Sequential Bayesian Semi-Parametric Binary Classification

†Christophe Andrieu - *Nando de Freitas - ‡Arnaud Doucet

†Department of Mathematics, Statistics Group,
University of Bristol,
University Walk, Bristol BS8 1TW, U.K.
Email: `C.Andrieu@bristol.ac.uk`

*Computer Science Division,
UC Berkeley,
387 Soda Hall, Berkeley, CA 94720-1776 USA.
Email: `jfgf@cs.berkeley.edu`

‡Department of Electrical and Electronic Engineering,
University of Melbourne,
Parkville, Victoria 3052, Australia.
Email: `doucet@ee.mu.oz.au`

DRAFT: June 2001

Abstract

In this paper, we present a dynamic Bayesian model for time-varying semiparametric binary classification. The posterior distribution evolves over time through the accumulation of data and needs to be estimated on-line. This is a complex optimal filtering problem. Recently, a large number of algorithms based on particle filtering methods, also known as sequential Monte Carlo (SMC), have appeared in the literature to solve this class of problems (see (Doucet, de Freitas and Gordon 2001)). We propose here an efficient particle filtering method which uses random mixtures of normal distributions to represent the posterior distributions. It is based on a variance reduction idea that is shown to lead to remarkable improvements over standard algorithms. We demonstrate our methodology on a simple example and on the problem of fault detection of dynamical operated marine diesel engines.

Keywords: Bayesian estimation; filtering; particle filtering; semi-parametric classification; sequential Monte Carlo; state space models.

1 Introduction

Sequential classification problems arise in a few areas of technology, including condition monitoring and real-time decision systems (Melvin 1996, Penny, Roberts, Curran and Stokes 1999). For example, when monitoring patients, we might wish to decide whether they require an increase in drug intake at several intervals in time. In the past, we addressed the problem of sequential fault detection in marine diesel engines (Højen-Sørensen, de Freitas and Fog 2000). This application is of great importance as early detection of incipient faults can improve safety and efficiency, as well as, help to reduce down-time and plant maintenance in many industrial and transportation environments.

In this article we present a dynamic (Bayesian) model for time-varying semi-parametric binary classification. The classifier is described by a linear combination of basis functions whose coefficients evolve according to a smoothness prior (Kitagawa and Gersch 1996). Our model is generic and any bases can be used, depending on the application: neural networks (Andrieu, de Freitas and Doucet 1999a), splines (Holmes and Denison 2000) or wavelets (Holmes and Mallick 2000) for example. Performing Bayesian inference in this context is a computationally challenging problem because the posterior distribution does not admit a closed form expression and, in addition, it evolves over time. The estimation of the posterior distribution at a given time could be undertaken with Markov chain Monte Carlo (MCMC) methods (Robert and Casella 1999). However, for practical reasons, we are interested in developing algorithms to process data arriving on-line. This is a complex optimal nonlinear filtering problem.

Many approximation algorithms, such as the extended Kalman filter and Gaussian sum approximations, have been proposed to surmount this problem (Anderson and Moore 1979, West and Harrison 1997). However, in many realistic problems, these approximating methods are notoriously unreliable and faults are difficult to diagnose on-line. Recently there has been a surge of interest in sequential Monte Carlo (SMC) methods (also known as particle filtering when the objective of the analysis is to compute the filtering distribution) for nonlinear/non-Gaussian time series analysis (Doucet et al. 2001). These methods, initiated in (Gordon, Salmond and Smith 1993) and (Kitagawa and Gersch 1996) utilize a random sample (or particle) based representation of the posterior probability distributions. The particles are propagated over time using a combination of sequential importance sampling and resampling steps; see (Doucet, Barat and Duvaut 1995) and (Isard and Blake 1996) for early related work.

One could apply standard SMC to carry out the sequential classification task; some results are reported in (Højen-Sørensen et al. 2000). However, this algorithm is generic and does not use all the salient structure of the model under consideration. Our algorithm is based on the introduction of an artificial latent variable that allows us to transform the classification problem to a Gaussian regression problem, where we can make efficient use of a marginalisation technique known as Rao-Blackwellisation (Robert and Casella 1999, pp. 116-119). The marginalisation is carried out using Kalman filtering methods. The Kalman filter has been previously used to improve the efficiency of particle filtering (Bergman, Doucet and Gordon 2001, Doucet, Godsill and Andrieu 2000). The algorithm presented here extends to analysis to more complex models. Our simulations demonstrate that it improves the performance of standard methods.

The rest of the paper is organised as follows. Section 2 describes the statistical model and the estimation objectives. In Section 3, we give a detailed description of our particle filtering method. Section 4 demonstrates the performance of the proposed algorithm.

2 Model Specification and Estimation Objectives

2.1 Statistical Model

Let us consider the following binary classification model. Given at time $t = 1, 2, \ldots$ an input x_t we observe $z_t \in \{0, 1\}$ such that

$$\Pr\left(z_t = 1 | \, x_t, \beta_t \right) = \Phi\left(f\left(x_t, \beta_t \right) \right), \tag{1}$$

where $\Phi\left(u\right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{u} \exp\left(-a^2/2\right) da$ is the cumulative function of the standard normal distribution. This is the so-called probit link. By convention, researchers tend to adopt a logistic (sigmoidal) link function $\varphi\left(u\right) = \left(1 + \exp\left(-u\right)\right)^{-1}$. However, from a Bayesian computational point of view, the probit link has many advantages and is equally valid. The unknown (scalar) function is modeled as

