate in the first 50 ms after latency is given by (80 Hz + background). While an increased background rate broadens the latency posterior, a latency is clearly detectable even in the 'per trial' evaluation with a 50 Hz background.

5. Trial-by-trial latency and firing rate estimation

Most of our previous analyses have assumed that there is a single 'correct' PSTH from which the data were generated. In other words, we presupposed that the experimentally controlled parameters (e.g. stimulus identity and presentation time) specified the spike train generating process up to a random element, which is fully modelled by the firing probability. It is certainly conceivable that, for example, latencies and firing rates vary between trials. We show here that our method allows computation of the posterior distributions of these parameters on a trial-by-trial basis. Fig. 9, left, shows the trial-by-trial latency posterior distribution marginalised across all trials. The high contrast latency posterior was calculated on the same data as those used in Fig. 6. While the posterior uncertainty is increased due to the trial-by-trial evaluation, the bulk of the probability is in the same post-stimulus time range (≈75–110 ms) as before, which indicates the usefulness of our approach for trial-by-trial evaluations. Moreover, the trial-by-trial latency tends to increase with decreasing stimulus contrast, as observed using estimates from stimulus based SDFs (Oram et al., 2002; van Rossum et al., 2008) using a statistical based approach to latency estimation.

Our trial-by-trial latency estimation method also captures inhibitory responses seen in the LGN data (Fig. 9, right). Note that the distribution is centred at the same time as the latency estimate obtained from assuming a single PSTH for the neuron/stimulus combination (see Fig. 6). The trial-by-trial latency posterior from the inhibitory LGN response declines more slowly than observed for the excitatory STSa responses. This reflects the difficulty in determining the absence of a response when the likelihood of the occurrence of a spike is low: you have to wait a long time to be sure that there isn't a response, whereas, for excitatory responses, a few spikes in quick succession are a good indicator of a change in neuronal activity.

6. Summary

We show that our exact Bayesian binning method treats uncertainty – a real problem with neurophysiological datasets – in a principled fashion, and that it outperforms competing methods on real neural data. It offers automatic complexity control because the model posterior can be evaluated. While its computational cost is significantly higher than that of the methods we compared it to, it is still fast enough to be useful: evaluating the predictive probability takes less than 1s on a modern PC,[2] with a small memory footprint (<10 MB for 512 spiketrains). A free software implementation is available at the machine learning open source software repository.[3]

We have extended our previous studies (Endres et al., 2008) to show how our approach allows extraction of characteristic features of neural responses in a Bayesian way, e.g. excitatory or inhibitory response latencies. We demonstrated the robustness of our latency estimation method against shifts in the

[2] 3.2 GHz Intel Xeon™, SuSE Linux 10.1.

[3] http://www.mloss.org, package 'binsdfc'.

