Semi-parametric dynamic time series modelling with applications to detecting neural dynamics.

Fabio Rigat[*]
Jim. Q. Smith[†]

October 1st, 2008

Abstract

This paper illustrates the theory and applications of a methodology for non-stationary time series modeling which combines sequential parametric Bayesian estimation with non-parametric change-point testing. A novel Kullback-Leibler divergence between posterior distributions arising from different sets of data is proposed as a non-parametric test statistic. A closed form expression of this test statistic is derived for exponential family models whereas Markov chain Monte Carlo simulation is used in general to approximate its value and that of its critical region. The effects of detecting a change-point using our method are assessed analytically for the one-step ahead predictive distribution of a linear dynamic Gaussian time series model. Conditions under which our approach reduces to fully parametric state-space modeling are illustrated.

The method is applied to estimating the functional dynamics of a wide range of neural data, including multi-channel electroencephalogram recordings, the learning performance in longitudinal behavioural experiments and in-vivo multiple spike trains. The estimated dynamics are related to the presentation of visual stimuli, to the generation of motor responses and to variations of the functional connections between neurons across different experiments.

Introduction

Stochastic modeling of dynamic processes is often implemented via models having time-dependent parameters (Hamilton [1994], West and Harrison

---
[*]CRiSM, Department of Statistics, University of Warwick; f.rigat@warwick.ac.uk
[†]CRiSM, Department of Statistics, University of Warwick; J.Q.Smith@warwick.ac.uk

1

[1997], Frühwirth-Shnatter [2006]). These dynamic time series models can effectively capture non-stationarities induced by the occurrence of change-points in the data dependence structure (Page [1955], Smith [1975], Carlin et al. [1992], Ferger [1995], Chib [1998]), by switches among different dependence regimes (Hamilton [1990], Shumway and Stoffer [1991], Robert et al. [1993], Albert and Chib [1993], McCulloch and Tsay [1994], Kim [1994], Ghahramani and Hinton [2000], Frühwirth-Shnatter [2001]) or by smooth changes of the model parameters through time (Harrison and Stevens [1976], West and Harrison [1986a]). This paper illustrates the theory and applications of a novel sequential method for estimating semi-parametrically the dynamics of such time-dependent model coefficients.

The distinctive characteristic of our approach with respect to state-space and hidden Markov models (Kalman [1960], West et al. [1985], West and Harrison [1997], Cappe et al. [2005]) and parametric change-point models (Muller [1992], Stephens [1994], Loader [1996], Mira and Petrone [1996], Bélisle et al. [1998], Jain et al. [2007], Fearnhead and Liu [2007]) is that change-points are defined through the discrepancy between one-step ahead predictive distributions, measured here by a novel Kullback-Leibler (KL) divergence (Kullback and Leibler [1951], Kullback [1997]). The distribution of this KL statistic reflects the concentration of the joint posterior distribution of the model's parameters when new data are generated using the same model structure as for past observations and parameter values drawn from their current joint posterior distribution. An important implication of this approach is that the occurrence of changes in the model coefficients is not assumed as in standard state-space frameworks but it is sequentially tested at each time point.

The motivation for adopting this sequential framework for dynamic time series modeling is twofold. First, in state-space models inferences and predictions are sensitive to the form of the state evolution equations (Frühwirth-Shnatter [1995], Bengtsson and Cavanaugh [2006]). Therefore, an exploratory non-parametric approach is a natural choice for a first analysis of the data when no reliable information about the latent evolution of a model's parameters are available (Robinson [1983], Härdle et al. [1997]). This is typically the case for many biological systems, where dynamic responses to novel experimental conditions are difficul to anticipate. Second, fully Bayesian sequential inference for a model's parameters and for a latent multiple change-point process is impractical unless marginal likelihoods can be calculated explicitly. This is because the Bayes factors measuring the strength of evidence in the data about the occurrence of change-points can only be approximated (Han and Carlin [2001]). The approach taken in this work represents a

2

practical alternative providing a non-parametric point estimate of a latent change-point process without computing marginal likelihoods.

Section 1 of this paper includes the methodological developments. A general time series framework is introduced and the KL test is illustrated. A closed form expression of the KL statistic for exponential family models is derived and examples are presented. Markov chain Monte Carlo simulation (Gelfand and Smith [1990], Tierney [1994]) is used to estimate in general the value of the KL statistic and of its critical region under the null hypothesis. A sequential algorithm integrating parametric Bayesian inference with a non-parametric change-point test using the KL statistic is presented. The behaviour of both location and spread of the one-step ahead predictive distribution is described analytically as a function of the timing of the last detected change-point for a linear Gaussian dynamic model with conjugate priors. Conditions are given so that our semi-parametric approach reduces to fully parametric state-space modeling.

In sections 2, 3 and 4 our method is applied to estimating three different types of neural dynamics. First we analyse a multivariate time series of electroencephalogram (EEG) recordings (Delorme et al. [2002], Makeig et al. [2002]) to reconstruct the dynamics functional relationships among different brain areas. Second, we estimate semi-parametrically a learning curve using a univariate binary time series arising from a longitudinal behavioural experiment (Smith et al. [2004]). Finally, our method is applied to estimating the functional dynamics of networks of neurons using in-vivo experimental multiple spike trains recordings (Buzsáki [2004]).

1 Sequential time series modelling and Kullback-Leibler change-point testing

Let $\{Y_i\}_{i=1}^N$ represent a sequence of N consecutive time series $Y_i \in \mathcal{Y}$ of random variables $Y_{i,k,t}$ with $k = 1, ..., K$ measured at the time points $t = t_{i,1} < t_{i,2} < ... < t_{i,n_i}$ with $t_{i,n_i} < t_{i+1,1}$. The distinction between the N time series is relevant when we allow for the occurrence of time gaps of possibly unknown length between them. This situation arises, for instance, when N consecutive trials are run sequentially interposed by resting periods.

The marginal probability of the data $(Y_1 = y_1, ..., Y_N = y_n)$ can always be written as

$$P(y_1, ..., y_N) = \prod_{i=1}^N P_i(y_i \mid y^{0:(i-1)}), \tag{1}$$

3

where $y^{0:(i-1)}$ includes fixed initial conditions y^0 and the observations $(y_1, ..., y_{i-1})$ (see, for instance Dawid [1984]). Explicit representations of the left-hand side of (1) are obtained by specifying suitable conditional distributions $P_1(y_1 \mid y^0), ..., P_N(y_N \mid y^{0:(N-1)})$. From a Bayesian perspective, these conditional distributions are often constructed by assuming that

$$P_i(y_i \mid y^{0:(i-1)}) = \int_{\Theta_i} P(y_i \mid \theta_{i-1}, y^{0:(i-1)}) f(\theta_{i-1} \mid y^{0:(i-1)}) d\theta_{i-1}, \qquad (2)$$

meaning that the multivariate time series Y_i are assumed to be generated by a fixed model $P(y_i \mid \theta_{i-1}, y^{0:(i-1)})$, such as a vector auto-regressive (VAR) model with shared parameters θ_{i-1} within each of the N periods. The probability density $f(\theta_{i-1} \mid y^{0:(i-1)})$ here represents the distribution of the model coefficients given the initial conditions y^0 and all past observations. Note that in (2), although the parameter values are allowed to vary in time, neither the functional form of the likelihood function nor the interpretation of its coefficients change over time.

Within this framework, dynamic modeling consists of specifying a transfer map taking as arguments the posterior density $f(\theta_{i-1} \mid y^{0:(i-1)})$, the time series data y_i and possibly additional hyper-parameters α and returning the density $f(\theta_i \mid y^{0:i})$ for any value $i = 1, ..., N$. Various characterisations of analogous maps are given, for instance, in Smith [1990] and Smith [1992]. In standard state-space models the coefficients are assumed to vary across each time period, so that their transfer map is defined implicitly via the transition equations defining the prior distribution for θ_i in terms of the coefficients θ_{i-1} and of fixed hyper-parameters. In Markov switching and finite mixture time series models, this transfer map is again derived from modelling the full joint density of the coefficients θ_i and θ_{i-1} conditional on the location of a sequence of change-points (Frühwirth-Shnatter [2006]). In this work we model the transfer map semi-parametrically by integrating sequential Bayesian inference for the model's parameters with a statistic detecting the occurrence of change-points. The latter statistic is similar in spirit to the cumulative Bayes factors proposed in West [1986] and West and Harrison [1986b], with the practical advantage that the computation of marginal likelihoods is not required.

When a change-point occurs upon observing the data y_i, we define a transfer prior

$$\theta_i \sim h(\theta_i \mid \hat{\theta}_{i-1}), \qquad (3)$$

taking as arguments functionals of the current posterior density, $\hat{\theta}_{i-1}$, and returning a prior density $h(\cdot)$ for the model's parameters θ_i. Among the

4

many possible formulations of this transfer prior we let the hyperparameters of the conditional prior density for θ_i be fully specified using estimates of the first two moments of the current posterior density. We note that similar forms of prior moment matching have been used for dynamic point process modeling by Gamerman [1992] and for multi-process dynamic linear models by West and Harrison [1997].

Equation (3) represents a partially specified state evolution density where neither the exact form of the prior nor the time of occurrence of the change-points are given a priori. Specific choices for the prior density will in fact depend on the structure of the time series model being entertained and on the interpretation of its parameters. When no change-points are detected prior to observing the data $Y_i = y_i$, under (3) the joint posterior density of the model's parameters is

$$f(\theta_i \mid y^{0:i}, \alpha) \propto \begin{cases} f(\theta_{i-1} \mid y^{0:(i-1)}) P(y_i \mid \theta_{i-1}, y^{0:(i-1)}) & \text{if } y_i \in \Psi_i(\alpha), \\ h(\theta_i \mid \hat{\theta}_{i-1}) P(y_i \mid \theta_i, y^{0:(i-1)}) & \text{if } y_i \notin \Psi_i(\alpha), \end{cases} \quad (4)$$

where $\Psi_1(\alpha) = \mathcal{Y}$ and for $i = 2, ..., N$ the sets $\Psi_i(\alpha) \subseteq \mathcal{Y}$ include the time series Y_i which are inconsistent with their observed past under the assumed model and the hyper-parameters α.

