Online Variational Inference for State-Space Models with Point-Process Observations

LETTER Communicated by Emery Brown

Online Variational Inference for State-Space Models with Point-Process Observations

Andrew Zammit Mangion
A.Zammit@shef.ac.uk
Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, U.K.

Ke Yuan
ky08r@ecs.soton.ac.uk
School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K.

Visakan Kadirkamanathan
visakan@sheffield.ac.uk
Department of Automatic Control and Systems Engineering, University of Sheffield, Sheffield S1 3JD, U.K.

Mahesan Niranjan
mn@ecs.soton.ac.uk
School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, U.K.

Guido Sanguinetti
gsanguin@inf.ed.ac.uk
School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, U.K.

We present a variational Bayesian (VB) approach for the state and parameter inference of a state-space model with point-process observations, a physiologically plausible model for signal processing of spike data. We also give the derivation of a variational smoother, as well as an efficient online filtering algorithm, which can also be used to track changes in physiological parameters. The methods are assessed on simulated data, and results are compared to expectation-maximization, as well as Monte Carlo estimation techniques, in order to evaluate the accuracy of the proposed approach. The VB filter is further assessed on a data set of taste-response neural cells, showing that the proposed approach can effectively capture dynamical changes in neural responses in real time.

1 Introduction

Many biomedical signal processing problems, such as neural spikes and heartbeats, are concerned with discrete events in time, separated by seemingly random intervals. They are often driven by continuous processes relating to the organ's physiology, whose charge-and-fire type of behavior results in observed discrete events. Conventional approaches to modeling such signals are largely based on modeling the time intervals between these discrete events, which, as continuous variables, are amenable to standard signal processing and system identification approaches (Ivanov et al., 1996; Jolivet et al., 2008).

An alternative approach is the state-space model with point-process observations (SSPP) recently proposed by Smith and Brown (2003), which avoids the somewhat artificial change to interspike times and handles the discrete events directly. This model assumes a first-order autoregressive process driven by an exogenous stimulus as state dynamics and a parameterized intensity function of an approximate Bernoulli process as its observation model. For simultaneous estimation of state and parameters of such a model, Smith and Brown derived an expectation-maximization (EM) algorithm, along the lines of similar formulations (Roweis & Ghahramani, 1999). In a recent study, Yuan and Niranjan (2010), showed that the expectation of the log-complete data likelihood (Q-function) of the SSPP is unimodal and highly nongaussian with respect to each of its parameters. The high skewness is indicative of parameter posteriors where simple maximum likelihood estimates of the parameters may be quite far from the actual posterior means, motivating a Bayesian treatment of the SSPP.

In this letter, we propose a variational Bayesian (VB) approach to solve this problem, extending the results of Beal (2003) to the SSPP case to obtain a variational smoother that offers a good compromise between distributional accuracy and computational efficiency. The developed techniques are demonstrated on a synthetic data set, showing good performance when compared to EM and fully Bayesian approaches based on Gibbs sampling. The details of a VB filter are also given, using ideas taken from dual filtering (Wan & Nelson, 2001), whereby parameters are allowed to evolve to track changes in the system's mode of operation. A case study based on real data of neural responses to different taste stimuli (di Lorenzo & Victor, 2003) is presented, showing that the online filter correctly predicts a change in the input gain or the background firing rate parameters when the input stimulus is changed.

2 Preliminaries

This work is concerned with point-process observation models, where events are recorded on an interval $(0, T]$ from C independent output channels. The observation length is discretized with a sampling interval $\Delta > 0$,

so that the incoming events are represented as a sequence of binary vectors $\mathbf{y}_k := \mathbf{y}(k\Delta) \in \mathbb{R}^C$, where $y^c(k\Delta) = 1, c = 1, \ldots, C$ indicates that an event has occurred at the cth output channel in the interval $((k-1)\Delta, k\Delta]$ and is zero otherwise. The sampling interval Δ is thus chosen small enough so that at most one event per sample for each output channel is present:

$$\Delta \in \left\{ r;\, y^c(kr) \in \{0, 1\}, k \in [1, \ldots, T/r], c \in [1, \ldots, C] \right\}. \tag{2.1}$$

Given a dynamic latent state $x_k := x(k\Delta)$, for the point-process we define a conditional intensity function (CIF) of the form

$$\lambda_k^c = \lambda(k\Delta | x_k, \mu, \beta^c) = \exp(\mu + \beta^c x_k). \tag{2.2}$$

Through the conditioning on x_k, the CIF renders the process an inhomogeneous Poisson process (see Smith & Brown, 2003). The parameter μ represents a background firing rate, which for simplicity is assumed to be the same for all channels. It can be shown that the observation model (or likelihood) at the kth time interval in the cth channel is given by the approximate probability mass function defined as

$$p\left(y_k^c \mid x_k, \mu, \beta^c\right) = \left[\Delta\lambda_k^c\right]^{y_k^c} \exp\left(-\Delta\lambda_k^c\right). \tag{2.3}$$

Equation 2.3 can be obtained from first principles by treating the binned event sequence as a series of correlated Bernoulli trials (Brown, Barbieri, Eden, & Frank, 2003) and is thus a realistic approximation only if equation 2.1 is ensured and, hence, Δ is sufficiently small. In practice, constraint 2.1 cannot be guaranteed before data collection, but Δ may be chosen such that the probability of expected arrival time within an interval Δ at the maximum expected intensity a priori is less than some predefined threshold.

The underlying state follows the standard linear evolution equation,

$$x_k = \rho x_{k-1} + \alpha I_k + \epsilon_k, \tag{2.4}$$

where $I_k := I(k\Delta)$ is 1 if an input is present at $k\Delta$ and zero otherwise. $\epsilon_k := \epsilon(k\Delta) \in \mathbb{R}$ is additive white gaussian noise with mean 0 and variance $\sigma_\epsilon^2 \in \mathbb{R}^+$. The initial state x_0 is assumed to be normally distributed with known mean $x_{0|0}$ and variance $\sigma_{0|0}^2$. The parameters $\rho \in \mathbb{R}$ and $\alpha \in \mathbb{R}$ are the propagation constant and input gain, respectively.

Since the CIF is itself probabilistic and time varying, equations 2.2 and 2.4 define a doubly stochastic process. Despite the simplicity of the underlying latent process used to describe the state evolution, this model has been applied several times in practice to represent the dynamics of a system variable, the behavior of which is not fully understood. For instance, equation 2.4 has been used successfully to model the spatial receptive field

of a pyramidal neuron in a rat hippocampus (Ergün, Barbieri, Eden, Wilson, & Brown, 2007), and, more recently, to model the arousal state in subjects receiving thalamic stimulation (Smith et al., 2009).

In practice, both the parameters governing the firing rate μ and $\boldsymbol{\beta} = \{\beta^c\}_{c=1}^C$, and the governing state equation parameters α and ρ are unknown. In this work, the noise variance σ_ϵ^2 is assumed to be fixed, and we are hence faced with the problem of having to estimate a set of unknown parameters $\boldsymbol{\theta} \in \mathbb{R}^d$, $d = C + 3$ with $\boldsymbol{\theta} = \{\alpha, \rho, \mu, \beta^1, \beta^2, \ldots, \beta^C\}$ in addition to an underlying hidden state x_k.

3 Batch VBEM for SSPP

The variational framework for the inference in the SSPP is developed in a similar way to Beal (2003). Let $\mathcal{X}_K, \mathcal{Y}_K$ be the set of states and observed data points, respectively, $\mathcal{X}_K = \{x_i\}_{i=0}^K$ and $\mathcal{Y}_K = \{\mathbf{y}_i\}_{i=1}^K$. The problem pivots on finding an approximation to the true posterior $p(\mathcal{X}_K, \boldsymbol{\theta}|\mathcal{Y}_K) \approx \tilde{p}(\mathcal{X}_K, \boldsymbol{\theta})$ such that the variational free energy (or log marginal likelihood) is maximized (Attias, 1999). The approximation is carried out by imposing independence between partitioned variables in the joint distribution. This is a well-known drawback when employing variational Bayesian methods; however, the ensuing factorization is rewarded with significant computational savings.

In this work, the approximate (joint) posterior is assumed to be a product of gaussian distributions:

$$\tilde{p}(\mathcal{X}_K, \boldsymbol{\theta}) = \tilde{p}(\mathcal{X}_K)\tilde{p}(\boldsymbol{\theta}) = \tilde{p}(\mathcal{X}_K)\tilde{p}(\rho|\alpha)\tilde{p}(\alpha)\tilde{p}(\mu)\prod_{i=1}^{C}\tilde{p}(\beta^i).$$

The dependency between the ρ and α parameters is retained since the interaction terms between them, which appear when deriving the log posterior distribution, are relatively easy to compute. As a result, α and ρ are dealt with jointly, and without loss in generality, we redefine the set $\boldsymbol{\theta} = \{(\alpha, \rho), \mu, \beta^1, \beta^2, \ldots, \beta^C\}$. The optimal choice for the variational posteriors $\tilde{p}(\mathcal{X}_K)$ and $\tilde{p}(\boldsymbol{\theta})$ is then given by (Šmídl & Quinn, 2005),

$$\tilde{p}(\mathcal{X}_K) \propto \exp(\langle \ln p(\mathcal{X}_K, \mathcal{Y}_K, \boldsymbol{\theta})\rangle_{\tilde{p}(\boldsymbol{\theta})}), \tag{3.1a}$$

$$\tilde{p}(\theta^i) \propto \exp(\langle \ln p(\mathcal{X}_K, \mathcal{Y}_K, \boldsymbol{\theta})\rangle_{\tilde{p}(\mathcal{X}_K)\tilde{p}(\boldsymbol{\theta}^{/i})}), \tag{3.1b}$$

where θ^i is the ith component in $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{/i}$ is the set of all $\boldsymbol{\theta}$ excluding θ^i. The notation $\langle\cdot\rangle_{p(x)}$ denotes the expectation operator taken with respect to $p(x)$. In the standard case of linear-gaussian dynamical systems, the variational posteriors can be computed exactly (Beal, 2003). For the model under consideration, because of the form of the observation process, this is not possible. However, the nongaussian densities that become apparent in the subsequent derivations are unimodal with respect to the underlying

states and parameters, and simulation studies have shown that they can be reasonably approximated by gaussian densities. We take advantage of this property and introduce approximations in a way similar to Smith and Brown (2003) (see also Friston, Mattout, Trujillo-Barreto, Ashburner, & Penny, 2007) to obtain analytically tractable forward and backward passes for state distribution updates and the subsequent parameter distribution updates.

3.1 Batch Update of $\tilde{p}(\mathcal{X}_K)$. Evaluating equation 3.1a and linearizing as in Smith and Brown (2003), one obtains the following equations governing the forward pass (see section A.1 in appendix A):

$$x_{k|k} = \tilde{x_k} + \tilde{\sigma}_k^2 \sum_{c=1}^{C} \left\{ \langle \beta^c \rangle_{\tilde{p}(\beta^c)} y_k^c - \Delta \langle \exp \mu \rangle_{\tilde{p}(\mu)} \times \frac{d}{dx_k} \left[\langle \exp x_k \beta^c \rangle_{\tilde{p}(\beta^c)} \right] \Big|_{x_k = x_{k|k}} \right\}, \tag{3.2a}$$

$$\sigma_{k|k}^2 = \left(\tilde{\sigma}_k^{-2} + \sum_{c=1}^{C} \left\{ \Delta \langle \exp \mu \rangle_{\tilde{p}(\mu)} \frac{d^2}{dx_k^2} \left[\langle \exp x_k \beta^c \rangle_{\tilde{p}(\beta^c)} \right] \Big|_{x_k = x_{k|k}} \right\} \right)^{-1}, \tag{3.2b}$$

where

$$\begin{aligned} \frac{\tilde{x_k}}{\tilde{\sigma}_k^2} &= \left(\left(\sigma_{k-1|k-1}^{-2} + \langle \rho^2 \rangle \sigma_\epsilon^{-2} \right)^{-1} \langle \rho \rangle \sigma_\epsilon^{-2} \left[x_{k-1|k-1} \sigma_{k-1|k-1}^{-2} - \langle \rho \alpha \rangle I_k \sigma_\epsilon^{-2} \right] \right. \\ &\quad \left. + \langle \alpha \rangle I_k \sigma_\epsilon^{-2} \right), \\ \tilde{\sigma}_k^2 &= \left(\sigma_\epsilon^{-2} - \langle \rho \rangle^2 \left(\sigma_{k-1|k-1}^{-2} + \langle \rho^2 \rangle \sigma_\epsilon^{-2} \right)^{-1} \sigma_\epsilon^{-4} \right)^{-1}. \end{aligned} \tag{3.3}$$

Equation 3.2a is composed of two terms—the first pertaining to the underlying linear dynamical model and the second to the observation point-process. Considering the nonlinear form of equation 3.2a, it can be shown that if each $\beta^c > 0$ and $\langle \beta^{c^2} \rangle_{\tilde{p}(\beta^c)} \approx \langle \beta^c \rangle_{\tilde{p}(\beta^c)}^2$, the forward estimate tends to be lowered by a lack of events (indicative of a decreasing intensity). On the other hand, $y_k^c = 1$ tends to increase the estimated $x_{k|k}$. The effect of the number of output channels C is also apparent by evaluating the sum in equation 3.2b, from which it is easily seen that the precision $\sigma_{k|k}^{-2}$ increases with increasing C (assuming β^c is constant across all channels).