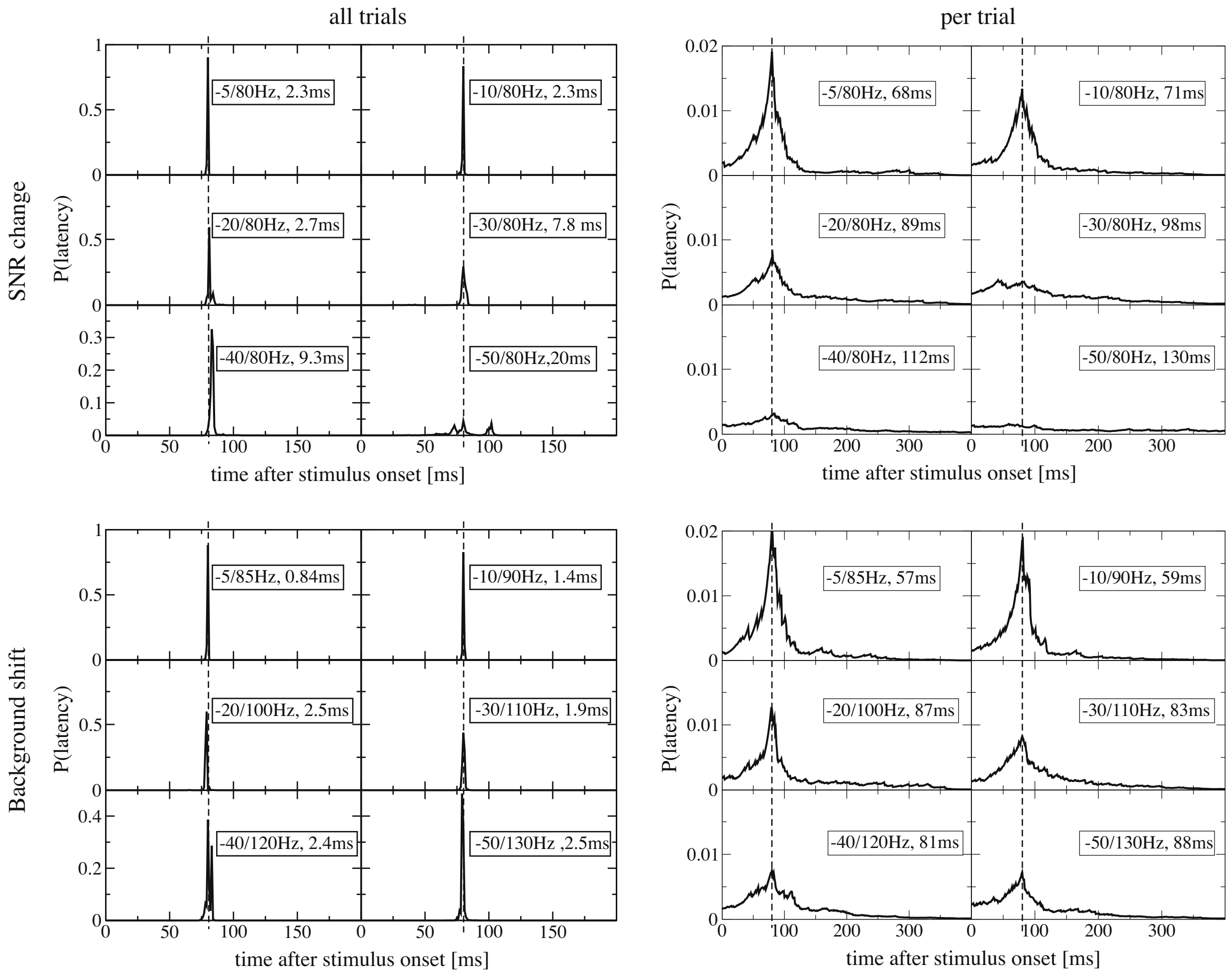

Fig. 8. Effect of signal-to-noise ratio (SNR) change (top half) and shift in the background firing rate (bottom half). Latency posteriors of artificial data drawn from an inhomogeneous Bernoulli process were computed from all trials (left column), i.e. one PSTH for all 30 trials, and per trial (right column). The latency is indicated by the dashed vertical line and was at 80 ms after the stimulus onset in all cases. The legends indicate the background/response firing rates and latency posterior standard deviation in ms. *Top half*: SNR was changed by increasing the background firing rate from 5 Hz to 50 Hz, while the response rate in the first 50 ms after latency was fixed at 80 Hz. A large increase in firing rate is easier to detect than a small one, hence the latency posteriors are more concentrated for large increases. Latency posteriors computed from all trials are concentrated in the vicinity of the generating latency up to a baseline of 50 Hz. When computing the latency posteriors per trial, the posterior standard deviations become too large to be useful for baselines above 30 Hz. *Bottom half*: Shifting the background rate and the response, i.e. the firing rate in the first 50 ms after latency is given by (80 Hz + background) for background rates between 5 Hz and 50 Hz. Here, a latency is clearly detectable even in the 'per trial' evaluation with a 50 ms baseline.