The implementation of (4) presents two related challenges. First, it is essential to reformulate the rejection sets $(\Psi_2(\alpha), ..., \Psi_N(\alpha))$ in terms of a low-dimensional statistic and of the hyper-parameters α. Second, it must be possible to derive the distribution of such statistic over the sample space at least approximately, so as to provide a sequential approximation of the rejection sets for any value of α. A natural way to overcome these challenges is to view the sets $(\Psi_2(\alpha), ..., \Psi_N(\alpha))$ as the α-level critical regions of a sequential change-point test based on an appropriate statistic. The transfer map is thus completely specified by the transfer prior (3) together with a choice of this test statistic. The next section illustrates the derivation of a suitable statistic using well established information theoretic principles.

1.1 A Kullback-Leibler change-point statistic

The KL divergence (Kullback and Leibler [1951]) is a well-known information-theoretic criterion with many applications in statistics, such as density estimation (Hall [1987], Hastie [1987]), model selection (Akaike [1978], Akaike [1981], Carota et al. [1996], Goutis and Robert [1998]), experimental design (Lindley [1956], Stone [1959]) and the construction of uninformative prior distributions (Bernardo [1979]). The geometric properties of the KL divergence have been throughly explored: see for instance Critchley et al. [1994].

5

The statistic proposed in this work has an analogous function to the KL divergence when used to support model selection. In the latter case, the null hypothesis specifies an assumed model structure which is compared to a typically more parsimonious formulation. In (4), under the hypothesis of no change sequential Bayesian learning is used to update the joint distribution of the model parameters. When this hypothesis is not sufficiently supported by the data, their distribution is updated using the conditional transfer prior (3). Therefore, instead of testing which of two competing model structures best predicts one given set of data, we construct a statistic detecting whether the same parameter values could have likely generated two sets of data given a common model structure.

As a statistic to measure the evidence in the data in favour of the hypothesis of no change we adopt a Kullback-Leibler divergence

$$KL(y^{0:(i+1)}) = \int_{\Theta_i} \log \left(\frac{f(\theta_i | y^{0:i})}{f(\theta_i | y^{0:(i+1)})} \right) f(\theta_i | y^{0:i}) d\theta_i,$$

$$= \log(E_{\theta_i | y^{0:i}}(P(y_{i+1} | \theta_i, y^{0:i}))) - E_{\theta_i | y^{0:i}}(\log(P(y_{i+1} | \theta_i, y^{0:i}))), \quad (5)$$

where the expectations in (5) are taken with respect to the posterior density $f(\theta_i | y^{0:i})$. The right-hand side of (5) is finite when the likelihood function is bounded away from zero for all values of the model's parameters and when their posterior density is proper. In this case (5) is a non-negative convex function measuring the discrepancy between the posterior densities $f(\theta_i | y^{0:i})$ and $f(\theta_i | y^{0:(i+1)})$, being equal to zero if and only if the likelihood function does not vary with θ_i over the range of the former posterior density. Prior to observing the data $Y_{i+1} = y_{i+1}$, (5) is a random variable which distribution under the null hypothesis depends on that of the future data Y_{i+1} via the time series model $P(Y_{i+1} | \theta_i, y^{0:i})$. The following sections focus on the interpretation and on the computation of (5).

1.1.1 Interpretation of the KL statistic and of the change-points

The hyper-parameter α of the joint posterior (4) has the interpretation of the type-1 error probability for the change-point test using the statistic (5). The rejection sets can be written explicitly as intervals $\Psi_i(\alpha) = (l_{i,\alpha}, u_{i,\alpha})$, each representing the α-level highest probability interval for the random variable (5) under the hypothesis of no change over period i.

When (5) lies below the value $l_{i,\alpha}$, the likelihood is almost a constant function of the parameters θ_i for non-negligible values of their current posterior density. In other terms, the parameter values maximising the likelihood of the observed y_{i+1} conditionally on the past data $y^{0:i}$ are given almost zero

6

probability by the posterior distribution under the hypothesis of no change. If the change-point statistic lies above $u_{i,\alpha}$, the parameter values maximising the likelihood of the data y_{i+1} are associated to substantial values of the current joint posterior density but they are far from its global maximum. In this case the joint posterior density of all data $y^{0:(i+1)}$ under the hypothesis of no change tends to be bimodal, signaling that the latest batch of data y_{i+1} are not adequately explained by the current parameter values.

In both the above cases, the detection of change-points using the statistic (5) indicate that, with probability $1 - \alpha$, the estimated shape of the posterior density of $y^{0:(i+1)}$ significantly departs from its assumed shape under the hypothesis that standard sequential Bayesian inference is adequate for estimating the model parameters in time.

When $\alpha = 0$ no change-point is ever detected, so that the only revisions in the estimates of the model's parameters are those induced by the sequential accrual of information via Bayes' rule. On the other end, if $\alpha = 1$ a change-point in the parameter values is systematically detected at every time point. In this second limiting case the methodology proposed in this work is equivalent to a fully parametric first order Markov state-space model with state evolution equation partially specified by the transfer prior (3).

1.1.2 Computation of the change-point statistic

In general neither the value of (5) nor the rejection sets $(\Psi_2(\alpha), ..., \Psi_N(\alpha))$ can be derived in closed form under the hypothesis of no change. At each time period $i = 2, ..., N$ this change-point statistic and its critcal interval can be approximated using a sequence of draws $\{\theta_i^m\}_{m=1}^M$ generated using a Markov chain Monte Carlo algorithm (Gelfand and Smith [1990], Smith and Roberts [1993], Tierney [1994]) having as target the posterior probability density of the model's parameters at period i conditional on the change-points detected in the past. Using this technique, the value of (5) can be approximated using the average

$$KL(y^{0:(i+1)}) \approx \log\left(\frac{\sum_{m=1}^M P(y_{i+1} \mid \theta_i^m, y^{0:i})}{M}\right) - \frac{\sum_{m=1}^M \log(P(y_{i+1} \mid \theta_i^m, y^{0:i}))}{M}. \quad (6)$$

An approximation of the distribution of (5) for varying values of Y_{i+1} under the hypothesis of no change can be constructed using the same sequence of draws as follows:

   i) for each draw θ_i^m generate a pseudo-realisation y_{i+1}^m using the joint sampling probability $P(Y_{i+1} \mid \theta_i^m, y^{0:i})$;

ii) compute the statistic $KL(y_m^{0:(i+1)})$, where $y_m^{0:(i+1)} = (y_0, ..., y_i, y_{i+1}^m)$, using its Monte Carlo approximation (6).

The empirical distribution of the sequence $\{KL(y_m^{0:(i+1)})\}_{m=1}^M$ approximates that of the KL statistic (5) under the hypothesis of no change. Therefore the empirical $(\frac{\alpha}{2}, 1 - \frac{\alpha}{2})$th percentiles of the sequence $\{KL(y_m^{0:(i+1)})\}_{m=1}^M$ approximate the rejection sets $\Psi_i(\alpha) = (l_{i,\alpha}, u_{i,\alpha})$ for any given value of α.

1.2 Sequential fitting and change-point testing algorithm

Given initial conditions (α, y_0) and a parametric time series model $P(Y_{i+1} \mid \theta_i, y^{0:i})$, time-dependent inferences for its coefficients θ_i can be derived by integrating Bayesian sequential learning with the change-point KL test described above. We illustrate this algorithm starting from the first sample y_1:

i) upon observing the data y_1, derive the posterior density

$$f(\theta_1 \mid y^{0:1}) \propto h(\theta_i \mid y_0)P(y_i \mid \theta_i, y_0),$$

ii) having observed data y_2, compute the statistic $KL(y^{0:2})$ and its rejection set $\Psi_i(\alpha) = (l_{1,\alpha}, u_{1,\alpha})$ as described in the previous section;

iii) if $l_{1,\alpha} < KL(y^{0:2}) < u_{1,\alpha}$, update the posterior density as

$$f(\theta_1 \mid y^{0:2}, \alpha) \propto f(\theta_1 \mid y^{0:1})P(y_2 \mid \theta_1, y^{0:1}).$$

Otherwise, derive the conditional posterior density

$$f(\theta_2 \mid \hat{\theta}_1, y^{0:2}, \alpha) \propto h(\theta_2 \mid \hat{\theta}_1)P(y_2 \mid \theta_2, y^{0:1}),$$

where $\hat{\theta}_1$ represents estimates of the first two moments of the $f(\theta_1 \mid y^{0:1})$.

In the latter case, the sequentially estimated process of change-points up to and including times $(1, 2)$ reports one change at time 2. Consistently with the interpretation of the KL statistic, the model parameters are iteratively re-estimated using all data starting from the last detected change-point, if any. When a change is detected at level $1 - \alpha$, the new parameter values are estimated via their conditional posterior distribution using the transfer prior (3) and the likelihood of the latest batch of data.