The forward state update equations depend not on the actual values of the parameters, but rather on their first and second moments under the approximating distribution. This averaging, which will be evident in all

of the following update equations, is at the core of mean field variational algorithms, which originated in statistical physics, where the interdependence between states was replaced by a dependence on the average (mean) value of the states. For conciseness, in equation 3.3 and in some of the following equations, the distributions with which the expectations are taken with respect to are omitted. The normal assumption for the variational distributions allows analytical computation of the expectations involved in equations 3.2a, 3.2b, and 3.3.

In a similar fashion, a backward recursion on the data is computed in order to obtain variational smoothed state estimates (see section A.2). The resulting equations are given as

$$x_{k|K} = \sigma^2_{k|K}\left(x_{k|k}\sigma^{-2}_{k|k} + x^*_k\sigma^{*^{-2}}_k\right), \qquad \sigma^2_{k|K} = \left(\sigma^{-2}_{k|k} + \sigma^{*^{-2}}_k\right)^{-1},$$

where

$$\begin{aligned}\frac{x^*_k}{\sigma^{*^2}_k} &= \left(\langle\rho\rangle x'_{k+1}\left(\sigma_\epsilon^{-2} + \sigma^{'-2}_{k+1}\right)^{-1}\sigma_\epsilon^{-2}\sigma^{\prime-2}_{k+1}\right.\\ &\quad\left. + \left(\sigma_\epsilon^{-2} + \sigma^{'-2}_{k+1}\right)^{-1}\langle\rho\rangle\langle\alpha\rangle I_{k+1}\sigma_\epsilon^{-4} - \langle\rho\alpha\rangle I_{k+1}\sigma_\epsilon^{-2}\right),\\ \sigma^{*^2}_k &= \left(\langle\rho^2\rangle\sigma_\epsilon^{-2} - \left(\sigma_\epsilon^{-2} + \sigma^{'-2}_{k+1}\right)^{-1}\langle\rho\rangle^2\sigma_\epsilon^{-4}\right)^{-1},\end{aligned}$$

and

$$\begin{aligned}x'_{k+1} &= x_{k+1|k+1} + \sigma^{\prime 2}_{k+1}\left(\frac{x^*_{k+1} - x_{k+1|k+1}}{\sigma^{2^*}_{k+1}} + \sum_{c=1}^{C}\left\{\langle\beta^c\rangle_{\tilde{p}(\beta^c)}y^c_{k+1}\right.\right.\\ &\quad\left.\left. - \Delta\langle\exp\mu\rangle_{\tilde{p}(\mu)}\frac{d}{dx_{k+1}}\left[\langle\exp x_{k+1}\beta^c\rangle_{\tilde{p}(\beta^c)}\right]\Big|_{x_{k+1}=x_{k+1|k+1}}\right\}\right), \qquad (3.4a)\\ \sigma^{\prime 2}_{k+1} &= \left(\sigma^{*^{-2}}_{k+1} + \sum_{c=1}^{C}\left\{\Delta\langle\exp\mu\rangle_{\tilde{p}(\mu)}\frac{d^2}{dx^2_{k+1}}\right.\right.\\ &\quad\left.\left.\times\left[\langle\exp x_{k+1}\beta^c\rangle_{\tilde{p}(\beta^c)}\right]\Big|_{x_{k+1}=x_{k+1|k+1}}\right\}\right)^{-1}. \qquad (3.4b)\end{aligned}$$

In equations 3.4a and 3.4b, the gaussian approximation is carried out around the filtered estimate to give a closed-form solution. As a consequence, the forward and backward passes need to be carried out sequentially. If the initial state distribution is not known when the backward pass is completed, it may be updated by setting $x_{0|0} = x_{0|K}$ and variance $\sigma^2_{0|0} = \sigma^2_{0|K}$ (see Beal, 2003).

Equation 3.2a is not available in closed form and needs to be solved by a deterministic optimization method. One can take advantage of the facts that the equation has a unique solution and that the prior $x_{k|k-1}$ (obtained from the predictive density) can be used as a good initialization for $x_{k|k}$ to solve the optimization method in an efficient manner. In practice it was found that replacing the state variable on the right-hand side by the prior (to obtain a closed-form solution) gave very good results and a marked decrease in computational requirements.

The required statistics needed for updating the parameter variational posteriors are $\langle x_k x_{k+1}\rangle_{\tilde{p}(\mathcal{X}_K)}$, $\langle x_k^2\rangle_{\tilde{p}(\mathcal{X}_K)}$, and $\langle x_k\rangle_{\tilde{p}(\mathcal{X}_K)}$ for all time k. The only quantity that is not readily available from the above is the first of these expectations. To obtain this, we invert the precision of the approximate pairwise marginal $p(x_k, x_{k+1}|\mathcal{Y}_K)$ to get

$$\langle x_k x_{k+1}\rangle_{\tilde{p}(\mathcal{X}_K)} = \langle\rho\rangle\sigma_\epsilon^{-2}\left(\left(\sigma_{k|k}^{-2} + \langle\rho^2\rangle\sigma_\epsilon^{-2}\right)\sigma^{l^{-2}} - \langle\rho\rangle^2\sigma_\epsilon^{-4}\right)^{-1} + x_{k+1|K}\,x_{k|K},$$

where

$$\sigma^{l^2} = \left(\sigma_{k+1}^{*^{-2}} + \sigma_\epsilon^{-2} + \sum_{c=1}^{C}\left\{\Delta\langle\exp\mu\rangle\,\frac{d^2}{dx_{k+1}^2} \times \left[\langle\exp x_{k+1}\beta^c\rangle_{\tilde{p}(\beta^c)}\right]\Big|_{x_{k+1}=x_{k+1|K}}\right\}\right)^{-1}.$$

After computing the state-sufficient statistics, one can update the parameter variational posteriors as described next.

3.2 Batch Update of $\tilde{p}(\theta)$. Equation 3.1b gives the updates for the parameter distributions. As a direct consequence of the underlying linear state evolution model, the optimal variational estimates for α and ρ become identical to those in a linear dynamical system, so we refer readers to Beal (2003) for details. The estimation of μ and β^c is somewhat more involved, and we refer the readers to appendix B for their treatment. Denoting the means and variances of μ and β^c as $\hat{\mu}$, $\hat{\beta}^c$ and σ_μ^2, $\sigma_{\beta^c}^2$, respectively, we have that

$$\hat{\beta}^c = \beta_p^c + \sigma_{\beta_p^c}^2 \times \sum_{i=1}^{K}\left(y_i^c\langle x_i\rangle_{\tilde{p}(\mathcal{X}_K)} - \Delta\langle\exp\mu\rangle_{\tilde{p}(\mu)}\frac{d}{d\beta^c}\left[\langle\exp x_i\beta^c\rangle_{\tilde{p}(\mathcal{X}_K)}\right]\Big|_{\beta^c=\hat{\beta}^c}\right),$$

$$\sigma^2_{\beta^c} = \left(1/\sigma^2_{\beta^c_p} + \Delta\langle\exp\mu\rangle_{\tilde{p}(\mu)} \sum_{i=1}^{K} \left[\frac{d^2}{d\beta^{c2}} \langle\exp x_i\beta^c\rangle_{\tilde{p}(\mathcal{X}_K)}\Big|_{\beta^c=\hat{\beta}^c}\right]\right)^{-1},$$

and

$$\hat{\mu} = \mu_p + \sigma^2_{\mu_p} \sum_{i=1}^{K}\sum_{c=1}^{C} \left(y_i^c - \Delta\exp(\hat{\mu})\langle\exp(\beta^c x_i)\rangle_{\tilde{p}(\mathcal{X}_K)\tilde{p}(\beta^c)}\right),$$

$$\sigma^2_{\mu} = \left(1/\sigma^2_{\mu_p} + \Delta\exp(\hat{\mu}) \sum_{i=1}^{K}\sum_{c=1}^{C}\langle\exp(\beta^c x_i)\rangle_{\tilde{p}(\mathcal{X}_K)\tilde{p}(\beta^c)}\right)^{-1},$$

where the subscript p denotes *prior*. For this study, we have taken gaussian prior distributions over all parameters, with hyperparameters assumed to be known.

All expectations required to compute $\tilde{p}(\boldsymbol{\theta})$ are standard except for $\langle\exp(\beta^c x_i)\rangle_{\tilde{p}(\mathcal{X}_K)\tilde{p}(\beta^c)}$, which can be calculated using moment-generating functions (see appendix B). As is standard in VB estimation, updates for specific variables depend on the expectations of the remaining variables, leading to a natural iterative algorithm. Convergence can be easily assessed by monitoring changes in the free energy or in the statistics of the variational distributions.

4 Online VB for SSPP

The offline VB algorithm of section 3 can be extended for use in an online scenario with some modifications. Using a standard technique in dual filtering (Wan & Nelson, 2001), we introduce a time evolution model for the parameters,

$$\boldsymbol{\theta}_k = \boldsymbol{\theta}_{k-1} + \mathbf{e}_k,$$

where $\mathbf{e}_k \in \mathbb{R}^d$ is additive white gaussian noise with diagonal covariance matrix $\Sigma^{\mathbf{e}}_{\mathbf{k}} \in \mathbb{R}^{d\times d}$, which is also time varying (see below). Let $\Theta_k = \{\boldsymbol{\theta}_i\}_{i=1}^k$. Equations 2.2 and 2.4 now become

$$\lambda_k^c = \exp(\mu_k + \beta_k^c x_k),$$
$$x_k = \rho_k x_{k-1} + \alpha_k I_k + \epsilon_k.$$

The online variational posteriors are given as

$$\tilde{p}(\mathcal{X}_k) \propto \exp(\langle[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \Theta_k)]\rangle_{\tilde{p}(\Theta_k)}), \tag{4.1a}$$

$$\tilde{p}(\theta_k^i) \propto \exp\left[\langle[\ln p(\mathcal{X}_k, \mathcal{Y}_k, \Theta_k)]\rangle_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\Theta_k^{/\theta_k^i})}\right], \tag{4.1b}$$

where $\tilde{p}(\Theta_k^{/\theta_k^i})$ is the joint $\tilde{p}(\Theta_k)$ without the variable θ_k^i. We choose the following variational posteriors,

$$\begin{aligned}\tilde{p}(\mathcal{X}_k, \Theta_k) &\approx \tilde{p}(\mathcal{X}_k)\prod_{j=1}^{k}\tilde{p}(\boldsymbol{\theta}_j)\\ &= \tilde{p}(\mathcal{X}_k)\tilde{p}(\Theta_k),\end{aligned} \tag{4.2}$$

that is, the parameters are approximated to be conditionally independent in time through the product distribution $\tilde{p}(\Theta_k)$. To facilitate recursion, the parameter variational posteriors are further restricted to be the filtered distributions. We hence redefine $\tilde{p}(\Theta_k)$ as

$$\tilde{p}(\Theta_k) = \prod_{j=1}^{k}\tilde{p}(\boldsymbol{\theta}_j|\mathcal{Y}_j). \tag{4.3}$$

At each time step the distributions $\tilde{p}(\mathcal{X}_k)$ and $\tilde{p}(\boldsymbol{\theta}_k)$ are variational posteriors in the conventional sense. We refer to $\{\tilde{p}(\boldsymbol{\theta}_j)\}_{j=1}^{k-1}$ as the restricted variational posteriors, as is typical in restricted variational Bayes methods (Šmídl & Quinn, 2006). A novel result for dual VB filtering is presented in the following theorem.