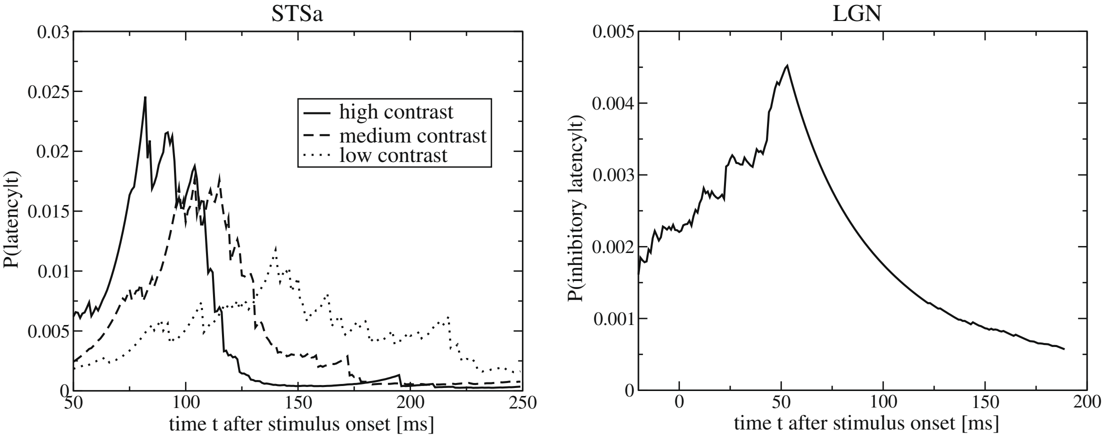

Fig. 9. Trial-by-trial latency posteriors. *Left*: Excitatory responses. Latency posterior was computed for each trial and then marginalised across all trials for each of three stimulus contrasts for an STSa neuron. The high contrast posterior was calculated on the same data as the latency posterior in Fig. 6. While the posterior uncertainty is increased due to the trial-by-trial evaluation, the bulk of the probability is in the same post-stimulus time range (≈75–110 ms) as before. Reducing stimulus contrast clearly increases latencies. *Right*: Inhibitory responses. The latency posterior was computed for each trial and then marginalised across all trials for the LGN neuron using the same data as the latency posterior in Fig. 6. As with excitatory responses, the posterior is centred in the same post-stimulus time range as seen before. Note, however, the gradual decline in the posterior probability after the peak (see text for details).

background firing rate and changes of the signal-to-noise ratio. However, we note that we are not restricted these features: our method can be used to compute expectations of any function of the PSTH, subject to the condition that the function depends on the PSTH in a bin-wise fashion. For example, it is possible to compute exact (up to roundoff errors) expectations of information-theoretic quantities, e.g. mutual informations between latencies and stimulus contrast.

We note that we can substitute our observation model Eq. (2) with any other distribution in a straightforward way, as long as the replacement is also comprised of bins. For example, one might model each spike train within a bin by a separate Bernoulli process and mix these with a suitable distribution to capture the inter-trial differences. Alternatively, one could use a model similar to that of Shinomoto and Koyama (2007): choose a Gamma process for the inter-spike intervals and model the time-dependent rate with a bin model. The relative value of such changes remains to be investigated.

It is clear that our Bayesian approach outperforms other methods in terms of capturing the changes in activity in early (LGN) and late (STSa) areas of the visual pathway, including a range of different background activity levels (LGN ≈30–50 Hz, STSa ≈2–20 Hz) and for both excitatory and inhibitory responses. However, a number of features found in neuronal recordings were not examined. The response onsets of visually responsive neurons are typically rapid, favouring binning over kernel smoothing methods. Neurons in, for example, the motor cortex have, on average, a relatively gradual increases in activity prior to muscle activation. Visual inspection of rastergrams from motor cortical recordings suggests that on individual trials the response onset may be brisk but the time of the onset is only loosely linked to muscle activation. It remains to be seen how our Bayesian binning method copes with these types of data and whether we find evidence that the trial-by-trial PSTHs provide a significantly better fit than assuming a single PSTH applicable to all trials.

While we have shown the advantages of the Bayesian binning method over other estimation methods with real data from the visual system, this does not address potential limitations and the range over which our conclusions are valid. Sensory systems nearly all show adaptation and, although our method works with real sensory data, we have not examined specifically how adaptation influences the performance. The robustness results in Section 4.5 indicate that our approach should be able to deal with adaptation. Further analysis using artificial data sets from known generators will allow us to examine the impact of adaptation on the range of validity and what happens outside the applicable range. Additionally, we note that there are several other approaches to PSTH/SDF estimation, the most noteworthy (from a Bayesian perspective) are (Shinomoto and Koyama, 2007), Bayesian Adaptive Regression Splines (BARS) (Kass et al., 2005) and a recent Gaussian process model (Cunningham et al., 2008). We have not yet directly compared our method to either of them, but (Cunningham et al., 2008) reports that their Gaussian process model performs frequently better than BARS on both simulated and real neural data. Thus, comparisons to BARS and the methods of Cunningham et al. (2008) and Shinomoto and Koyama (2007) using both real and artificial data will be interesting future work.