8

1.3 Change-point KL statistic for exponential family models

The properties of the KL divergence within the exponential family have been explored by McCulloch [1988]. Also the divergence (5) has a closed form when the data Y_i is generated by an exponential family model. In this case stochastic simulation is necessary to approximate only the value of the interval $\Psi_i(\alpha) = (l_{i,\alpha}, u_{i,\alpha})$ under the hypothesis of no change. Without loss of generality, in what follows we assume that no change-point is detected prior to priod i and we let Y_i be a $1 \times n_i$ dimensional sample of conditionally independent observations with joint density (Diaconis and Ylvisaker [1979])

$$P(Y_i \mid \theta_i) = \prod_{j=1}^{n_i} a(Y_{i,j}) e^{Y_{i,j}\theta_i - b(\theta_i)}, \tag{7}$$

where θ_i is a scalar canonical parameter. Diaconis and Ylvisaker [1979] show that each elment of Y_i has mean and variance

$$E(Y_{i,j} \mid \theta_i) = \frac{\partial b(\theta_i)}{\partial \theta_i}, V(Y_{i,j} \mid \theta_i) = \frac{\partial^2 b(\theta_i)}{\partial \theta_i' \partial \theta_i}.$$

Using the prior

$$f(\theta_i \mid n_0, y_0) = c(n_0, S_0) e^{S_0 \theta_i - n_0 b(\theta_i)},$$

where $S_0 = n_0 y_0$ for scalar n_0 and y_0, the posterior for θ_i given the past data $y^{0:i}$ has conjugate density

$$f(\theta_i \mid n(i), y^{0:i}) = c(n(i), S(i)) e^{n(i)\left(\frac{S(i)}{n(i)}\theta_i - b(\theta_i)\right)} \tag{8}$$

where $n(i) = \sum_{j=0}^{i} n_j$, $S(i) = \sum_{j=0}^{i} n_j \bar{y}_j$ and $\bar{y}_j$ represents the arithmetic mean of sample y_j. Using the results of Gutiérrez-Peña [1997], the posterior mean and variance of θ_i are

$$E(\theta_i \mid n(i), S(i)) = \frac{\partial H(n(i), S(i))}{\partial S(i)}, V(\theta_i \mid n(i), S(i)) = \frac{\partial H(n(i), S(i))}{\partial S(i)^2},$$

and the posterior mean and variance of the function $b(\theta_i)$ are

$$E(b(\theta_i) \mid n(i), S(i)) = \frac{\partial H(n(i), S(i))}{\partial n(i)}, V(b(\theta_i) \mid n(i), S(i)) = \frac{\partial H(n(i), S(i))}{\partial n(i)^2},$$

where $H(n(i), S(i)) = -\log\left(c(n(i), S(i))\right)$. Using these results we can derive an explicit form for the KL divergence (5), which is stated in the following

theorem.

Theorem: When the posterior density for the coefficients θ_i has form (8), given the data up to and including y_{i+1} the Kullback-Leibler statistic (5) is

$$
\begin{aligned}
KL(y^{0:(i+1)}) &= \log\left(\frac{c(n(i), S(i))}{c(n(i+1), S(i+1))}\right) - S_{i+1}\frac{\partial H(n(i), S(i))}{\partial S(i)} + \\
&+ n_{i+1}\frac{\partial H(n(i), S(i))}{\partial n(i)},
\end{aligned}
\tag{9}
$$

where the terms on the right hand side of (9) are defined above.

Proof: by letting the posterior densities $f(\theta_i \mid n(i), S(i))$ and $f(\theta_i \mid n(i+1), S(i+1))$ have form (8), the KL (5) becomes

$$
KL(y^{0:(i+1)}) = \log\left(\frac{c(n(i), S(i))}{c(n(i+1), S(i+1))}\right) - S_{i+1}E(\theta_i) + n_{i+1}E(b(\theta_i)). \tag{10}
$$

For exponential family models, the expectations of θ_i and $b(\theta_i)$ with respect to $f(\theta_i \mid y^{0:i})$ in equation (10) are given in Gutiérrez-Peña [1997], as reported above. By substituting these expressions in (10), equation (9) obtains. $\diamond$

Example 1.1: when Y_i is a Gaussian random variable with mean μ_i and precision λ_i, its distribution can be written in the form (8) using the two-dimensional statistic

$$
Y_i^* = [Y_i, Y_i^2],
$$

and the canonical parameter

$$
\theta_i = [\theta_{1,i}, \theta_{2,i}] = \left[\lambda_i\mu_i, -\frac{\lambda_i}{2}\right].
$$

with

$$
\begin{aligned}
a(Y_i^*) &= (2\pi)^{-\frac{1}{2}}, \\
b(\theta_i) &= = -\frac{1}{2}\log(-2\theta_{2,i}) - \frac{\theta_{1,i}^2}{\theta_{2,i}}.
\end{aligned}
$$

The conjugate prior for (μ_i, λ_i) is Normal-Gamma $N(\mu_i \mid \gamma, \lambda_i(2\alpha-1))Ga(\lambda_i \mid \alpha, \beta)$ with coefficients $\alpha > 0.5, \beta > 0, \gamma \in \mathcal{R}$ and normalising constant (Bernardo and Smith [2007])

$$
c(n_0, S_0) = \left(\frac{2\pi}{n_0}\right)^{\frac{1}{2}}\frac{\frac{1}{2}S_{2,0}^{\frac{S_{1,0}}{2}}}{\Gamma\left(\frac{n_0+1}{2}\right)},
$$

10

where $n_0 = 2\alpha - 1$, $y_0^* = [y_{1,0}^*, y_{2,0}^*] = [\gamma, \frac{2\beta}{2\alpha-1} + \gamma^2]$, $S_{1,0} = n_0 y_{1,0}^*$ and $S_{2,0} = n_0 y_{2,0}^*$. Upon observing the realisation $(y_1, ..., y_i)$, the normalising constant of the corresponding conjugate posterior is

$$c(n(i), S(i)) = \left(\frac{2\pi}{n(i)}\right)^{\frac{1}{2}} \frac{\frac{1}{2} S(2,i)^{\frac{S(1,i)}{2}}}{\Gamma\left(\frac{n(i)+1}{2}\right)},$$

where $n(i) = n_0 + i$, $S(1,i) = S_{1,0} + \sum_{j=1}^{i} y_j$ and $S(2,i) = S_{2,0} + \sum_{j=1}^{i} y_j^2$. When also y_{i+1} is observed, using (9) the KL statistic can be written as

$$KL(y^{0:(i+1)}) = \log\left(\frac{\Gamma\left(\frac{n(i+1)+1}{2}\right)}{\Gamma\left(\frac{n(i)+1}{2}\right)}\right) + \frac{1}{2}\log\left(\frac{n(i+1)}{n(i)}\right) + \log\left(\frac{S(2,i)^{\frac{S(1,i)}{2}}}{S(2,i+1)^{\frac{S(1,i+1)}{2}}}\right) -$$

$$- \frac{y_{i+1}}{2}\log\left(\frac{S(2,i)}{2}\right) - y_{i+1}^2 \frac{S(1,i)}{S(2,i)} + \frac{1}{2n(i)} + \Gamma\left(\frac{n(i)+1}{2}\right)\frac{\partial\Gamma\left(\frac{n(i)+1}{2}\right)}{\partial n(i)}.$$

Example 1.2: let Y_i be a sample of size n_i of conditionally independent Bernoulli random variables with success probabilities $\{\pi_i\}_{i=1}^{N}$. The canonical representation of the Bernoulli probability mass function obtains by letting $\theta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$, $b(\theta_i) = \log\left(1 + e^{\theta_i}\right)$ and $a(Y_i) = 1$. The conjugate prior for π_i is $Beta(S_0, m_0)$ where $m_0 = n_0 - S_0$. Upon observing $(y_1, .., y_i)$ the conjugate posterior is $Beta(S(i), m(i))$, where $S(i) = \sum_{j=0}^{i} S_j$, $n(i) = \sum_{j=0}^{i} n_j$ and $m(i) = n(i) - S(i)$. When also y_{i+1} is observed, the KL statistic (9) has form

$$KL(y^{0:(i+1)}) = \log\left(\frac{\prod_{k=1}^{n_i}(n(i)+k)\prod_{w=1}^{n_i-S_i}(n(i)-S(i)+w)}{\prod_{j=1}^{S_i}(S(i)+j)}\right) -$$

$$- S_{i+1}\frac{\Gamma(S(i))}{\Gamma(n(i)-S(i))}\frac{\partial\frac{\Gamma(n(i)-S(i))}{\Gamma(S(i))}}{\partial S(i)} + n_{i+1}\frac{\frac{\partial(\Gamma(n(i))\Gamma(n(i)-S(i)))}{\partial n(i)}}{\Gamma(n(i))\Gamma(n(i)-S(i)))}.$$

Example 1.3: let Y_i represent the random number of events of a given kind observed within a time interval $(t_{i,1}, t_{i,n_i}]$ of fixed length. For this example we assume that the latter is identical for all samples $i = 1, ..., N$. Let the random times at which the events take place be distributed according to a homogeneous Poisson process with intensity λ_i, so that the distribution of Y_i is Poisson with parameter $\lambda_i^* = \lambda_i(t_{i,n_i} - t_{i,1})$. The canonical form of the Poisson distribution has parameter $\theta_i = \log(\lambda_i^*)$ and functions $a(Y_i) = \frac{1}{Y_i!}$, $b(\theta_i) = e^{\theta_i}$. The conjugate prior for λ_i^* is Gamma with parameters $Ga(S_0, n_0)$ having mean y_0 and variance $\frac{y_0}{n_0}$. Upon observing $(y_1, .., y_i)$ the

11

conjugate posterior for λ_i^* is $Ga(S(i), n(i))$ with $S(i) = S_0 + \sum_{j=1}^{i} y_j$, $n(i) = n_0 + i$. When also y_{i+1} is observed, using (9) the KL statistic has form

$$KL(y^{0:(i+1)}) = \log\left(\frac{S(i)n(i)^{S(i)}}{n(i+1)^{S(i+1)}}\right) + y_{i+1}\left(\log(n(i)) - \frac{\frac{\partial \Gamma(S(i))}{\partial S(i)}}{\Gamma(S(i))}\right) - \frac{S(i)}{n(i)}.$$