Theorem 1. *For the SSPP described by equations 2.3 and 2.4, given the factorization in equation 4.2, the restriction in equation 4.3, and the maximizers in equations 4.1a and 4.1b, the recursive updates for the state and parameter variational distributions $\tilde{p}(\mathcal{X}_k)$ and $\tilde{p}(\boldsymbol{\theta}_k)$ are given by*

$$\tilde{p}(x_k) \propto \int dx_{k-1}\tilde{p}(x_{k-1})\exp(\langle\ln p(x_k|x_{k-1}, \boldsymbol{\theta}_k)p(\mathbf{y}_k|x_k, \boldsymbol{\theta}_k)\rangle_{\tilde{p}(\boldsymbol{\theta}_k)}), \tag{4.4a}$$

$$\begin{aligned}\tilde{p}(\theta_k^i) \propto{}& \exp(\langle\ln p(\mathbf{y}_k|x_k, \boldsymbol{\theta}_k)p(x_k|x_{k-1}, \boldsymbol{\theta}_k)\rangle_{\tilde{p}(\mathcal{X}_k)\tilde{p}(\boldsymbol{\theta}_k^{/i})})\\ &\times \exp(\langle\ln p(\theta_k^i|\theta_{k-1}^i)\rangle_{\tilde{p}(\theta_{k-1}^i)}), \quad i = 1\ldots d.\end{aligned} \tag{4.4b}$$

Proof. We start by considering the variational approximation of the state marginal, which is given by

$$\begin{aligned}\tilde{p}(x_k) \propto{}& \int d\mathcal{X}_{k-1}\exp(\langle\ln p(\mathcal{X}_k, \Theta_k, \mathcal{Y}_k)\rangle)\\ ={}& \exp(\langle\ln p(\mathbf{y}_k|x_k, \boldsymbol{\theta}_k)\rangle)\int d\mathcal{X}_{k-1}\exp(\langle\ln p(x_k|x_{k-1}, \boldsymbol{\theta}_k)\\ &\times p(\mathcal{X}_{k-1}, \Theta_k, \mathcal{Y}_{k-1})\rangle),\end{aligned} \tag{4.5}$$

where the above expectations are taken with respect to the unknown parameters. The second term of the integrand can also be expanded, and by treating the conditional parameter distributions as constants relative to the distribution of interest, it can be shown that

$$\begin{aligned}\tilde{p}(x_k) \propto \exp(\langle \ln p(\mathbf{y}_k|x_k, \boldsymbol{\theta}_k)\rangle) \int dx_{k-1} \Bigg(& \exp(\langle \ln p(x_k|x_{k-1}, \boldsymbol{\theta}_k)\rangle) \\ \times \Bigg[& \exp(\langle \ln p(\mathbf{y}_{k-1}|x_{k-1}, \boldsymbol{\theta}_{k-1})\rangle) \int d\mathcal{X}_{k-2} \exp(\langle \ln p(x_{k-1}|x_{k-2}, \boldsymbol{\theta}_{k-1})\rangle) \\ & \times \exp(\langle \ln p(\mathcal{X}_{k-2}, \Theta_{k-1}, \mathcal{Y}_{k-2})\rangle) \Bigg] \Bigg). \end{aligned} \tag{4.6}$$

Recall that since the approximate parameter posteriors have been restricted to be conditional on the data up to the instant at which they were estimated, the distributions of the parameters do not need to be recomputed using the latest data which is available. In particular for any function $\psi(\cdot)$,

$$\langle \psi(\boldsymbol{\theta}_{k-1})\rangle_{\tilde{p}(\Theta_k)} = \langle \psi(\boldsymbol{\theta}_{k-1})\rangle_{\tilde{p}(\boldsymbol{\theta}_{k-1}|\mathcal{Y}_{k-1})},$$

which was computed at the previous time step. Hence, in comparison to equation 4.5, it is clear that the terms in the square brackets of equation 4.6 constitute the exact variational posterior marginal of the state at the previous time instant to give equation 4.4a. Equation 4.4b follows by application of the chain rule on equation 4.1b where the joint $p(\mathcal{X}_{k-1}, \Theta_k^{/\theta_k^i}, \mathcal{Y}_{k-1})$ is constant relative to the distribution of interest.

Theorem 1 does not constitute an online algorithm in the strictest sense since equations 4.4a and 4.4b are evidently coupled, and, as in the offline case, some form of iteration between the solutions is required for convergence. However, iterations are required only between the marginals at the last time instant, making the algorithm fast and efficient, and in practice, few iterations often suffice. It should also be noted that the online algorithm does not necessarily maximize the variational free energy because the restricted VB assumption is an approximation to the correct update rule. Based on this result, one can find the update equations for the variational posteriors of interest.

4.1 Online Update of $\tilde{p}(\mathcal{X}_k)$. By comparing equation 4.1a to equation 3.1a, it is evident that $\tilde{p}(\mathcal{X}_k)$ is updated exactly in the same way as in the offline case, with these two differences:

- The expectations in this case are taken with respect to the parameters at the latest time instant.

- From equation 4.4b, it is evident that only the variational posteriors over the pair (x_k, x_{k-1}) are required to be evaluated at each time step.

The parameter distribution updates require the smoothed distribution of x_{k-1} at each time instant and the cross-covariance between (x_k, x_{k-1}) (see section A.2). The required sufficient statistics are denoted as

$$U_k = I_k^2, \quad G_k = I_k \langle x_{k-1} \rangle, \quad M_k = I_k \langle x_k \rangle, \quad W_k = \langle x_{k-1}^2 \rangle, \quad S_k = \langle x_k x_{k-1} \rangle.$$

4.2 Online Update of $\tilde{p}(\boldsymbol{\theta}_k)$. The variational posteriors can be obtained using similar computations to those for the offline case. The only alteration is the time evolution of the parameters driven by the noise $\mathbf{e}_k$. Following standard practice in signal processing (Wan & Nelson, 2001), $\mathbf{e}_k$ is modeled to have zero mean and slowly varying variance,

$$\left\langle e_k^{i^2} \right\rangle = (\eta^i)^{-1} \sigma^2_{\theta^i_{k-1}}, \quad i = 1, \ldots, d,$$

where the term $\eta^i \in (0, 1], i \in \{\alpha, \rho, \mu, \beta^1, \beta^2, \ldots, \beta^C\}$ is a user-defined forgetting factor. Effectively the prior is no longer fixed (although an additional fixed prior can be introduced); rather, according to the parameter evolution equation, it is a gaussian distribution with the mean of the previous estimate and a precision weighted by η.

4.2.1 Online Update of $\tilde{p}(\alpha_k)$. The joint distribution $\tilde{p}(\rho_k, \alpha_k)$ is first found from equation 4.4b. The conditional distribution may then be obtained from $\tilde{p}(\rho_k, \alpha_k) = \tilde{p}(\rho_k | \alpha_k) \bar{p}(\alpha_k)$. Ignoring terms independent of ρ_k, this is given as

$$\ln \tilde{p}(\rho_k | \alpha_k) = \langle \ln p(\rho_k | \rho_{k-1}) \rangle + \langle \ln p(x_k | x_{k-1}, \rho_k, \alpha_k) \rangle,$$

from which the following expressions are obtained:

$$\begin{aligned} \sigma^2_{\rho_k | \alpha_k} &= \left[\frac{1}{\eta^{\rho^{-1}} \sigma^2_{\rho_{k-1}}} + \frac{W_k}{\sigma_\epsilon^2} \right]^{-1}, \\ \langle \rho_k \rangle_{\tilde{p}(\rho_k | \alpha_k)} &= \sigma^2_{\rho_k | \alpha_k} \left[\frac{S_k}{\sigma_\epsilon^2} + \frac{\hat{\rho}_{k-1}}{\eta^{\rho^{-1}} \sigma^2_{\rho_{k-1}}} - \frac{\alpha_k G_k}{\sigma_\epsilon^2} \right]. \end{aligned} \tag{4.7}$$

The marginal $\tilde{p}(\alpha_k)$ may be found by marginalizing ρ_k from the $\tilde{p}(\rho_k|\alpha_k)\bar{p}(\alpha_k)$. This is given by

$$\sigma^2_{\alpha_k} = \left(\frac{1}{\eta^{\alpha^{-1}}\sigma^2_{\alpha_{k-1}}} + \frac{U_k}{\sigma^2_\epsilon} - \frac{\sigma^2_{\rho_k|\alpha_k}G_k^2}{\sigma^4_\epsilon}\right)^{-1},$$

$$\hat{\alpha}_k = \sigma^2_{\alpha_k}\left(\frac{\hat{\alpha}_{k-1}}{\eta^{\alpha^{-1}}\sigma^2_{\alpha_{k-1}}} + \frac{M_k}{\sigma^2_\epsilon} - \frac{G_k}{\sigma^2_\epsilon}\left[\frac{S_k\sigma^2_{\rho_k|\alpha_k}}{\sigma^2_\epsilon} + \frac{\sigma^2_{\rho_k|\alpha_k}\hat{\rho}_{k-1}}{\eta^{\rho^{-1}}\sigma^2_{\rho_{k-1}}}\right]\right). \tag{4.8}$$

4.2.2 Online Update of $\tilde{p}(\rho_k)$. The statistics over ρ_k are obtained by marginalizing α_k from the joint distribution as

$$\tilde{p}(\rho_k) = \int d\alpha_k \tilde{p}(\rho_k|\alpha_k)\bar{p}(\alpha_k).$$

The variational posterior $\tilde{p}(\rho_k|\alpha_k)$ is computed from equation 4.7, and $\tilde{p}(\alpha_k)$ is known from equation 4.8. The marginalization is straightforward to give the following expressions:

$$\sigma^2_{\rho_k} = \sigma^2_{\rho_k|\alpha_k} + \frac{\sigma^2_{\alpha_k}\sigma^4_{\rho_k|\alpha_k}G_k^2}{\sigma^4_\epsilon},$$

$$\hat{\rho}_k = \sigma^2_{\rho_k|\alpha_k}\left[\frac{S_k}{\sigma^2_\epsilon} + \frac{\hat{\rho}_{k-1}}{\eta^{\rho^{-1}}\sigma^2_{\rho_{k-1}}} - \frac{G_k\hat{\alpha}_k}{\sigma^2_\epsilon}\right].$$

4.2.3 Online Update of $\tilde{p}(\mu_k)$. Following equation 4.4b and ignoring terms independent of μ_k, we have that

$$\begin{aligned}\ln\tilde{p}(\mu_k) &= \langle \ln p(\mu_k|\mu_{k-1})\rangle + \langle \ln p(\mathbf{y}_k|x_k, \mu_k, \boldsymbol{\beta}_k)\rangle,\\ &= -\frac{\langle(\mu_k-\mu_{k-1})^2\rangle}{2\eta^{\mu^{-1}}\sigma^2_{\mu_{k-1}}} + \left\langle \sum_{c=1}^{C} y_k^c\left[\mu_k + \beta_k^c x_k\right] - \exp(\mu_k)\exp(\beta_k^c x_k)\Delta\right\rangle,\end{aligned}$$

where the state evolution density is omitted since it is independent of μ_k. On expanding and approximating around $\hat{\mu}_k$, the following update equations are obtained:

$$\hat{\mu_k} = \hat{\mu}_{k-1} + \eta^{\mu^{-1}}\sigma^2_{\mu_{k-1}}\sum_{c=1}^{C}\left(y_k^c - \Delta\left\langle \exp\left(\beta_k^c x_k\right)\right\rangle\exp(\hat{\mu}_k)\right),$$

$$\sigma^2_{\mu_k} = \left(\eta^{\mu}\sigma^{-2}_{\mu_{k-1}} + \Delta\exp(\hat{\mu_k})\sum_{c=1}^{C}\left\langle \exp\left(\beta_k^c x_k\right)\right\rangle\right)^{-1}.$$