Acknowledgements

D. Endres was supported by MRC special training fellowship in bioinformatics G0501319. M.W. Oram was supported by Fogerty International Reseach Fellowship and EU Framework Grant (FP5-MIRROR). We would like to thank the unknown reviewers for their helpful comments. D. Endres also thanks H. Shimazaki and A. Leblois for stimulating discussions.

References

Barraclough, N., Xiao, D., Baker, C., Oram, M., Perrett, D., 2005. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. Journal of Cognitive Neuroscience 17.

Baylis, G., Rolls, E., Leonard, C., 1987. Functional subdivisions of the temporal lobe neocortex. Journal of Neuroscience 7, 330–342.

Berger, J., 1985. Statistical Decision Theory and Bayesian Analysis. Springer, New York.

Bruce, C., Desimone, R., Gross, C., 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. Journal of Neurophysiology 46, 369–384.

Cunningham, J., Yu, B., Shenoy, K., Sahani, M., 2008. Inferring neural firing rates from spike trains using Gaussian processes. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems, vol. 20. MIT Press, Cambridge, MA.

Endres, D., Földiák, P., 2005. Bayesian bin distribution inference and mutual information. IEEE Transactions on Information Theory 51.

Endres, D., Oram, M., Schindelin, J., Földiák, P., 2008. Bayesian binning beats approximate alternatives: estimating peri-stimulus time histograms. In: Platt, J., Koller, D., Singer, Y., Roweis, S. (Eds.), Advances in Neural Information Processing Systems, vol. 20. MIT Press, Cambridge, MA.

Földiák, P., Xiao, D., Keysers, C., Edwards, R., Perrett, D.I., 2004. Rapid serial visual presentation for the determination of neural selectivity in area STSa. Progress in Brain Research 144, 107–116.

Kass, R., Ventura, V., Brown, E., 2005. Statistical issues in the analysis of neuronal data. Journal of Neurophysiology 94, 8–25.

Loader, C. 1997. LOCFIT: An Introduction. Statistical Computing and Graphics Newsletter 8:11–17. <http://www.stat-computing.org/newsletter>.

Loader, C., 1999. Local Regression and Likelihood. Springer, New York.

Luczak, A., Bartho, P., Marguet, S., Buzsáki, G., Harris, K., 2007. Sequential structure of neocortical spontaneous activity in vivo. Proceedings of the National Academy of Sciences 104, 347–352.

Oram, M., Perret, D., 1996. Integration of form and motion in the anterior superior temporal polysensory area (stpa) of the macaque monkey. Journal of Neurophysiology 19, 109–129.

Oram, M.W., Perrett, D.I., 1992. Time course of neural responses discriminating different views of the face and head. Journal of Neurophysiology 68, 70–84.

Oram, M.W., Wiener, M.C.R.L., Richmond, B., 1999. The stochastic nature of precisely timed spike patterns in visual system neural responses. Journal of Neurophysiology 81, 3021–3033.

Oram, M.W., Xiao, D., Dritschel, B., Payne, K., 2002. The temporal precision of neural signals: a unique role for response latency? Philosophical Transactions of the Royal Society, Series B 357, 987–1001.

Perrett, D., Rolls, E., Caan, W., 1982. Visual neurons responsive to faces in the monkey temporal cortex. Experimental Brain Research 47, 329–342.

Richmond, B.J., Optican, L.M., 1987. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex.ii. quantification of response waveform. Journal of Neurophysiology 57, 147–161.

Shimazaki, H., Shinomoto, S. 2007a. Kernel width optimization in the spike-rate estimation. In: Budelli, R., Caputi, A., Gomez, L. (Eds.), Neural Coding, pp. 143–146.

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

Shimazaki, H., Shinomoto, S., 2007c. A recipe for optimizing a time-histogram. In: Schölkopf, B., Platt, J., Hoffman, T. (Eds.), Advances in Neural Information Processing Systems, vol. 19. MIT Press, Cambridge, MA, pp. 1289–1296.

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

Tanaka, K., Saito, H., Fukada, Y., Moriya, M., 1991. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. Journal of Neurophysiology, pp–189.

van Rossum, M.C.W., van der Meer, M.A.A., Xiao, D., Oram, M.W., 2008. Adaptive integration in the visual cortex by depressing recurrent cortical circuits. Neural Computation 20, 1847–1872.