1.4   Effect of change-points for predictive densities

In this section we illustrate analytically the effect of detecting a change-point on the one-step ahead predictive density using the transfer prior (3) and a conjugate Gaussian model. As in the previus section, a univariate time series is considered so that $K = 1$ and $n_i = 1$ for all $i = 1, ..., N$. For each value of i, in what follows we let Y_i be distributed as $N(\mu_i, \sigma_i^2)$. Analogously to example 1.1 in the previous section, the prior distribution for $\theta_i = (\mu_i, \sigma_i^2)$ is taken as the conjugate form

$$\mu_i \sim N(\hat{\mu}_{i^*-1}, \sigma_i^2),$$
$$\sigma_i^2 \sim IGa(\frac{\nu}{2}, \frac{\nu}{2}\hat{\sigma}_{i^*-1}^2),$$

where $1 \le i^* < i$ is the time of the last detected change-point and $(\hat{\mu}_{i^*-1}, \hat{\sigma}_{i^*-1}^2)$ represent the estimated mean and variance of the joint posterior density at time i^*. If $i^* = 1$, (μ_0, σ_0^2) represents a fixed initial condition. The variance here has marginal inverse Gamma prior with density

$$f(\sigma_i^2 \mid \nu, \hat{\sigma}_{i^*-1}^2) = \frac{(\frac{\nu}{2}\hat{\sigma}_{i^*-1}^2)^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})}\sigma_i^{-2(\frac{\nu}{2}+1)}e^{-\frac{\nu\hat{\sigma}_{i^*-1}^2}{2\sigma_i^2}}.$$

Under this formulation, the prior expectation of the mean is $\hat{\mu}_{i^*-1}$ and that of the variance is $\frac{\frac{\nu}{2}}{\frac{\nu}{2}-1}\hat{\sigma}_{i^*-1}^2$. If follows that the one-step ahead marginal predictive distribution is non-central Student-t. In absence of a change-point previous to time i, the predictive density is

$$Y_{i+1} \mid y^{1:i}, \mu_0, \sigma_0^2 \sim t_{\nu+i}\left(\tilde{\mu}_i, \frac{i+1}{i+2}\frac{\nu+i}{2}\frac{1}{\tilde{\sigma}_i^2}\right), \tag{11}$$

where

$$\tilde{\mu}_i = \frac{1}{1+i}\mu_0 + \frac{i}{1+i}\bar{y}^{(1:i)},$$
$$\tilde{\sigma}_i^2 = \frac{\nu}{2}\sigma_0^2 + \frac{i}{2}s_{(1:i)}^2 + \frac{i}{i+1}(\mu_0 - \bar{y}^{(1:i)})^2,$$

12

and $(\bar{y}^{(1:i)}, s^2_{(1:i)})$ represent respectively the sample mean and variance of the data $y^{1:i}$. If a change-point is detected by the KL statistic (5) at time $1 < i^* < i$, under the transfer prior (3) the conditional predictive density is

$$Y_{i+1} \mid \hat{\mu}_{i^*-1}, \hat{\sigma}^2_{i^*-1}, y^{i^*:i} \sim t_{\nu+i-(i^*-1)}\left(\tilde{\mu}_i^*, \frac{i-i^*+2}{i-i^*+3}\frac{\nu+i-i^*+1}{2}\frac{1}{(\tilde{\sigma}_i^2)^*}\right), (12)$$

where

$$\tilde{\mu}_i^* = \frac{1}{i-i^*+2}\hat{\mu}_{i^*-1} + \frac{i-(i^*-1)}{i-i^*+2}\bar{y}^{(i^*:i)},$$

$$(\tilde{\sigma}_i^2)^* = \frac{\nu}{2}(\hat{\sigma}_{i^*-1})^2 + \frac{i-(i^*-1)}{2}s^2_{(i^*:i)} + \frac{i-(i^*-1)}{i-i^*+2}(\hat{\mu}_{i^*-1} - \bar{y}^{(i^*:i)})^2.$$

Since the mean and variance of the non-central Student-t random variable with density $t_\nu(\mu, \sigma^2)$ are respectively equal to μ and to $\frac{\nu}{\nu+2}\sigma^2$, equations (11) and (12) provide a characterisation of the one-step ahead posterior predictive moments as a function of the time of the last detected change-point and of the inverse-Gamma prior coefficient ν. For $i^* > 1$ the predictive mean is less influenced by the sample mean of the data preceding the change-point, $\bar{y}^{1:(i^*-1)}$, and it is more influenced by $\bar{y}^{i^*:i}$, that is the sample mean of the data from the change-point on. When a change-point is detected the predictive variance is larger with respect to the case of no change. Its relative increase is a decreasing function of the difference $(i - i^*)$ and it is an increasing function of the coefficient ν, which measures the strenght of the prior at the initial time.

This behaviour is consistent with the intuition that predictions ensuing from a dynamic time series model should appropriately discount the information content of remote data and focus on more recent data when significant dynamics occur. In absence of an autoregressive model structure, as in the present section, the distinction between remote and recent data is entirely left to the timing of the detected change-points.

2 Analysis of multivariate EEG recordings

This section presente the first application of the methods discussed above for the estimation of neural functional dynamics. The multivariate electroencephalogram (EEG) recordings analysed in this section arise from a sequence of 80 short tests each having length of three seconds. During each test, a subject's task was to press a button when a green square appeared in a specific screen location (Makeig et al. [2002], Delorme et al. [2002]).

13

Previous analyses of these data have emphasized an increased synchronization of different brain areas during the generation of the motor response to the visual stimulus. Here we use this data to investigate the relationships among seven functionally distinct brain areas. The multidimensional EEG time series are modeled as a discrete time Gaussian stochastic process, which randomness is thought of as arising from the intrinsic variability of the brain activity and from the presence of experimenatal artifacts.

The recordings from all cranial EEG channels are first averaged within each brain area for each trial. The rationale for this data reduction pre-processing is that the variability across trials of the EEG recordings within each brain region is small. The main features of the trial-averaged EEG recordings are illustrated in the top-left plot of Figure 1. The activity of the different brain areas prior to the presentation of the visual cue appears to be tightly synchronised, exhibiting a markedly periodic pattern occurring respectively at 10Hz and 60Hz for all seven areas. The lower frequency is consistent with the so-called α band, reflecting eye movements. The higher frequency is an experimental artifact suggestive of poor electrode grounding.

The seven-dimensional trial-averaged signal at time i, Y_i, is modelled as $N_7(\mu_i, \Sigma_i)$. To derive Bayesian inferences for the mean vector and for the covariance matrix, we use the conjugate Normal-Inverse-Wishart prior:

$$\mu_i \sim N_7(\hat{\mu}_{i^*-1}, \Sigma_i), \tag{13}$$

$$\Sigma_i \sim IW_7(9, \hat{\Sigma}_{i^*-1}), \tag{14}$$

where $1 \leq i^* < i$ is the time index of the last detected change-point prior to time i. The marginal prior expectations are matched to the corresponding estimated posterior moments $(\hat{\mu}_{i^*-1}, \hat{\Sigma}_{i^*-1})$ consistently with the transfer prior (3). The initial conditions μ_0 and Σ_0 were set respectively equal to the null vector and to the identity matrix. The hyper-parameter of the posterior density was set at $\alpha = 0.01$ so as to dectect only the most prominent changes. The number of degrees of freedom of the Inverse Wishart density is set so that predictive intervals of length consistent with the set value of α are not excessively inflated when a change is detected. The distribution of the KL statistic and its value were approximated at each time i using the last 500 Gibbs sampler draws of the mean and of the covariance matrix.

Figure 1 shows the estimated change-point process, the one-step ahead marginal posterior point predictions and their 95% highest posterior pre-dictive intervals for each of the seven brain areas. The periodograms of these predictions indicate that the periodicities at 10Hz and 60Hz frequency have been effectively filtered out. The predictions emphasize a down-ward

14

shift in brain activity taking place roughly at half of the initial phase of the experiment, followed by a sharp increase corresponding to the cue presentation, a down-ward trend in the activity following the motor response and a stabilisation of the EEG signals towards the end of the experiment. The sharp increases in the activity of the frontal and central areas during the generation of the response are consistent with their characterisation as executive and motor centres of the brain.

Figure 2 depicts the estimates of the time-dependent variance and covariance functions for each brain area. Whole segments represent periods during which their 95% highest posterior intervals do not intersect zero. All estimated variances and covariances vary over time, indicating that a time-dependent covariance matrix is an appropriate modelling assumption for this data. All estimated covariances are positive, suggesting that the activity of the seven brain areas is dynamically cooperative as found by Delorme et al. [2002]. The covariances between most brain areas are increased upon detecting a change-point, suggesting a temporary increase in their mutual coordination. An important feature of the estimated covariance functions is their spatial ordering over time, the strongest relationships being estimated between adjacent brain areas. Since neither in the Gaussian likelihood nor the priors (13) include a spatial component, these estimates suggest a close correspondence between the detected functional relationships and the anatomical structure of the brain.