4.2.4 Online Update of $\tilde{p}(\beta_k^c)$. Following the same reasoning as that for updating $\tilde{p}(\mu_k)$ the resulting equations are given as

$$\hat{\beta}_k^c = \hat{\beta}_{k-1}^c + \eta^{\beta^{-1}} \sigma_{\beta_{k-1}^c}^2 \left(y_k^c \langle x_k \rangle - \Delta \langle \exp \mu_k \rangle \frac{d}{d\beta_k^c} \left[\langle \exp x_k \beta_k^c \rangle \right]\big|_{\beta_k^c = \hat{\beta}_k^c} \right),$$

$$\sigma_{\beta_k^c}^2 = \left(\eta^{\beta^c} \sigma_{\beta_{k-1}^c}^{-2} + \Delta \langle \exp \mu_k \rangle \left[\frac{d^2}{d\beta_k^{c^2}} \langle \exp x_k \beta_k^c \rangle \big|_{\beta_k^c = \hat{\beta}_k^c} \right] \right)^{-1}.$$

5 Results

5.1 Multiple Point-Process Outputs Driven by a Shared Underlying State. We first considered the offline inference problem illustrated by Smith and Brown (2003) and Yuan and Niranjan (2010), where outputs from multiple neurons sharing a common hidden state were simulated. We set the number of neurons $C = 20$ and considered the response to a spike input applied every 1 s over a time interval of $T = 10$ s with a sampling rate of 100 Hz. We set $\rho = 0.8, \alpha = 4, \mu = 0$, and β^c to a randomly generated number in the interval [0.9 1.1].

All priors on the parameters and states, except for that over β^c, were set to normal distributions with variances: $\sigma_{\rho_p}^2 = 5, \sigma_{\alpha_p}^2 = 50, \sigma_{\mu_p}^2 = 1$. The prior over β^c was set to a normal distribution centered at 1 with a 99% confidence between 0.7 and 1.3; this was done to remedy the identifiability issues stemming from the fact that the likelihood, equation 2.3, involves only the product $\beta^c x_k$ (a problem related to the parameter offsetting observed by Smith & Brown, 2003).

The estimation of the state variational posterior describing the latent process using the VBEM algorithm can be seen in Figure 1, where at each time step, the variational posterior's mean and 99% confidence limits are given. Graphical results for the corresponding estimation of the 23 unknown parameters are in Figure 2, showing rapid convergence to good estimates.

We further compared our results to those obtained using EM (Smith & Brown, 2003) and those given by a Gibbs sampler on the same data set (see appendix C). To avoid identifiability issues, we also ran experiments with $\boldsymbol{\beta}$ fixed to its true value. Table 1 shows that all methods are effective in estimating parameters for these data, with Gibbs and VBEM also providing confidence intervals that are in good agreement with the true values. It took 5 s for the EM algorithm (50 iterations), 12 s for the VBEM algorithm (50 iterations), and 279 s for the Gibbs sampler (5000 iterations) to converge.[1]

A more informative test of the model's performance is its ability to capture the spike train distribution. A quantitative measure of this can

[1]Simulations carried out on an Intel Core 2 Quad Q6600 @ 2.40 GHZ with 4 GB of RAM.

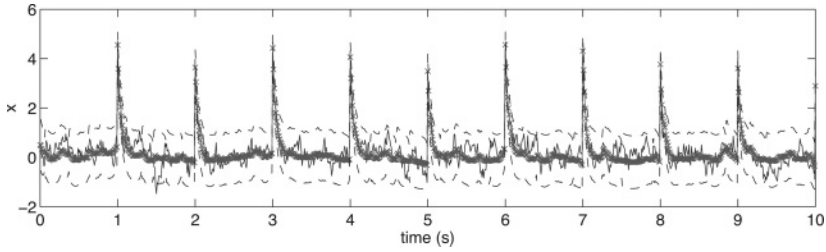

Figure 1: True state (continuous unmarked line) and mean estimated state (marked line) as given by the batch VB algorithm in the final iteration. The true state lies consistently within the 99% confidence intervals (dashed line).

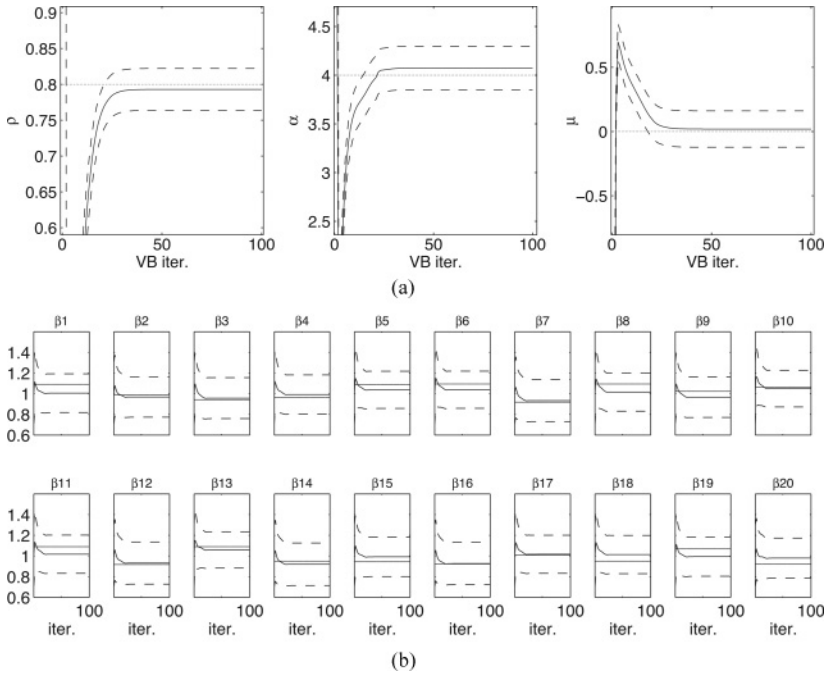

Figure 2: Mean estimates (continuous varying line) and 99% confidence intervals (dashed line) over 100 VBEM iterations for the parameters (a) ρ, α, μ and (b) $\beta^c, c = 1, \ldots, 20$ using the batch VB algorithm. The parameters converge in distribution to reasonable estimates regardless of the initial conditions, and the true (solid level line) values are seen to lie well within the 99% confidence intervals at steady state.

Table 1: Parameter Estimation by the EM Algorithm, Gibbs Sampler, and VBEM Algorithm.

θ	True	EM	Gibbs	VBEM	VBEM (free $\boldsymbol{\beta}$)
ρ	0.80	0.82	0.79 ± 0.06	0.79 ± 0.03	0.79 ± 0.03
α	4.00	4.08	3.81 ± 0.48	4.04 ± 0.22	4.07 ± 0.22
μ	0.00	-0.19	0.06 ± 0.24	0.01 ± 0.14	0.02 ± 0.14
avr($\boldsymbol{\beta}$)	1.00	-	-	-	0.99 ± 0.19

Note: Unless stated, $\boldsymbol{\beta}$ was fixed to the true value during simulation.

Table 2: Mean Squared Maximum KS Distances for the 20 Neurons with Different Event-Rate Models (Lower Is Better) for One Data Set.

	Gibbs	VBEM	EM	VBEM (free $\boldsymbol{\beta}$)	EM (free $\boldsymbol{\beta}$)	SW
MSE	0.0046	0.0058	0.0076	0.0077	0.0136	0.0336

Note: Unless stated, $\boldsymbol{\beta}$ was fixed to the true value during simulation.

be achieved using the time-rescaling theorem of Brown, Barbieri, Ventura, Kass, and Frank (2002) in conjunction with a Kolmogorov-Smirnov (KS) test, following the same procedure as Smith and Brown (2003) and Barbieri, Matten, Alabi, and Brown (2005). As a goodness-of-fit measure, the mean squared maximum distance between the model rate and the true rate over all output channels was found. The results for this KS measure on a synthetic data set are given in Table 2. For completeness we also compare with a sliding window (SW) empirical rate estimator of 100 ms width, which is often used in these applications (Riehle, Grün, Diesmann, & Aertsen, 1997). The Bayesian methods (VBEM and Gibbs sampler) obtain a considerably better goodness of fit than the EM algorithm (which, in turn, is much better than the simple SW heuristic), indicating that retaining distributional information over the parameters leads to an improvement in the modeling of the spike distribution. To further validate the result, we ran a two-sample t-test on the KS measures from 20 different data sets. The mean-square maximum KS distance for all these runs was 0.0070 for the VBEM algorithm (fixed $\boldsymbol{\beta}$) and 0.0089 for the EM algorithm (also with fixed $\boldsymbol{\beta}$). The test rejected the null hypothesis that the decrease in error occurred by chance at the 5% significance level.

5.2 Online Parameter Tracking from Point-Process Observations. In this section we present a simulation study of the VB online algorithm derived in section 4. The nature of the data typical in these types of models requires some further intervention for correct estimation when using filters. In regions where no input is present, the observed events in the output are

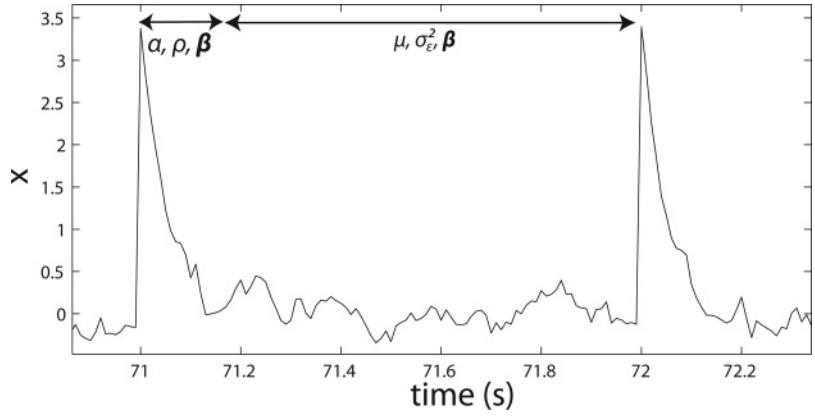

Figure 3: Selective updating of parameter estimates in an online framework is carried out in accordance with the areas where the state bears most information about the relevant parameters of interest. In this case, the narrow stretch close to an input spike bears a lot of information on the state decay factor ρ and the input gain α. The noise parameters μ and σ_ϵ^2 are more evident in regions of no input.

predominantly due to the background firing rate μ and state noise σ_ϵ^2, and there is little or no information about ρ and α in these regions. On the other hand, the deterministic component of the hidden state governs the output in time intervals close to an input. In these areas, there is significant information about ρ and α. Parameter distributions were thus updated only in regions where there is ample information about the relevant parameters, as illustrated in Figure 3. This procedure is standard in online filtering in other areas, such as speech enhancement by spectral subtraction, in which noise levels are estimated in regions of the signal where speech is not present (Boll, 1979).

For this study, we assumed $C = 20$ and that $\boldsymbol{\beta}$ and μ were predetermined from a previous offline analysis and assumed to be constant. The choice of the forgetting factors was carried out by trial and error such that a parameter change could be tracked without compromising stability in the online estimates. We subsequently chose $\eta^\rho = 0.8$ and $\eta^\alpha = 0.9$. The dual VB filter was compared to a standard particle filter (PF), which makes use of an augmented state vector $\mathbf{z}_k = [x_k, \rho_k, \alpha_k]^T$ and implements what is effectively a standard sequential importance sampling with resampling (SISR) algorithm (see Kitagawa, 1998; Doucet, de Freitas, & Gordon, 2001; de Freitas, Niranjan, Gee, & Doucet, 2000). The prior distribution was chosen as the importance distribution so that the weights were updated in time according to the likelihood. That is, if $w_k^{(i)}$ denotes the weight of the ith

particle at time k and $\mathbf{z}_k^{(i)}$ the ith particle at time k, the weight update is given as

$$w_k^{(i)} \propto w_{k-1}^{(i)} p\left(\mathbf{y}_k | \mathbf{z}_k^{(i)}\right).$$

The selective estimation process described was adapted to the PF by using selective SISR, as shown in Figure 4. In this figure, the case where only the input gain α_k and the state x_k are to be estimated at one time instant is shown. In regions where α_k does not affect the likelihood (or *importance factor*), propagation and subsequent resampling take place only in the state-space. The respective parameter marginal distribution is retained and propagated through time unchanged. Formally, after resampling, in this region, we have that the full joint distribution is given by

$$p(\alpha_k, x_k | \mathcal{Y}_k) \approx \frac{1}{N} \sum_{i=1}^{N} \delta \begin{pmatrix} x_k - x_k^{(i)} \\ \alpha_k - \alpha_{k-1}^{(i)} \end{pmatrix}$$

and the subsequent marginal distribution by

$$p(\alpha_k | \mathcal{Y}_k) = \int dx_k \, p(\alpha_k, x_k | \mathcal{Y}_k) \approx \frac{1}{N} \sum_{i=1}^{N} \delta\left(\alpha_k - \alpha_{k-1}^{(i)}\right) \approx p(\alpha_{k-1} | \mathcal{Y}_{k-1}),$$

where N denotes the number of particles and $\delta(\cdot)$ the delta Dirac mass. The result for the successful tracking of a sudden change in the true value of ρ from 0.8 to 0.6 by both the VB filter, and the PF with $N = 5000$ particles, is shown in Figure 5 (the number of particles chosen was the minimum required for consistent posterior distribution approximations across several trials). The results corroborate each other, indicating that the VB filter gives a realistic description of what can be termed the ground truth.[2] Complete results are shown in Table 3. Despite the parameter distributions estimated being very similar, the PF took on the order of 10 times longer than the VB filter to execute. Indeed, the computational time required by the PF in this example was more than the duration of the data stream itself, rendering it impractical for the real-time application of this case study scenario.

5.3 Online Characterization of Taste Stimuli. As an example application of our online algorithm on real data, we modeled spiking patterns of

[2]Since filters are particularly sensitive to the chosen parameter evolution model, the online parameter posterior distribution is highly dependent on the forgetting factor in the VB filter and the corresponding parameter noise statistics in the PF. In the latter case, the variance was tuned to give a similar learning rate as that of the VB filter.

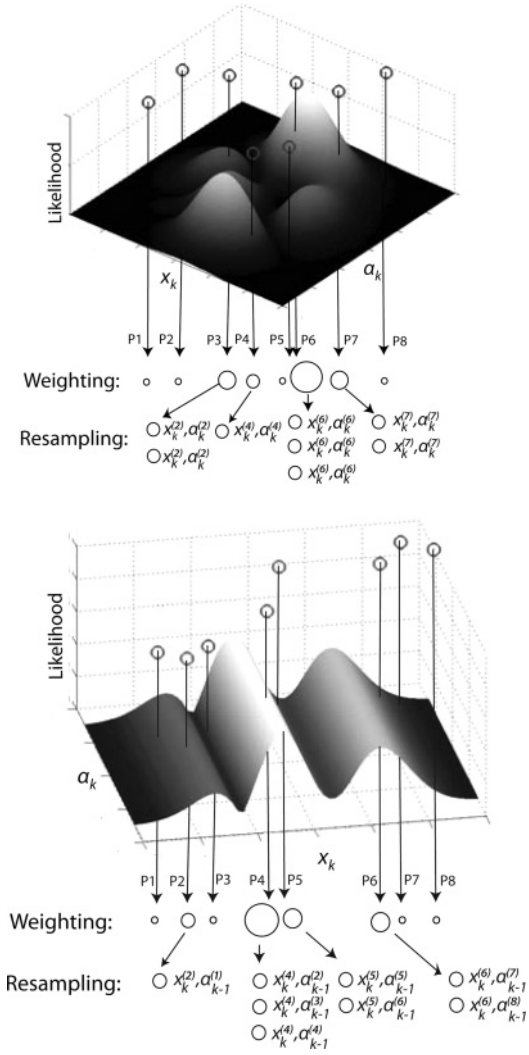

Figure 4: (Top) The likelihood function is used to appropriately weight the particles (P#) representing the posterior distribution, which are then resampled into N particles of equal weight. (Bottom) In this case, the likelihood is practically independent of α_k, and thus the weighing and resampling steps depend solely on the x_k component of the particles. In order to maintain the posterior distribution with fewer particles than would be necessary otherwise, the prior particle parameter set is redistributed after resampling, with equal weight among the resampled particles. The figures (top) and (bottom) correspond to the two areas marked in Figure 3, respectively (likelihood surfaces shown are for illustration only and do not represent actual surfaces).

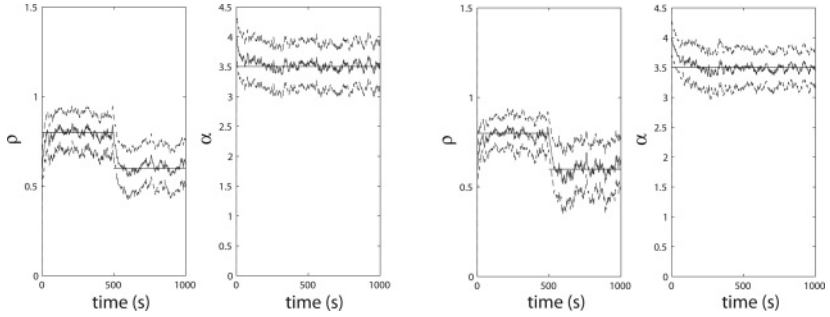

Figure 5: Online tracking of a sudden change in the true parameter (level black line) ρ at time $t = 500$ s. In this example, μ and $\boldsymbol{\beta}$ were assumed constant and known from previous offline analysis of the system. The 99% confidence intervals (outer traces) are seen to enclose the true value on the filter reaching a steady behavior for both the (left) VB filter and (right) particle filter with 5000 particles.

Table 3: Comparison Between the VB Filter and a PF for SSPP with 5000 Particles.

	ρ	α	$mean(\hat{\rho_k})$	$mean(\sigma_{\rho_k})$	$mean(\hat{\alpha_k})$	$mean(\sigma_{\alpha_k})$
VB ($t \leq 500$)	0.8	3.5	0.799	0.037	3.52	0.13
PF ($t \leq 500$)			0.797	0.031	3.49	0.12
VB ($t > 500$)	0.6	3.5	0.607	0.041	3.51	0.12
PF ($t > 500$)			0.602	0.049	3.49	0.10

taste response cells in the nucleus tractus solitarii (NTS) of Sprague-Dawley rats following the application of different taste stimuli (di Lorenzo & Victor, 2003). The attraction of the online approach is that it provides a method for stimulus chemical discrimination by tracking changes in underlying parameters on the presentation of different stimuli. The experimental data were obtained from trials where different compounds dissolved in distilled water were delivered to the oropharyngeal area. Taste-evoked spike train data used in this study were delivered via neurodatabase.org, a neuroinformatics resource funded by the Human Brain Project.

Although the SSPP was primarily developed for implicit stimuli, it provides a neat way of parameterizing a dynamic CIF to model variable-rate neural responses to explicit stimuli. Such is the case considered here, where ample evidence suggests that for some of the cells in the NTS, rate coding is used for interstimulus discrimination (Roussin, Victor, Chen, & di Lorenzo,

2008).[3] Some of these are so finely tuned to different stimuli that one can use spike count alone to discriminate between different tastes (e.g., cell 9 in the study). Others are not so finely tuned, and spike count cannot be used to discriminate between the tastants (e.g., cell 11). Nonetheless, spike count gives no information on the time-varying event rate (or the rate envelope) itself. Moreover, many alternatives (such as the conventional sliding window) do not provide a plausible model for the underlying neural dynamics. The SSPP applied to these cells can give not only the descriptive powers required for taste discrimination but also additional information that may be of physiological use. Here we also show how the VB filter can infer the varying SSPP parameters governing the underlying dynamics, which for the same neuron appear to vary in a structured manner with the application of different stimuli.

Each experimental trial consisted of three phases: (1) a 10 s baseline period in the absence of any stimulus, (2) 5 s of stimulus presentation, and (3) a 5 s wait. Each trial was separated by rinsing and a 1.5 min wait. The data used in the analysis were those recorded in the second and third phases (10 s segments) in which the neural response to the four tastants used—NaCl, HCl, quinine, and sucrose (each of which represents a different taste quality; salty, sour, bitter and sweet, respectively)—is present. The learning data set was formed by first grouping the 10 s segments according to stimulus and then concatenating them into four sets (one per stimulus). Combinations of these spike trains were then joined together to form the data sets on which learning was carried out.

Data were gathered at a resolution of 1 ms, and we hence initially organized the spikes into bins of $\Delta = 1$ ms such that the condition shown in equation 2.1 was satisfied. However, we then increased the bin size to $\Delta = 10$ ms to speed up the algorithm. This resulted in some bins (less than 5%) containing more than one output spike (e.g., for cell 9, maximum HCl with 3.4% and minimum sucrose with 1.5%), which were subsequently repositioned to the closest empty bin in forward time. Preanalysis of the data was carried out by studying the poststimulus histograms (PSTHs) of the responses to the four stimuli. These histograms suggested an approximate linear increase in firing rate for the first 250 ms and also a response latency that was not considered in the simulation study. To cater for these effects, we treated the input signal as a pulse of width 250 ms.

It was evident from preliminary studies that the dominant rate coding characteristics that differed across tastants were attributed to the input gain α and the background firing rate μ. We thus chose to monitor these two

[3] As opposed to temporal coding, where the specific arrangement in time is of particular relevance to discrimination and deemed to play an important role, particularly in the initial (phasic) phase of the response.

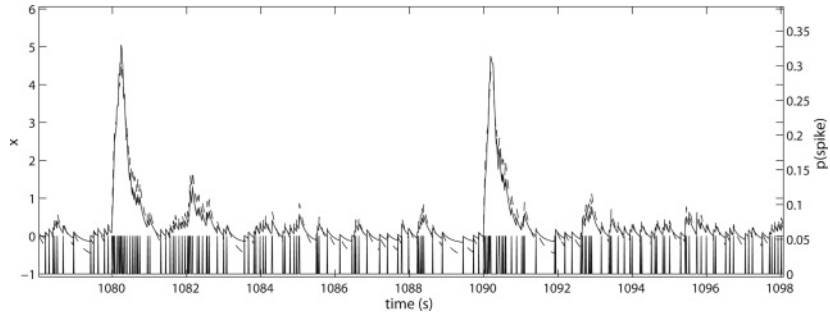

Figure 6: A 20 s segment of training data taken from the cell 9 response to NaCl. The time duration shown spans across two trials with the rinsing and phase 1 periods removed. The estimated state (dashed line) and probability of a spike occurring (solid line) are seen to be indicative of the frequency of spike events (shown on the bottom axis).

parameters online (in addition to the underlying state) in order to study the response behavior while discriminating among the tastants in real time. With the use of offline methods, we chose to fix the unknown parameters $\beta = 0.5$, $\sigma_\epsilon^2 = 0.05$, and $\rho = 0.97$, which was representative of all tastants. The dual VB filter was, however, found to be robust and resistant to changes in state noise and fixed parameter estimates. The relevant forgetting factors were set to $\eta^\mu = 0.999$ and $\eta^\alpha = 0.9$, respectively.

As discussed in section 5.2, to ensure identifiability, the online parameter updates were carried out only in the regions where ample information is present, so that α was updated only in regions of input application and μ in regions between the application of the respective inputs. We show a representative filtered state and output probability of a spike occurring for the tastant NaCl in cell 9 in Figure 6. Note how the firing probability adequately captures the behavior of the spike train.

The changes in both α and μ were very evident across the different experiments. In some cases, monitoring μ is sufficient to characterize the difference in response to different tastants (see Figure 7 for a comparison of sucrose with HCl in cell 9). However, this is not the general case, as shown by the trajectories of the mean parameter estimates of α and μ in Figures 8 and 9. For instance, while μ seems to vary across tastants in cell 9 (see Figure 8), the background firing rates in response to NaCl and HCl for cell 11 are fairly similar (see Figure 9). It is the input gain α that is different between these two responses. When the parameters μ and α are monitored, the responses cluster in distinct and separate regions characteristic of the stimulus being applied.