3   Estimation of a learning curve

Smith et al. [2004] introduced a parametric state-space model for inferring the learning performance of macaque monkeys using longitudinal behavioural experiments. The learning curve is thereby modeled using univariate binary time series data along with a logit link for each trial's success probability and a Gaussian state evolution equation for the parameters' dynamics. In this section we use the same Bernoulli sampling distribution for the binary trial outcomes as in Smith et al. [2004] and we estimate the dynamics of its success probability over time using the semi-parametric method illustrated in section 1. The main difference between our model and that of Smith et al. [2004] is that we do not need a link function, thus imposing fewer parametric constraints on the shape of the learning curve.
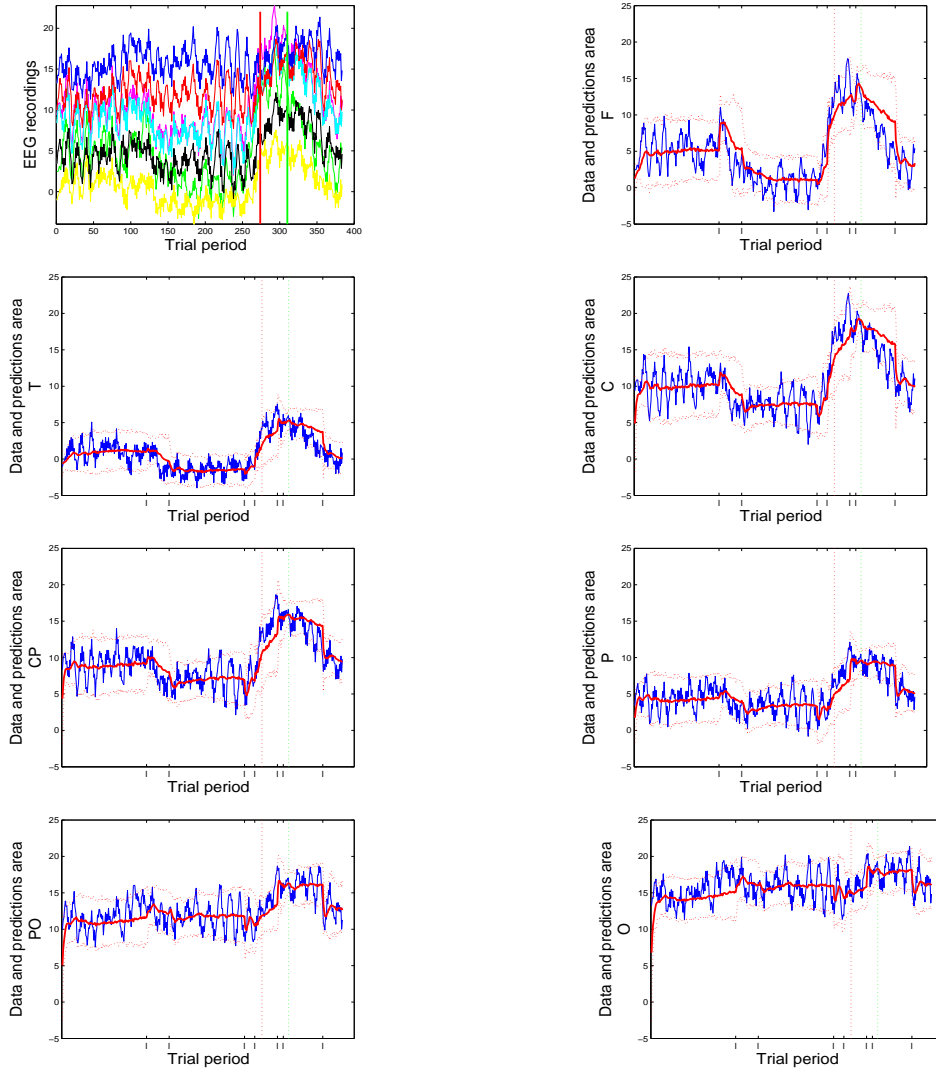
15

Figure 1: trial-averaged EEG recordings (top-left) from the frontal (green), temporal (yellow), central (magenta), central-parietal (cyan), parietal (black), parietal-occipital (red) and occipital (blue) lobes. The red and green vertical lines represent the average stimulus and response times. The remaining seven plots show the estimated change-points, one-step ahead marginal posterior point predictions and their 95% predictive intervals for each lobe. The brain activity is reduced roughly at half of the initial phase of the experiment and it increases when the cue is presented. The sharpest increases are detected in the frontal and central lobes, followed by the central-parietal, parietal and temporal lobes. The estimated change in activity in the parietal-occipical and occipital areas is less pronounced.

16
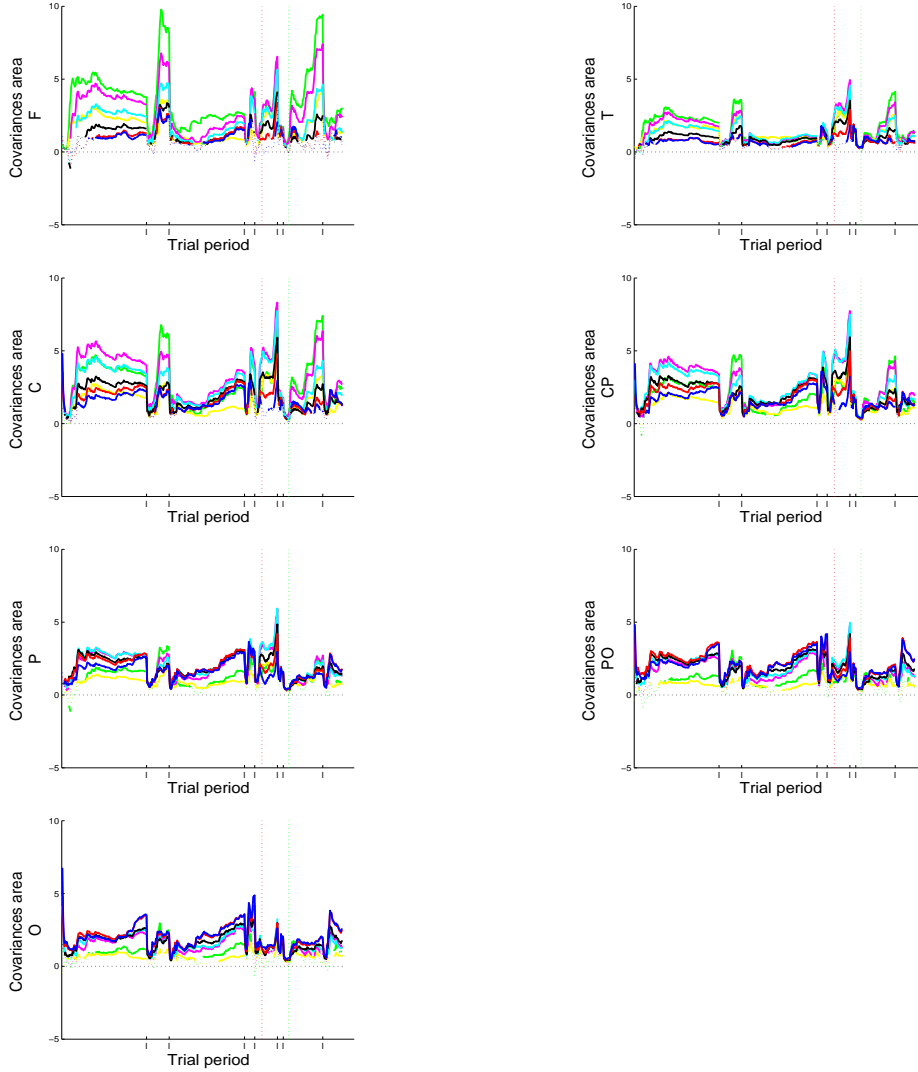
Figure 2: estimates of the time-dependent variance and covariance functions for each brain area. Whole segments represent periods during which their 95% posterior intervals do not intersect zero. All estimated covariances are positive and time-varying, representing different levels of cooperative activity of the seven brain areas over time. The detection of a change-point generally corresponds to an increase in the mutual coordination of their activities. The estimated covariance functions are spatially ordered consistently over time, the strongest relationships being estimated between adjacent brain areas.

17

The data analysed in this section consist of a sequence of 55 binary trials during which a macaque monkey performed a location-scene association task (Wirth et al. [2003]). The learning curve is represented by the time-dependent estimates of the trials'success probabilities. We derived these estimates using a uniform prior for the success probability of the first trial. The transfer prior (3) was implemented using a conjugate Beta prior. The hyper-parameter α was set at 0.7, requiring weak evidence in order to detect a change-point. This hyper-parameter setting was mainly motivated by the limited amount of data available for the analysis. The distribution of the KL statistic under the null hypothesis of no change was approximated using ten thousand Monte Carlo samples from the Beta posterior distribution of each trial's success probability. For this data, the unsmoothed state-space estimates of Smith et al. [2004] suggest that, with 90% confidence, the success probability significantly exceed its chance value 0.25 from trial 24 onwards. Figure 3 shows that our estimates of the success probability are more conservative suggesting that, with 90% conditional posterior probability, learning has taken place from trial 29 onwards.
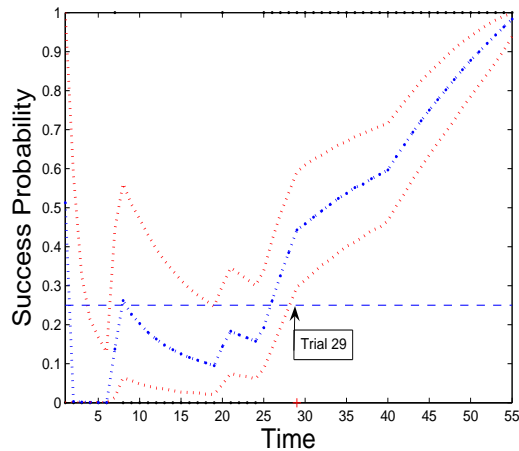

Figure 3: macaque monkey binary data and semi-parametric estimates of their time-dependent success probabilities. The sequence of estimates of the success probabilities describe the macaque's learning curve over time. The first trial at which the learning curve lies above its chance level 0.25, indicating that learning has effectively taken place, is number 29.

18

4 Dynamic modelling of functional neuronal networks

This example illustrates the application of the method presented in section 1 in the context of a model for networks of spiking neurons. Introductions to the neuronal physiology and to neuronal modelling are presented in Fienberg [1974] and Brillinger [1988]. Recent surveys of the state-of-the-art in multiple spike trains modelling can be found in Iyengar [2001], Brown et al. [2004], Kass et al. [2005], Okatan et al. [2005], Rao [2005] and Rigat et al. [2006]. Dynamic point process neuronal models based on fully parametric state-space representations have been proposed by Eden et al. [2004], Truccolo et al. [2005], Brown and Barbieri [2006], Srinivansan et al. [2006] and Eden and Brown [2008].

During the experiments analysed in this section part of the neural activity of a sheep's temporal cortex is observed at discrete times. The goal of the experiments is to investigate the activity of brain areas associated with memory. At each experiment the sheep is shown either a blank screen or two images. In the latter case, a reward is given when one of a set of "familiar faces" is correctly identified. A sequence of 77 disconnected experimental in-vivo multi-electrode array (MEA) recordings is generated (Kendrick et al. [2001]).

4.1 A binary network model

In what follows each element of the experimental time series $\{Y_i\}_{i=1}^{77}$ is $Y_{i,k,t_{i,j(i)}} = 1$ if neuron k fires at time $t_{i,j(i)}$ during trial i and $Y_{i,k,t_{i,j(i)}} = 0$ otherwise with $j(i) = 1, ..., n_i$. We model the joint sampling distribution of the multiple spike data for trial i, Y_i, as a Bernoulli process with renewal (Rigat et al. [2006]). The joint probability of a given realisation y_i is

$$P(Y_i = y_i \mid \pi_i) = \prod_{t=t_{i,1}}^{t_{i,n_i}} \prod_{k=1}^{K} \pi_{i,k,t}^{y_{i,k,t}} (1 - \pi_{i,k,t})^{1-y_{i,k,t}}. \qquad (15)$$

For model (15) to be biologically interpretable, the firing probability of neuron k at time $t_{i,j(i)}$ during trial i, $\pi_{i,k,t_{i,j(i)}}$, is defined as a one-to-one non-decreasing mapping of a real-valued voltage function $v_{i,k,t_{i,j(i)}}$ onto the interval $(0,1)$. The function $v_{i,k,t_{i,j(i)}}$ represents the unnormalised difference of electrical potential across the membrane of neuron k at time $t_{i,j(i)}$. Let $\tau_{i,k,t_{i,j(i)}}$ be the last spiking time of neuron k prior to time $t_{i,j(i)}$ during trial

19

i, that is

$$
\tau_{i,k,t_{i,j(i)}} = \begin{cases} 1 & \text{if } \sum_{\tau=1}^{t_{i,j(i)}} Y_{i,k,\tau} = 0 \text{ or } t_{i,j(i)} = 1, \\ \max\{1 \leq \tau < t_{i,j(i)} : Y_{i,k,\tau} = 1\} & \text{otherwise,} \end{cases}
$$

and the voltage function is modelled as

$$
v_{i,k,t_{i,j(i)}} = \sum_{l=1}^{K} \beta_{i,k,l} \sum_{w=\tau_{i,k,t_{i,j(i)}}}^{t_{i,j(i)}-1} y_{i,l,w}. \tag{16}
$$

The spiking probabilities are linked to (16) via the logistic mapping

$$
\pi_{i,k,t_{i,j(i)}} = \frac{e^{v_{i,k,t_{i,j(i)}}}}{1 + e^{v_{i,k,t_{i,j(i)}}}}.
$$

The coefficients $\beta_{i,k,l}$ represent the strenght of the functional relationship from neuron l to neuron k during trial i. When $\beta_{i,k,l}$ is positive during trial i, the firing activity of neuron l promotes that of neuron k whereas when it is negative firing of l inhibits that of k. When $k = l$, the coefficients $\beta_{i,k,k}$ represent the spontaneous spiking rate of neuron k during trial i. The last summation term in equation (16) indicates that the membrane potential of a neuron is assumed to be influenced only by the spiking activity of the other neurons during its last inter-spike interval.

For each trial $i = 1, ..., N$ we use a Metropolis sampler to produce approximate posterior inferences for the K^2 model parameters. For each experiment, we run a random scan neuron-wise update with independent Gaussian random-walk proposals for twenty-five thousand iterations. The initial prior for the parameters of all experiments is Gaussian with zero mean, standard deviation 1 and zero covariance for all pairs of neurons. Conditionally on the data $y^{0:i}$ and on the current posterior estimates, upon observing the outcome of the ith+1 experiment, y_{i+1}, we use the KL statistic (5) to test whether a significant change occurred in any of the model's parameters. The occurrence of such changes and the corresponding parameter estimates indicate statistically significant variations of different aspects of the neural activity.

4.2 Analysis of sheep multiple spike trains

In this section we analyse the spiking activity of the 7 most active neurons among the 64 recorded cells. The panel on the left in Figure 4 shows the

20

activity of these 7 neurons during all 77 experiments. The panel on the right shows their mean firing rates, which reflect the overall low spiking frequency typical of this type of measurements. Brighter vertical bands mark experiments during which the mean firing rate for all neurons is relatively high. The co-occurrence of these high firing rates suggest that the most prominent connections among the seven neurons are mutually excitatory functional relationships.
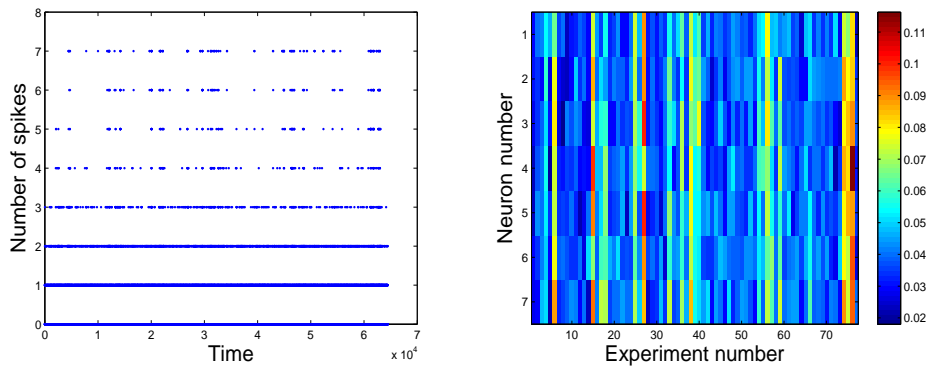

Figure 4: spiking activity (left) and mean firing rates (right) for the 7 most active neurons among the 64 recorded cells. Each dot in the upper panel marks the number of spiking neurons for each millisecond of the 77 experiments. The range of the mean firing rates is $0.02 - 0.12$, reflecting the low overall spiking frequency typical for this type of recordings. Vertical bright bands mark experiments during which the spiking activity of all neurons is relatively high, suggesting that the seven neurons are funciontally connected mostly via mutually excitatory relationships.

Figure 5 displays summaries of the point estimates of the network coefficients β across all experiments. The top-left panel shows the proportion of experiments during which each of the network coefficients were found not significant using their estimated 95% highest posterior probability intervals. The top-right and bottom panels respectively show the proportion of experiments where each coefficient was found either significantly inhibitory or excitatory. Most functional pair-wise interactions are consistently found not significant, whereas the self-dependence coefficients on the main diagonal are always found significant and negative, representing the well-known property of neural refractoriness. The excitatory functional connection from neuron 3 towards neuron 6 is most prominent, being significant over approximately

21

65% of the experiments. Figure 6 illustrates in detail the point estimates and the 95% highest posterior intervals of this pair-wise connection together with those of each neuron's self-dependence and of the connection from neuron 6 towards neuron 3. Consistently with the high proportion of experiments when it is found significant, the excitatory effect of neuron 3 to neuron 6 is the most stable estimate over time.
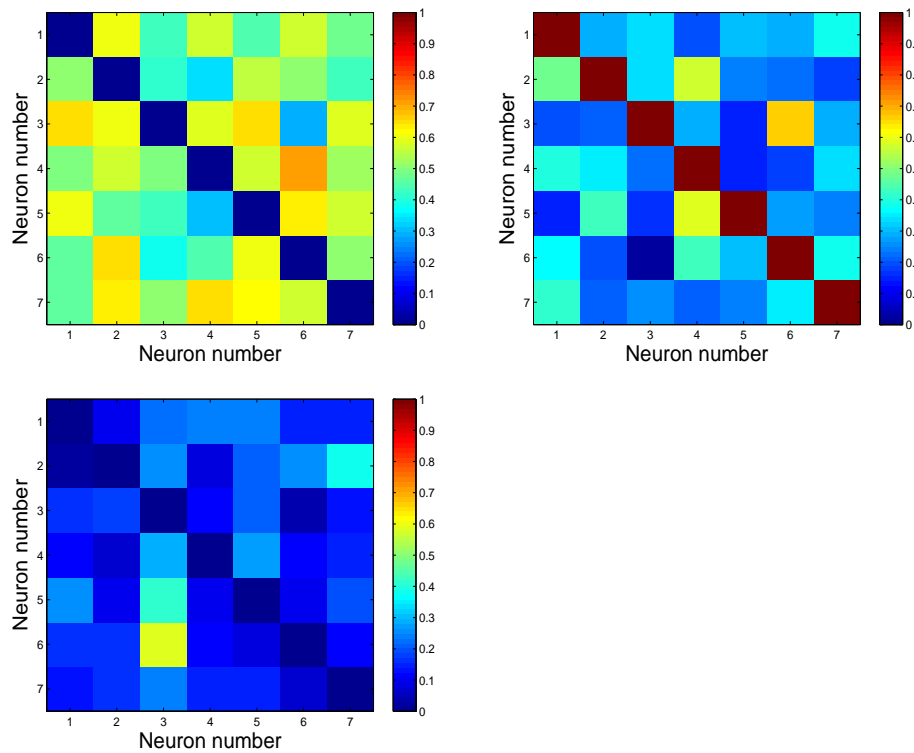