It is also interesting to note that except for sucrose, neither response can be considered to be passive, that is, with both a low α and a low μ.

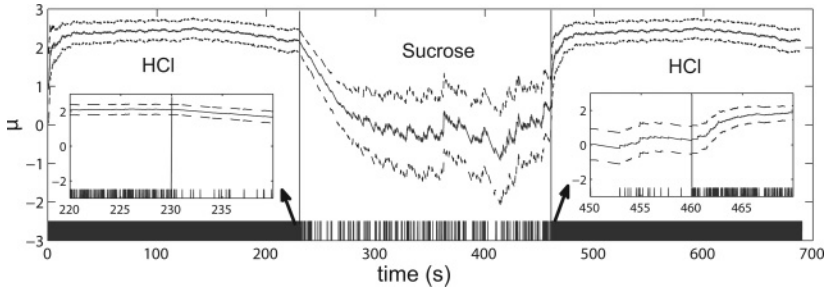

Figure 7: Tracking the mean (solid) and corresponding 99% intervals (dashed) of μ indicating a change in stimulus from HCl to sucrose and back to HCl in cell 9. The parameter change is indicative of a change in the spike train pattern (inset) when the stimulus is changed. The solid vertical lines indicate where the change in applied chemical stimulus took place. For this trial, α was fixed to 0.1.

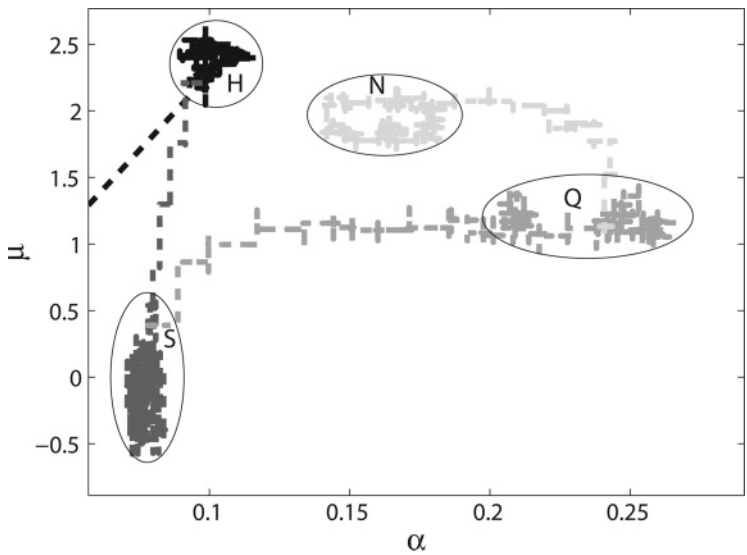

Figure 8: Cell 9—temporal progression of the estimated mean of α and μ indicating a change of stimulus from (in order of decreasing contrast) HCl (H) to sucrose (S) to quinine (Q) to NaCl (N). Although the cell is, overall, less responsive (μ) to quinine, the immediate effect of its application (α) is relatively more substantial than in the case of both HCl and NaCl. The ellipses define arbitrarily chosen classification boundaries.

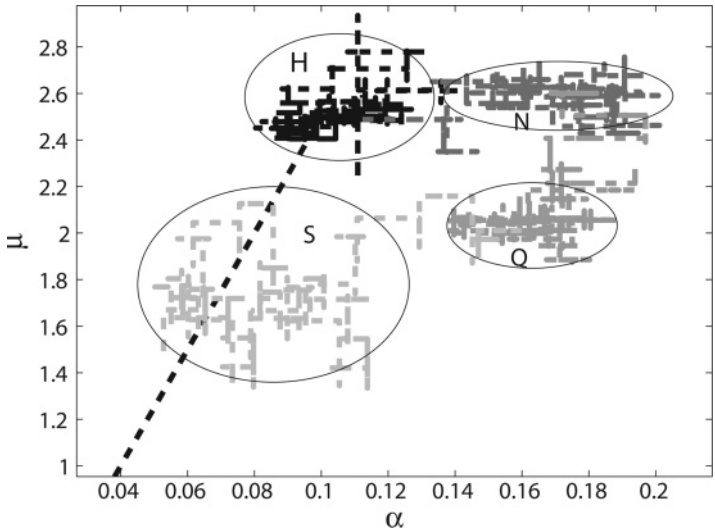

Figure 9: Cell 11—temporal progression of the estimated mean of α and μ indicating a change of stimulus from (in order of decreasing contrast) HCl (H) to NaCl (N) to quinine (Q) to sucrose (S). From this chart, it is evident that α or μ on their own cannot capture the difference in response to the different tastants. The ellipses define arbitrarily chosen classification boundaries.

The responses exhibit prominent activity in either the initial stage or the steady-state stage (the phasic and tonic stages respectively), or both. The considerable activity in the initial stage even when the overall reponse μ is low (particularly with quinine), is also somewhat of a testimony to the hypothesis that the initial neural response to every tastant may contain some additional information, encoding for instance a measure of taste acceptance (known as the hedonic value; see di Lorenzo & Victor, 2003).

Finally, we conclude by showing how the online algorithm also manages to accurately give a rate envelope over the responses as indicated by the multiple overlays on the PSTHs in Figure 10. The VB filter manages to approximate the PSTH in each trial, validating the appropriateness of this model for characterization of the rate-encoding properties of this neuron.

6 Discussion

In this letter, we have proposed a variational Bayesian method for filtering and smoothing within state-space models with point-process observations. This class of models provides a physiologically plausible signal processing framework for event-based observations and has proved a popular

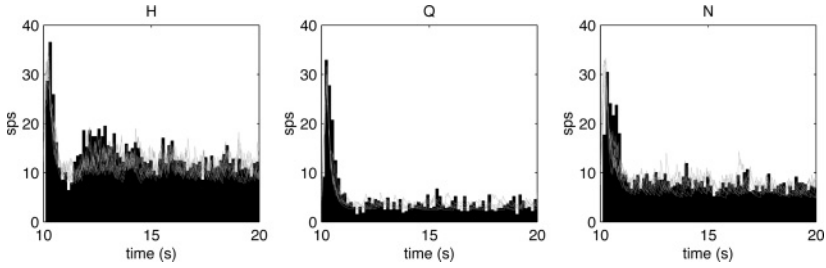

Figure 10: The estimated firing rate in spikes per second (sps) from five randomly selected trials (gray) in the online data, overlaying the PSTH (black) of responses to the respective stimuli, H (HCl), Q (quinine) and N (NaCl) in cell 9. The approximate firing rate is computed as $p(y_k = 1|x_k, \boldsymbol{\theta}_k)/\Delta$.

framework for analyzing and decoding spike train data. Experiments on realistic simulated data show that the Bayesian treatment (by either VB or computationally expensive sampling methods) does indeed lead to an improvement in the modeling of the spike train distribution while retaining very good accuracy in estimating the parameter posteriors. A major contribution of this work is the introduction of an online estimation framework. This allows considerable computational savings, potentially paving the way for real-time biomedical applications. It also allows the monitoring of online changes in a system mode of operation, as exemplified in our study of neural responses to different taste stimuli.

Filtering of doubly stochastic point-process may be carried out directly in continuous time (Snyder & Miller, 1991), in which case the stochastic intensity is generally assumed to be a function of a diffusion (Segall, Davis, & Kailath, 1975; Solo, 2000). Solutions are given as normalized or unnormalized conditional intensities that take the form of partial differential equations. Analytical solutions can be found in special cases, such as when the intensity is given as the square of an Ornstein-Uhlenbeck process (Boel & Benes, 1980). Nonetheless, in the general case, computationally expensive numerical methods are still required for implementation. The case is similar in discrete time. Manton, Krishnamurthy, and Elliott (1999), for instance, showed that an exact (strictly) finite-dimensional filter exists for equation 2.4 with $\lambda_k^c = (\gamma_k x_k)^2$, $\gamma_k > 0 . \forall k$, but the treatment quickly becomes intractable for different forms of the intensity. This work, and most of the literature that focuses on state estimation from point-process observations, uses models where the parameters are assumed to be known. This motivates investigation into new, more versatile methods such as that first proposed by Smith and Brown (2003), now extended into a variational setting in this letter.

VB provides a neat, deterministic way for approximating the joint posterior distribution online. We have compared the performance of the VB

filter to a stochastic approximation method through a standard PF and seen that it performs very well comparatively, with a marked decrease in computational requirements. Previous to this work, sequential Monte Carlo (SMC) methods had been applied to the state estimation problem in the SSPP framework. Ergün et al. (2007) modeled the underlying state dynamics by a random walk process, but the underlying parameters were assumed to be known. The authors introduced point-process adaptive filters (Eden, Frank, Barbieri, Solo, & Brown, 2004) for proposing new particles to increase computational efficiency. The method showed good performance on both a synthetic and a real data set, where the problem of tracking the evolution of a hippocampal spatial receptive field was studied. The extension of these results to online parameter learning SMC approaches (see also Storvik, 2002) was thus a natural step. It should be noted that the highly linear substructure (through the underlying AR latent process) also allows for Rao-Blackwellized PFs (Doucet, de Freitas, Murphy, & Russell, 2000) to be applied. In this case, the state forward filtering step may by approximated by that of Smith and Brown (2003) or Fahrmeir and Tutz (1994). However, preliminary results show that even in this case, SMC methods may still prove to be too time-consuming for any interesting biomedical application where data need to be handled in real time.

Online variational Bayes was first proposed for model selection of static conjugate-exponential (CE) models by Sato (2001), where the recursive updates at each time step describe the solution to successive maximizations of a discounted free energy. Unlike the online VB algorithm presented here, Sato's approach has the advantage that the algorithm behaves as a stochastic approximator for the maximum expected free energy for a fixed number of data points, obviating the requirement of VB iterations at each datum. However, Sato's algorithm relies on the favorable properties of the family of static CE models that SSPP clearly do not form part of. Moreover, it is envisioned that the algorithm proposed in this work will find potential in application to a continuous stream of data, where the maximization of a fixed objective functional loses its appeal. In the proposed solution, we have made use of a static forgetting factor to discount the use of "old" information in the estimation process. This bears similarity to the time-varying discount factor for variable learning rate as used in Sato's work.

The application to online tracking suggests naturally an extension to consider state-space models with switching parameters, which would formally incorporate abrupt changes in mode of operation into the model. These have proved a popular tool in biomedical applications (see, e.g., Quinn, Williams, & McIntosh, 2008), and would also be suitable for the application described in section 5.3. This additional complexity is likely, however, to come at some computational cost. A further interesting extension would be to improve on the observation model by using more

advanced models for spike generation such as integrate and fire; parameter estimation within these models has recently been explored using search-type algorithms (MacGregor, Williams, & Lang, 2009), but the complexity of the likelihood model means that it is likely that considerable work will be needed before they can be used in signal processing applications.