Figure 5: proportion of experiments during which each of the network coefficients were found not significant (top-left), significantly inhibitory (top-right) or excitatory (bottom). Most functional interactions are not significant over most experiments, whereas the self-dependence coefficients are always significant and negative, representing the well-known property of neural refractoriness.
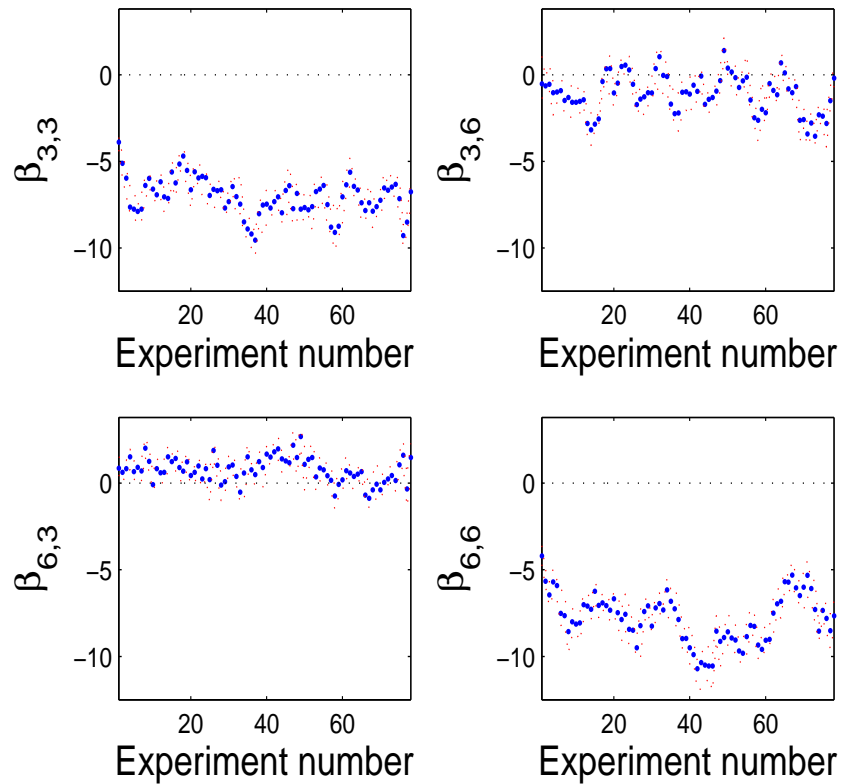
22

Figure 6: point estimates and 95% highest posterior intervals for the self-dependence parameters of neurons 3 and 6 (main diagonal) and of their pair-wise functional connections. The two neurons exhibit a comparable level of refractoriness over all experiments. Consistently with the high proportion of experiments when it is significant, the excitatory effect of neuron 3 to neuron 6 is the most stable over time.

23

5 Discussion

This work is motivated by the difficulties encountered in constructing time series models when the niether the factors driving the dynamics of their parameters nor the relationship between the resolution of the data and such dynamics are known. The semi-parametric method illustrated here provides flexible time-dependent estimates which may then suggest specific evolution dynamics. For exploratory data analyses, such as those presented in sections 2, 3 and 4, these estimates may suffice to address specific scientific questions. Otherwise, appropriate measures of dependence between these time-dependent estimates and experimental factors of interest provide a principled basis for more precise formulations of the parameters' dynamics. Describing the exact form of such dependence measures is very much context-dependent and it lies outside of the scope of this work.

A distinctive feature of the modeling approach proposed here is that it combines elements of sequential Bayesian learning and conditional frequentist inference along the lines of Guttman [1967], Box [1980], Berger et al. [1994], Meng [1994], Gelman et al. [1996], Berger and Bayarri [1997], Spiegelhalter et al. [2002], Bayarri and Morales [2003], Kuhnert et al. [2003] and Bayarri and Berger [2004] among others. A general treatment of such pragmatic combination of frequentist and Bayesian ideas for model criticism can be found in Chapter 8 of O'Hagan and Forster [1999]. From this perspective, our method is a "Bayesianly justifiable" procedure (Rubin [1984]) because only those future data that are consistent with the current conditional posterior distribution of the model parameters are relevant for approximating the distribution of the change-point statistic.

The latter statistic reflects a notion of a change-point as an observation which, on the basis of the chosen model with its prior and the observations so far, is "surprising" from a predictive point of view. Note that this characterization does not depend on the parametrization of the state space nor on the unobservable sample paths of latent states, but it depends only on the predictives on observables. Defining models and their properties via their one step ahead predictive statements but has been recommended, among others, by Geisser and Eddy [1979] for predictive model selection, Dawid [1984] in his prequential inference, San Martini and Spezzaferri [1984] for model selection, West and Harrison [1986b] for monitoring the adequacy of Bayesian forecasting models and by Smith [1992] for comparing the characteristics of different forecasting models. More recently, optimal predictive model selection criteria have been proposed by Barbieri and Berger [2004].

In this work, one change-point process common to all model's param-

eters is used to define their conditional posterior distribution. Should the data provide evidence of changes of only some parameters, the posterior distributions for the unchanging coefficients would not reflect an efficient use of the data. It is important to note that while in principle any subset of model parameters can be associated to a distinct change-point process, the limitations for implementing multivariate change-point process inference within our framework are eminently practical. This is because marginal likelihoods for each subset of model parameters having a different change-point process are required to approximate the distribution of their change-point test statistic. For classes of models where marginal likelihoods are available in closed form, this work can be extended by introducing a random variable identifying groups of coefficients sharing a common change-point process.

Posterior simulation via Markov chain Monte Carlo algorithms has been used in this work to fit multivariate time series models and to approximate critical values of the KL statistic. Although the current implementation of our method is operationally realistic, these computationally intensive methods are in fact rather impractical for an iterative process of model formulation and criticism. Currently two directions are being pursued to improve the computational efficiency of our method. On the one hand, faster resampling methods such as particle filters (Doucet et al. [2001]) and approximate Bayesian computation (Marjoram et al. [2003]) can be adopted. Alternatively, analytical posterior approximations can be adopted (Tierney and Kadane [1986]). For instance, in the context of sequential time series modeling Koyama et al. [2008] recently proposed a Laplace-Gauss posterior approximation that obviates the use of cumbersome resampling techniques.

Acknowledgements

References

H. Akaike. Likelihood of a model and information criteria. *Journal of Econometrics*, 16:3–14, 1981.

H. Akaike. On the likelihood of a time series model. *The Statistician*, 27:217–235, 1978.

J.H. Albert and S. Chib. Bayes inference via Gibbs sampling for autoregressive time series subject to Markov mean and variance shifts. *Journal of Business and Economic Statistics*, 11:1 – 15, 1993.

M.M. Barbieri and J.O. Berger. Optimal predictive model selection. *The Annals of Statistics*, 32:870–897, 2004.

M.J. Bayarri and J.O. Berger. The interplay between Bayesian and frequentist analysis. *Statistical Science*, 19:58 – 80, 2004.

M.J. Bayarri and J. Morales. Bayesian measures of surprise for outlier detection. *Journal of Statistical Planning and Inference*, 111:3 – 22, 2003.

P. Bélisle, L. Joseph, B. MacGibbon, D.B. Wolfson, and R. du Berger. Change-point analysis of neuron spike train data. *Biometrics*, 54:113 – 123, 1998.

T. Bengtsson and J.E. Cavanaugh. An improved Akaike information criterion for state-space model selection. *Computational Statistics and Data Analysis*, 50:2635 – 2654, 2006.

J.O. Berger and M.L. Bayarri. Measures of surprise in bayesian analysis. *ISDS Discussion Paper*, 46, 1997.

J.O. Berger, L. Brown, and R.L. Wolpert. A unified conditional frequentist and Bayesian test for fixed and sequential hypothesis testing. *The Annals of Statistics*, 22:1787 – 1807, 1994.

J. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7: 686–690, 1979.

J.M. Bernardo and A.F.M. Smith. *Bayesian theory*. Wiley, 2007.

G. E. P. Box. Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the American Statistical Association A*, 143:383–430, 1980.

D.R. Brillinger. Some statistical methods for random processes data from seismology and neurophysiology. *The Annals of Statistics*, 16:1–54, 1988.

E. Brown, R.E. Kass, and P.P. Mitra. Multiple neural spike train data analysis: state-of-the-art and future challenges. *Nature Neuroscience*, 7:456–461, 2004.

E.N. Brown and R. Barbieri. Dynamic analyses of neural representations using the state-space modeling paradigm. *In: Madras, B., Von Zastrow, M. Colvis, C., Rutter, J. Shurtleff, D. and Pollock, J. eds. "The Cell Biology of Addiction", New York, Cold Spring Harbor Laboratory press*, 2006.

G. Buzsáki. Large scale recording of neuronal ensembles. *Nature Neuroscience*, 7: 446–451, 2004.

O. Cappe, E. Moulines, and T. Ryden. *Inference in Hidden Markov Models.* Springer, 2005.

B.P. Carlin, A.E. Gelfand, and A.F.M. Smith. Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, 41:389–405, 1992.

C. Carota, G. Parmigiani, and N. Polson. Diagnostic measures for model criticism. *Journal of the American Statistical Association*, 91:753–762, 1996.

S. Chib. Estimation and comparison of multiple change-point models. *Journal of Econometrics*, 86:221 – 241, 1998.

F. Critchley, P. Marriott, and M. Salmon. Preferred point geometry and the local differential geometry of the Kullback-Leibler divergence. *The Annals of Statistics*, 22:1587–1602, 1994.

A.P. Dawid. Present position and potential developments: some personal views. statistical theory: the prequential approach. *Journal of the Royal Statistical Society A*, 147:278–292, 1984.

A. Delorme, S. Makeig, M. Fabre-Thorpe, and T. J. Sejnowski. From single trial EEG to brain area dynamics. *Neurocomputing*, 44:1057–1064, 2002.

P. Diaconis and D. Ylvisaker. Conjugate priors for exponential families. *The Annals of Statistics*, 7:269–281, 1979.

A. Doucet, N. De Freitas, and N.J. Gordon. *Sequential Monte Carlo Methods in Practice.* Springer-Verlag, New York, 2001.