Appendix A: Derivation of the Update Equations for $\tilde{p}(\mathcal{X}_K)$

A.1 The Forward Pass. Initialize $x_{0|0}$ and set $\sigma^2_{0|0} = \kappa$ where κ is indicative of the uncertainty on the initial state. The forward pass is given by Beal (2003),

$$p(x_k|\mathcal{Y}_k) \propto \int dx_{k-1}\, p(x_{k-1}|\mathcal{Y}_{k-1}) \exp\langle \ln p(x_k|x_{k-1}, \boldsymbol{\theta})\, p(\mathbf{y}_k|x_k, \boldsymbol{\theta})\rangle,$$

where $p(x_k|x_{k-1}, \boldsymbol{\theta}) = \mathcal{N}_{x_k}(\rho x_{k-1} + \alpha I_k, \sigma^2_\epsilon)$ and $p(x_{k-1}|\mathcal{Y}_{k-1}) = \mathcal{N}_{x_{k-1}}(x_{k-1|k-1}, \sigma^2_{k-1|k-1})$. The product $p(x_{k-1}|\mathcal{Y}_{k-1}) \exp\langle \ln p(x_k|x_{k-1}, \boldsymbol{\theta})\rangle$ is normal in x_{k-1} with precision $\overline{\sigma}^{-2}_{k-1} = \sigma^{-2}_{k-1} + \langle \rho^2\rangle \sigma^{-2}_\epsilon$ and mean

$$\overline{x}_{k-1} = \overline{\sigma}^2_{k-1}(x_{k-1|k-1}\sigma^{-2}_{k-1|k-1} + \langle \rho\rangle x_k \sigma^{-2}_\epsilon - \langle \rho\alpha\rangle I_k \sigma^{-2}_\epsilon).$$

Marginalizing out x_{k-1}, we get

$$p(x_k|\mathcal{Y}_k) \propto \mathcal{N}_{x_k}(\tilde{x_k}, \tilde{\sigma}^2_k) \exp(\langle \ln p(\mathbf{y}_k|x_k, \boldsymbol{\theta})\rangle),$$

where $\tilde{\sigma}^{-2}_k = \sigma^{-2}_\epsilon - \langle \rho\rangle^2 \overline{\sigma}^2_{k-1}\sigma^{-4}_\epsilon$, and

$$\tilde{x_k} = \tilde{\sigma}^2_k\big(\overline{\sigma}^2_{k-1}\langle \rho\rangle \sigma^{-2}_\epsilon [x_{k-1|k-1}\sigma^{-2}_{k-1|k-1} - \langle \rho\alpha\rangle I_k \sigma^{-2}_\epsilon] + \langle \alpha\rangle I_k \sigma^{-2}_\epsilon\big).$$

Since the observation equation is nonlinear, we choose to approximate the product of the distributions to a gaussian with Laplace's method so that

$$\mathcal{N}_{x_k}\big(\tilde{x_k}, \tilde{\sigma}^2_k\big) \exp(\langle \ln p(\mathbf{y}_k|x_k, \boldsymbol{\theta})\rangle) \approx \mathcal{N}_{x_k}\big(x_{k|k}, \sigma^2_{k|k}\big),$$

where, we recall,

$$p(\mathbf{y}_k|x_k, \boldsymbol{\theta}) = \prod_{c=1}^{C} \Delta \exp(\mu + \beta^c x_k)^{y^c_k} \exp(-\exp(\mu + \beta^c x_k)\Delta).$$

As shown in the main text, a nonlinear optimizer is needed to evaluate $x_{k|k}$.

A.2 The Backward Pass. Initialize with $\sigma^{*^2} = \kappa$ where κ is large and $x_K^* = x_{K|K}$ if carried out after the forward pass (see below). The backward pass is given by the recursion as in Beal (2003),

$$p(\mathbf{y}_{k+1:K}|x_k) = \int dx_{k+1}\, p(\mathbf{y}_{k+2:K}|x_{k+1}) \times \exp\langle \ln p(x_{k+1}|x_k, \boldsymbol{\theta}) p(\mathbf{y}_{k+1}|x_{k+1}, \boldsymbol{\theta})\rangle,$$

where $p(x_{k+1}|x_k,\ \boldsymbol{\theta}) = \mathcal{N}_{x_{k+1}}(\rho x_k + \alpha I_{k+1},\ \sigma_\epsilon^2)$ and $p(\mathbf{y}_{k+2:K}|x_{k+1}) = \mathcal{N}_{x_{k+1}}(x_{k+1}^*,\ \sigma_{k+1}^{*^2})$. We find $p(\mathbf{y}_{k+2:K}|x_{k+1}) \exp(\langle \ln p(\mathbf{y}_{k+1}|x_{k+1},\ \boldsymbol{\theta})\rangle) \approx \mathcal{N}_{x_{k+1}}(x'_{k+1},\ \sigma'^2_{k+1})$ by taking the quadratic Taylor expansion around an arbitrary $\hat{x}_{k+1}$ to obtain the expressions

$$x'_{k+1} = \hat{x}_{k+1} + \sigma'^2_{k+1}\left(\frac{x^*_{k+1} - \hat{x}_{k+1}}{\sigma^{*^2}_{k+1}} + \sum_{c=1}^{C}\left\{\langle \beta^c\rangle_{\bar{p}(\beta^c)} y^c_{k+1} - \Delta\langle \exp\mu\rangle_{\bar{p}(\mu)} \frac{d}{dx_{k+1}}\left[\langle \exp x_{k+1}\beta^c\rangle_{\bar{p}(\beta^c)}\right]\Big|_{x_{k+1}=\hat{x}_{k+1}}\right\}\right),$$

$$\sigma'^2_{k+1} = \left(\sigma^{*^{-2}}_{k+1} + \sum_{c=1}^{C}\left\{\Delta\langle \exp\mu\rangle_{\bar{p}(\mu)} \frac{d^2}{dx^2_{k+1}}[\langle \exp x_{k+1}\beta^c\rangle_{\bar{p}(\beta^c)}]|_{x_{k+1}=\hat{x}_{k+1}}\right\}\right)^{-1}.$$

The choice of $\hat{x}_{k+1}$ bears a lot of weight on the performance of the algorithm. One can set $\hat{x}_{k+1} = x'_{k+1}$, resulting in a nonlinear optimization problem. Or, one can linearize around the filtered estimate $x_{k+1|k+1}$ instead, and this is what is done in the main text. The advantage is that no nonlinear optimization is required to compute the backward pass; the drawback is that the backward pass can no longer be carried out in parallel with the forward pass.

The next step is to find the product of this approximate distribution with $\exp(\langle \ln p(x_{k+1}|x_k, \boldsymbol{\theta})\rangle)$, which is easily shown to be proportional to

$$\exp\big(-x^2_{k+1}(\sigma_\epsilon^{-2} + \sigma'^{-2}_{k+1})/2 + x_{k+1}\big[\langle\rho\rangle x_k\sigma_\epsilon^{-2} + \langle\alpha\rangle I_{k+1}\sigma_\epsilon^{-2} + x'_{k+1}\sigma'^{-2}_{k+1}\big] - \langle\rho^2\rangle x_k^2\sigma_\epsilon^{-2}/2 - \langle\rho\alpha\rangle x_k I_{k+1}\sigma_\epsilon^{-2}\big).$$

The required normal distribution in x_k with mean x_k^* and variance $\sigma_k^{*^2}$ is found by marginalizing out x_{k+1}. The smoothed estimate is computed by considering the product distribution of the forward pass and the backward pass. In particular, we find that

$$p(x_k|\mathcal{Y}_K) \propto p(x_k|\mathcal{Y}_k)\, p(\mathbf{y}_{k+1:K}|x_k).$$

Since this is a product of gaussian distributions, the state estimate conditioned on all the data can be found and can be readily computed in the backward pass if this is carried out sequential to the forward pass. The results are shown in the main text. The pairwise marginals are given as

$$p\,(x_k, x_{k+1}|\mathcal{Y}_K) \propto p(x_k|\mathcal{Y}_K)\,p(\mathbf{y}_{k+2:K}|x_{k+1}) \times \exp(\langle \ln p(x_{k+1}|x_k, \boldsymbol{\theta}) \quad p(\mathbf{y}_{k+1}|x_{k+1}, \boldsymbol{\theta})\rangle).$$

We expand the logarithm of this quantity and approximate it to a multivariate normal distribution about the smoothed state estimate. The required second moment is then found by adding the product of the smoothed pair to the cross-covariance. The result is shown in the main text.

Appendix B: Derivation of the Update Equations for $\tilde{p}(\mu)$ and $\tilde{p}(\beta)$

B.1 Batch Update of $\tilde{p}(\mu)$. The variational posterior over μ, ignoring terms independent of μ, is given by

$$\ln \tilde{p}(\mu) = \ln p(\mu) + \left\langle \sum_{c=1}^{C} \sum_{i=1}^{K} y_i^c [\mu + \beta^c x_i] - \exp(\mu) \exp(\beta^c x_i) \Delta \right\rangle,$$

where $p(\mu)$ is the prior over μ with mean μ_p and variance σ_p^2. We restrict the variational posterior to be gaussian with mean $\hat{\mu}$ and variance σ_μ^2. By application of the standard Laplace method, we obtain the expressions given in the main text. In these expressions it is required to evaluate the quantity $\langle \exp(x_i \beta^c) \rangle$. From moment-generating functions, we know that

$$\int \exp(x_i \beta^c) \mathcal{N}_{x_i}(x_{i|K}, \sigma_{i|K}^2) dx_i = \exp(x_{i|K} \beta^c + \sigma_{i|K}^2 \beta^{c^2}/2).$$

However, we are concerned with the quantity

$$\begin{aligned} \langle \exp(x_i \beta^c) \rangle &= \int dx_i \left[\int d\beta^c \mathcal{N}_{\beta^c}(\hat{\beta}^c, \sigma_{\beta^c}^2) \right] \mathcal{N}_{x_i}(x_{i|K}, \sigma_{i|K}^2) \\ &= \int dx_i \exp(\hat{\beta}^c x_i + \sigma_{\beta^c}^2 x_i^2/2) \mathcal{N}_{x_i}(x_{i|K}, \sigma_{i|K}^2) \\ &= \frac{1}{\sqrt{2\pi\sigma_{i|K}^2}} \int dx_i \exp(\hat{\beta}^c x_i + \sigma_{\beta^c}^2 x_i^2/2 - (x_i - x_{i|K})^2/2\sigma_{i|K}^2). \end{aligned}$$

After marginalizing out x_i and some algebraic manipulation, the final result is

$$\langle \exp(\beta^c x_i)\rangle_{\tilde{p}(\mathcal{X}_K)\tilde{p}(\beta^c)} = \sqrt{\frac{1}{1-\sigma^2_{\beta^c}\sigma^2_{i|K}}}\exp\left(\frac{x^2_{i|K}\sigma^2_{\beta^c}+\hat{\beta^c}^2\sigma^2_{i|K}+2\hat{\beta^c}x_{i|K}}{2(1-\sigma^2_{\beta^c}\sigma^2_{i|K})}\right).$$

B.2 Batch Update of $\tilde{p}(\beta^c)$. The variational posterior over β^c, ignoring terms independent of β^c, is given by

$$\ln\tilde{p}(\beta^c) = \ln p(\beta^c) + \left\langle \sum_{i=1}^{K} y_i^c[\mu+\beta^c x_i] - \exp(\mu)\exp(\beta^c x_i)\Delta \right\rangle,$$

where $p(\beta^c)$ denotes the prior over β^c. Effecting the required derivatives, we once again restrict the variational posterior to be gaussian with mean and variance as given in the main text. The expectations required in this case are those of log-normal distributions, which are easy to compute. In particular, we have

$$\langle \exp(\beta^c x_i)\rangle_{\tilde{p}(\mathcal{X}_K)} = \exp(\beta^c x_{i|K} + \beta^{c^2}\sigma^2_{i|K}/2)$$

and $\langle \exp(\mu)\rangle_{\tilde{p}(\mu)} = \exp(\hat{\mu}+\sigma^2_\mu/2)$.

Appendix C: Gibbs Sampler for SSPP

Unlike with variational methods, in a full Bayesian treatment for the SSPP, we attempt to find the full joint distribution of the underlying states and parameters given the observations ($p(\mathcal{X}_K, \boldsymbol{\theta}|\mathcal{Y}_K)$) by employing Markov chain Monte Carlo (MCMC) approximation methods. The Gibbs sampler is a standard technique used when approaching this problem (Carter and Kohn, 1994; Geweke & Tanizaki, 2001) by iteratively drawing samples from two conditional probability distributions; $p(\mathcal{X}_K|\mathcal{Y}_K, \boldsymbol{\theta})$ and $p(\boldsymbol{\theta}|\mathcal{Y}_K, \mathcal{X}_K)$.

For sampling the latent states, the distribution $p(\mathcal{X}_K|\mathcal{Y}_K, \boldsymbol{\theta})$ can be found by sequentially drawing samples from $p(x_k|\mathcal{X}'^{k}_K, \mathcal{Y}_K, \boldsymbol{\theta})$, where $\mathcal{X}'^{k}_K$ denotes the joint $\mathcal{X}_K$ without x_k, and $k = 1, \ldots, K$. It is easy to show that

$$p\,(x_k|\mathcal{X}'^{k}_K, \mathcal{Y}_K, \boldsymbol{\theta}) \tag{C.1}$$
$$\propto \begin{cases} p(x_k|x_{k-1}, \alpha, \rho, \sigma^2_\varepsilon)\,p(x_{k+1}|x_k, \alpha, \rho, \sigma^2_\varepsilon)\prod_{c=1}^{C} p(y_k^c|x_k, \mu, \beta^c) \\ \qquad k = 1, \ldots, K-1 \\ p(x_k|x_{k-1}, \alpha, \rho, \sigma^2_\varepsilon)\prod_{c=1}^{C} p(y_k^c|x_k, \mu, \beta^c) \qquad k = K. \end{cases}$$

In a similar fashion to the estimation of the state distribution, the parameters posterior distribution $p(\boldsymbol{\theta}|\mathcal{Y}_K, \mathcal{X}_K)$ is approached by single-site updates, where parameters are updated one at a time. These conditionals are

Algorithm 1: A Gibbs Sampler for SSPP.

1. Initialize $\mathcal{X}_K^{(0)}, \boldsymbol{\theta}^{(0)}$

2. Sampling and updating

for $n = 0, \ldots, N-1$ **do**

  for $k = 1, \ldots, K$ **do**

    Sample $x_k^{(n+1)}$ from $p(x_k|\mathcal{X}_K^{/k^{(n)}}, \mathcal{Y}_K, \boldsymbol{\theta}^{(n)})$ using (C.1)

  end for

  Sample $\boldsymbol{\theta}^{(n+1)}$ from $p(\boldsymbol{\theta}|\mathcal{Y}_K, \mathcal{X}_K^{(n+1)})$ using (C.2)-(C.5).

end for

given as

$$p(\rho|\mathcal{Y}_k, \mathcal{X}_K, \alpha, \sigma_\varepsilon^2, \mu, \boldsymbol{\beta}) \propto \prod_{k=1}^{K} p(x_k|x_{k-1}, \rho, \alpha, \sigma_\varepsilon^2) p(\rho), \tag{C.2}$$

$$p(\alpha|\mathcal{Y}_k, \mathcal{X}_K, \rho, \sigma_\varepsilon^2, \mu, \boldsymbol{\beta}) \propto \prod_{k=1}^{K} p(x_k|x_{k-1}, \rho, \alpha, \sigma_\varepsilon^2) p(\alpha), \tag{C.3}$$

$$p(\mu|\mathcal{Y}_k, \mathcal{X}_K, \rho, \alpha, \sigma_\varepsilon^2, \boldsymbol{\beta}) \propto \prod_{c=1}^{C} \prod_{k=1}^{K} p(y_k^c|x_k, \mu, \beta^c) p(\mu), \tag{C.4}$$

$$p(\beta^c|\mathcal{Y}_k, \mathcal{X}_K, \rho, \alpha, \sigma_\varepsilon^2, \mu) \propto \prod_{k=1}^{K} p(y_k^c|x_k, \mu, \beta^c) p(\beta^c). \tag{C.5}$$

In this work, the prior for each parameter is set to be the uniform distribution. The Gibbs sampler for the above conditional distributions is shown in algorithm 1. Clearly for the SSPP, it is not possible to directly draw samples from the above distributions. We hence use the Metropolis-Hastings algorithm with a random walk proposal (as in Geweke & Tanizaki, 2001) to overcome such a difficulty.[4]

We tested our Gibbs sampler based on synthetic data with the parameters set as in the main text. Figure 11 shows that the Gibbs sampler is able to converge within a burn-in period of 3000 iterations.

[4] Using the state transition density as the proposal density gives similar results in this case.

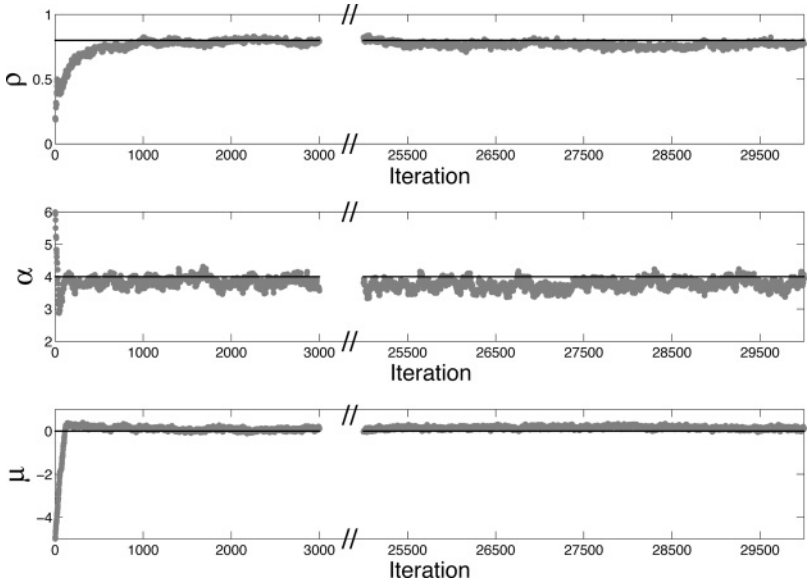

Figure 11: The trajectory of the Gibbs sampler for the unknown parameters ρ, α, and μ. The solid level line denotes the true parameter value.

References

Attias, H. (1999). Inferring parameters and structure of latent variable models by variational Bayes. In *Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence* (pp. 21–30). San Francisco: Morgan Kaufmann.

Barbieri, R., Matten, E., Alabi, A., & Brown, E. (2005). A point-process model of human heartbeat intervals: New definitions of heart rate and heart rate variability. *AJP—Heart and Circulatory Physiology, 288*(1), 424–435.

Beal, M. (2003). *Variational algorithms for approximate Bayesian inference*. Unpublished doctoral dissertation, University College London.

Boel, R., & Benes, V. (1980). Recursive nonlinear estimation of a diffusion acting as the rate of an observed Poisson process. *IEEE Transactions on Information Theory, 26*(5), 561–575.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing, 27*(2), 113–120.

Brown, E., Barbieri, R., Eden, U., & Frank, L. (2003). Likelihood methods for neural spike train data analysis. In J. Feng (Ed.), *Computational neuroscience: A comprehensive approach* (pp. 253–286). Boca Raton, FL: CRC.

Brown, E., Barbieri, R., Ventura, V., Kass, R., & Frank, L. (2002). The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation, 14*(2), 325–346.

Carter, C., & Kohn, R. (1994). On Gibbs sampling for state space models. *Biometrika, 81*(3), 541–553.

de Freitas, J., Niranjan, M., Gee, A., & Doucet, A. (2000). Sequential Monte Carlo methods to train neural network models. *Neural Computation, 12*(4), 955–993.

di Lorenzo, P., & Victor, J. (2003). Taste response variability and temporal coding in the nucleus of the solitary tract of the rat. *Journal of Neurophysiology, 90*, 1418–1431.

Doucet, A., de Freitas, N., & Gordon, N. (2001). An introduction to sequential Monte Carlo methods. In A. Doucet, N. de Freitas, & N. Gordon (Eds.), *Sequential Monte Carlo methods in practice* (pp. 3–14). New York: Springer-Verlag.

Doucet, A., de Freitas, N., Murphy, K., & Russell, S. (2000). Rao-Blackwellised particle filters for dynamic Bayesian networks. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence* (pp. 176–183). San Francisco: Morgan Kaufmann.

Eden, U., Frank, L., Barbieri, R., Solo, V., & Brown, E. (2004). Dynamic analysis of neural encoding by point process adaptive filtering. *Neural Computation, 16*(5), 971–998.

Ergün, A., Barbieri, R., Eden, U., Wilson, M., & Brown, E. (2007). Construction of point process adaptive filter algorithms for neural systems using sequential Monte Carlo methods. *IEEE Transactions on Biomedical Engineering, 54*(3), 419–428.

Fahrmeir, L., & Tutz, G. (1994). Dynamic stochastic models for time-dependent ordered paired comparison systems. *Journal of the American Statistical Association, 89*(428), 1438–1449.

Friston, K., Mattout, J., Trujillo-Barreto, N., Ashburner, J., & Penny, W. (2007). Variational free energy and the Laplace approximation. *NeuroImage, 34*, 220–234.

Geweke, J., & Tanizaki, H. (2001). Bayesian estimation of state-space models using the Metropolis-Hasting algorithm within Gibbs sampling. *Computational Statistics and Data Analysis, 37*(2), 151–170.

Ivanov, P., Rosenblum, M., Peng, C., Mietus, J., Havlin, S. Stanley, H., et al. (1996). Scaling behavior of heartbeat intervals obtained by wavelet-based time-series analysis. *Nature, 383*, 323–327.

Jolivet, R., Kobayashi, R., Rauch, A., Naud, R., Shinomoto, S., & Gerstner, W. (2008). A benchmark test for a quantitative assessment of simple neuron models. *Journal of Neuroscience Methods, 169*(2), 417–424.

Kitagawa, G. (1998). A self-organizing state-space model. *Journal of the American Statistical Association*, 93(443):1203–1215.

MacGregor, D., Williams, C., & Leng., G. (2009). A new method of spike modelling and interval analysis. *Journal of Neuroscience Methods, 176*(1), 45–56.

Manton, J., Krishnamurthy, V., & Elliott, R. (1999). Discrete time filters for doubly stochastic Poisson processes and other exponential noise models. *International Journal of Adaptive Control and Signal Processing, 13*(5), 393–416.

Quinn, J., Williams, C., & McIntosh, N. (2008). Factorial switching linear dynamical systems applied to physiological condition monitoring. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 31*(9), 1537–1551.

Riehle, A., Grün, S., Diesmann, M., & Aertsen, A. (1997). Spike synchronization and rate modulation differentially involved in motor cortical function. *Science, 278*(5345), 1950–1953.

Roussin, A., Victor, J., Chen, J.-Y., & di Lorenzo, P. (2008). Variability in responses and temporal coding of tastants of similar quality in the nucleus of the solitary tract of the rat. *J. Neurophysiol., 99*(2), 644–655.

Roweis, S., & Ghahramani, Z. (1999). A unifying review of linear gaussian models. *Neural Computation, 11*(2), 305–345.

Sato, M. (2001). On-line model selection based on the variational Bayes. *Neural Computation, 13*(7), 1649–1681.

Segall, A., Davis, M. H. A., & Kailath, T. (1975). Nonlinear filtering with counting observations. *IEEE Transactions on Information Theory, 21*(2), 143–149.

Šmídl, V., & Quinn, A. (2005). *The variational Bayes method in signal processing*. New York: Springer-Verlag.

Šmídl, V., & Quinn, A. (2006). The restricted variational Bayes approximation in Bayesian filtering. In *Proceedings of the IEEE Nonlinear Statistical Signal Processing Workshop* (pp. 224–227). Piscataway, NJ: IEEE.

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

Smith, A., Shah, S., Hudson, A., Purpura, K., Victor, J., Brown, E. et al. (2009). A Bayesian statistical analysis of behavioral facilitation associated with deep brain stimulation. *Journal of Neuroscience Methods, 183*, 267–276.

Snyder, D. L., & Miller, M. I. (1991). *Random point processes in time and space*. New York: Springer-Verlag.

Solo, V. (2000). "Unobserved" Monte Carlo method for identification of partially observed nonlinear state space systems. Part II: Counting process observations. In *Proceedings of the 39th IEEE Conference on Decision and Control* (Vol. 4, pp. 3331–3336). Piscataway, NJ: IEEE.

Storvik, G. (2002). Particle filters for state-space models with the presence of unknown static parameters. *IEEE Transactions on Signal Processing, 50*(2), 281–289.

Wan, E., & Nelson, A. (2001). Dual extended Kalman filter methods. In S. Haykin (Ed.), *Kalman filtering and neural networks* (pp. 123–173). Hoboken, NJ: Wiley.

Yuan, K., & Niranjan, M. (2010). Estimating a state-space model from point process observations: A note on convergence. *Neural Computation, 22*(8), 1993–2001.

Received October 5, 2010; accepted January 25, 2011.