U.T. Eden and E.N. Brown. Continuous-time filters for state estimation from point process models of neural data. *Statistica Sinica*, 2008.

U.T. Eden, L.M. Frank, R. Barbieri, V. Solo, and E.N. Brown. Dynamic analysis of neural encoding by point process adaptive filtering. *Neural Computation*, 16: 971–998, 2004.

P. Fearnhead and Z. Liu. On-Line Inference for Multiple Change Points. *Journal of the Royal Statistical Society B*, 69:589–605, 2007.

D. Ferger. Nonparametric tests for nonstandard change-point problems. *The Annals of Statistics*, 23:1848–1861, 1995.

S. E. Fienberg. Stochastic models for single neuron firing trains: a survey. *Biometrics*, 30:399–427, 1974.

S. Frühwirth-Shnatter. Bayesian model discrimination and Bayes factors for linear gaussian state-space models. *Journal of the Royal Statistical Society B*, 1:237 – 246, 1995.

S. Frühwirth-Shnatter. Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *Journal of the American Statistical Association*, 96:194 – 209, 2001.

S. Frühwirth-Shnatter. *Finite Mixture and Markov Switching Models*. Springer, 2006.

D. Gamerman. A dynamic approach to the statistical analysis of point processes. *Biometrika*, 79:39–50, 1992.

S. Geisser and W.F. Eddy. A predictive approach to model selection. *Journal of the American Statistical Association*, 74:153 – 160, 1979.

A.E. Gelfand and A. F. M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85:398–409, 1990.

A. Gelman, X.L. Meng, and H.S. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–807, 1996.

Z. Ghahramani and G.E. Hinton. Variational learning for switching state-space models. *Neural Computation*, 12:831 – 864, 2000.

C. Goutis and C. Robert. Model choice in generalised linear models: A Bayesian approach via Kullback-Leibler projections. *Biometrika*, 85:29–37, 1998.

E. Gutiérrez-Peña. Moments for the canonical parameter of an exponential family under a conjugate distribution. *Biometrika*, 84:727–732, 1997.

I. Guttman. The use of the concept of a future observation in goodness-of-fit problems. *Journal of the Royal Statistical Society B*, 29:83 – 100, 1967.

P. Hall. On Kullback-Leibler loss and density estimation. *The Annals of Statistics*, 15:1491–1519, 1987.

J.D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

J.D. Hamilton. Analysis of time series subject to changes in regime. *Journal of Econometrics*, 45:39 – 70, 1990.

C. Han and B.P. Carlin. Markov chain Monte Carlo methods for computing Bayes factors: A comparative review. *Journal of the American Statistical Association*, 96:1122 – 1132, 2001.

W. Härdle, H. Lütkepohl, and R. Chen. A review of nonparametric time series analysis. *International Statistical Review*, 65:49 – 72, 1997.

P.J. Harrison and C.F. Stevens. Bayesian forecasting. *Journal of the Royal Statistical Society B*, 38:205–247, 1976.

T. Hastie. A closer look at the deviance. *The American Statistician*, 41:16–20, 1987.

S. Iyengar. The analysis of multiple neural spike trains. *In: Advances in Methodological and Applied Aspects of probability and Statistics; N. Bolakrishnan editor; Gordon and Breack*, pages 507–524, 2001.

M. Jain, M. Elhilali, N. Vaswami, J. Fritz, and S. Shamma. A particle filter for tracking adaptive neural responses in auditory cortex. *Submitted for publication*, 2007.

R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

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

K.M. Kendrick, A.P. da Costa, A.E. Leigh, M.R. Hinton, and J.W. Peirce. Sheep don't forget a face. *Nature*, 414:165–166, 2001.

C.J. Kim. Dynamic linear models with Markov-switching. *Journal of Econometrics*, 60:1 – 22, 1994.

S. Koyama, L. Castellanos Perez-Bolde, and R.E. Kass. Approximate Methods for State-Space Models: The Laplace-Gaussian Filter. *Submitted for publication*, 2008.

P.M. Kuhnert, K. Mergesen, and P. Tesar. Bridging the gap between different statistical approaches: an integrated framework for modelling. *International Statistical Review*, 71:335 – 368, 2003.

S. Kullback. *Information theory and Statistics*. Dover; New York, 1997.

S. Kullback and R.A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86, 1951.

D. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27:986–1005, 1956.

C.R. Loader. Change point estimation using nonparametric regression. *The Annals of Statistics*, 24:1667–1678, 1996.

S. Makeig, M. Westerfield, T. P. Jung, S. Enghoff, J. Towsend, E. Courchesne, and T. J. Sejnowski. Dynamic brain sources of visual evoked responses. *Science*, 295: 690–694, 2002.

P. Marjoram, J. Molitor, V. Plagnol, and S. Tavaré. Markov Chain Monte Carlo without likelihoods. *Proceedings of the National Academy of Science*, 100:15324 – 15328, 2003.

R.E. McCulloch. Information and the likelihood function in exponential families. *The American Statistician*, 42:73–75, 1988.

R.E. McCulloch and R.S. Tsay. Statistical analysis of economic time series via Markov switching models. *Journal of Time Series Analysis*, 15:523 – 539, 1994.

X.L. Meng. Posterior predictive p-values. *The Annals of Statistics*, 22:1142–1160, 1994.

A. Mira and S. Petrone. Bayesian hierarchical nonparametric inference for change point problems. *Bayesian Statistics*, 5:693–703, 1996.

H.G. Muller. Change-points in nonparametric regression analysis. *The Annals of Statistics*, 20:737–761, 1992.

T. O'Hagan and J. Forster. *Kendall's Advanced Theory of Statistics*, volume 2B. Arnold, 1999.

M. Okatan, M.A. Wilson, and E.N. Brown. Analyzing functional connectivity using a network likelihood model of ensemble neural spiking activity. *Neural computation*, 17:1927–1961, 2005.

E.S. Page. A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–527, 1955.

R.P.N. Rao. Hierarchical Bayesian inference in networks of spiking neurons. *Advances in NIPS; MIT press*, 17, 2005.

F. Rigat, M. de Gunst, and J. ven Pelt. Bayesian modelling and analysis of spatio-temporal neuronal networks. *Bayesian Analysis*, 1:733–764, 2006.

C.P. Robert, G. Celeux, and J. Diebolt. Bayesian estimation of hidden Markov chains: a stochastic implementation. *Statistics and Probability Letters*, 16:77 – 83, 1993.

P.M. Robinson. Non-parametric estimation for time series models. *Journal of Time Series Analysis*, 4:185 – 208, 1983.

D.B. Rubin. Bayesianly justifiable and relevant frequency calclulations for the applied Statistician. *Annals of Statistics*, 12:1151 – 1172, 1984.

A. San Martini and F. Spezzaferri. A predictive model selection criterion. *Journal of the Royal Statistical Society B*, 46:296–383, 1984.

R.H. Shumway and D.S. Stoffer. Dynamic linear models with switching. *Journal of the American Statistical Association*, 86:763–769, 1991.

A. F.M. Smith and G.O. Roberts. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society B*, 55:3–23, 1993.

A.C. Smith, M.F. Loren, S. Wyrth, M. Yanike, D. Hu, Y. Kubota, A.M. Graybiel, W.A. Suzuki, and E.M. Brwon. Dynamic analysis of learning in behavioural experiments. *Journal of Neuroscience*, 24:447–461, 2004.

A.F.M. Smith. A Bayesian approach to inference about a change-point in a sequence of random variables. *Biometrika*, 62:407–416, 1975.

J.Q. Smith. Non-linear state space models with partially specified distributions on states. *Journal of Forecasting*, 9:137 – 149, 1990.

J.Q. Smith. A comparison of the characteristics of some Bayesian forecasting models. *International Statistical Reviews*, 60:75 – 85, 1992.

D.J. Spiegelhalter, N.G. Best, P.B. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, 64:583–639, 2002.

L. Srinivansan, U.T. Eden, A.S. Willsky, and E.N. Brown. A state-space analysis for reconstruction of goal-directed movements using neural signals. *Neural Computation*, 18:2465–2494, 2006.

D.A. Stephens. Bayesian retrospective multiple-changepoint identification. *Applied Statistics*, 43:159–178, 1994.

M. Stone. Application of a measure of information to the design and comparison of regression experiments. *The Annals of Mathematical Statistics*, 30:55–70, 1959.

L. Tierney. Markov chains for exploring posterior distributions. *The Annals of Statistics*, 22:1701–1762, 1994.

L. Tierney and J.B. Kadane. Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81:84 – 86, 1986.

W. Truccolo, U.T. Eden, M.R. Fellows, J.P. Donoghue, and E.N. Brown. A point process framework for relating neural spiking activity to spiking history, neural ensemble and extrinsic covariate effects. *Journal of Neurophysiology*, 93:1074–1089, 2005.

M. West. Bayesian model monitoring. *Journal of the Royal Statistical Society B*, 48:70–78, 1986.

M. West and P.J. Harrison. *Bayesian forecasting and dynamic models*. Springer; New York, second edition, 1997.

M. West and P.J Harrison. Monitoring and adaptation in Bayesian forecasting models. *Journal of the American Statistical Association*, 81:741–750, 1986a.

M. West and T.J. Harrison. Monitoring and adaptation in Bayesian forecasting models. *Journal of the American Statistical Association*, 81:741–50, 1986b.

M. West, P.J. Harrison, and H.S. Migon. Dynamic generalised linear models and Bayesian forecasting. *Journal of the American Statistical Association*, 80:73–83, 1985.

S. Wirth, M. Yanike, M.F. Loren, A.C. Smith, E.M. Brwon, and W.A. Suzuki. Single neurons in the monkey hyppocampus and learning of new associations. *Science*, 300:1578–1584, 2003.