$$f\left(x_t, \beta_t\right) = \sum_{k=1}^{K} \beta_{t,k} \Psi_k\left(x_t\right) = \Psi^{\mathrm{T}}\left(x_t\right) \beta_t,$$

where we have assumed that the basis functions $\Psi\left(x_t\right) \triangleq \left(\Psi_1\left(x_t\right), \ldots, \Psi_K\left(x_t\right)\right)^{\mathrm{T}}$ do not depend on unknown parameters; see (Andrieu, de Freitas and Doucet 1999b) for the more general case. $\beta_t \triangleq \left(\beta_{t,1}, \ldots, \beta_{t,K}\right)^{\mathrm{T}} \in \mathbb{R}^K$ is a set of unknown time-varying regression coefficients. To complete the model, we assume that the regression coefficients satisfy

$$\beta_t = A_t \beta_{t-1} + B_t v_t, \ \beta_0 \sim \mathcal{N}\left(m_0, P_0\right) \tag{2}$$

where $v_t \overset{i.i.d.}{\sim} \mathcal{N}\left(0, I_{n_v}\right)$. Typically K is rather large, say 10 or 100, and the basis functions $\Psi_k\left(\cdot\right)$ are multivariate splines, wavelets or radial basis functions (Holmes and Denison 2000, Holmes and Mallick 2000).

2.2 Augmented Statistical Model

We shall artificially augment the original probabilistic model. This augmentation has several important practical consequences from an algorithmic point of view, and will be detailed in the next section. In particular, we augment the model by introducing the set of independent variables y_t, such that

$$y_t = f\left(x_t, \beta_t\right) + n_t, \tag{3}$$

where $n_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Next, we define

$$z_t = \begin{cases} 1 & \text{if } y_t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to check that one has

$$\Pr(z_t = 1 | x_t, \beta_t) = \Pr(y_t > 0 | x_t, \beta_t) = \Pr(n_t > -f(x_t, \beta_t)) = \Phi(f(x_t, \beta_t)).$$

That is, we have defined a hierarchical model such that

$$\int \Pr(z_t = 1 | x_t, \beta_t, y_t)\, p(y_t | x_t, \beta_t)\, dy_t = \int \Pr(z_t = 1 | y_t)\, p(y_t | x_t, \beta_t)\, dy_t = \Phi(f(x_t, \beta_t)).$$

This data augmentation trick was first introduced in econometrics by Daniel McFadden (Gouriéroux and Monfort 1996, McFadden 1989)[1]. In the MCMC context, it has been adopted by several authors to design efficient samplers (Albert and Chib 1993, Holmes and Denison 2000, Robert and Casella 1999, Wood and Kohn 1998). Here, we will show how to take advantage of this trick in a sequential Monte Carlo setting.

2.3 Estimation objectives

Given, at time t, the observations $o_{1:t} \triangleq (x_{1:t}, z_{1:t})$, any Bayesian inference is based on the posterior distribution[2] $P(d\beta_{0:t} | o_{1:t})$. We are, therefore, interested in estimating sequentially in time this distribution and some of its features, such as $\mathbb{E}(f(x_t, \beta_t) | o_{1:t})$ or the marginal predictive distribution at time t for new input data x_{t+1}

$$\Pr(z_{t+1} = 1 | o_{1:t}, x_{t+1}) = \int \Pr(z_{t+1} = 1 | o_{1:t}, x_{t+1}, \beta_{t+1})\, p(\beta_{t+1} | \beta_t)\, p(\beta_{0:t} | o_{1:t})\, d\beta_{0:t+1}.$$

The posterior density satisfies the following recursion

$$p(\beta_{0:t} | o_{1:t}) = p(\beta_{0:t-1} | o_{1:t-1}) \frac{\Pr(z_t | x_t, \beta_t)\, p(\beta_t | \beta_{t-1})}{\Pr(z_t | o_{1:t-1}, x_t)}.$$

It is clear that this multivariate density does not admit an analytical expression and, consequently, we need to resort to numerical methods to approximate it.

3 Sequential Bayesian Estimation via Particle Filtering

A straightforward application of SMC methods to the model (1)-(2) would focus on sampling from the high-dimensional distribution $P(d\beta_{0:t} | o_{1:t})$ (Højen-Sørensen et al. 2000). A substantially more efficient strategy is to exploit the augmentation of the model to sample only from the low-dimensional distribution $P(dy_{1:t} | o_{1:t})$. The low-dimensional samples allow us then to compute the remaining estimates analytically, as shown in the following subsection. In Section 3.2 we present the SMC algorithm to sample the augmentation variable $y_{1:t}$.

[1]Daniel McFadden was awarded the Nobel Prize of Economics in 2000 for its contribution to the analysis of related econometrics models, where this trick is used extensively.

[2]For any θ, we use $P(d\theta_{0:t} | o_{1:t})$ to denote the distribution and $p(\theta_{0:t} | o_{1:t})$ to denote the density, where $P(d\theta_{0:t} | o_{1:t}) = p(\theta_{0:t} | o_{1:t})\, d\theta_{0:t}$. Also, $\theta_{0:t} \triangleq \{\theta_0, \theta_1, \ldots, \theta_t\}$.

3.1 Marginalisation

Consider the extended model defined by equations (1)-(2)-(3). One has

$$p\left(\beta_{0:t}\right|o_{1:t}) = \int p\left(\beta_{0:t}\right|x_{1:t}, y_{1:t})\, p\left(y_{1:t}\right|o_{1:t})\, dy_{1:t}.$$

Thus if one applies SMC to obtain an approximation of $P\left(dy_{1:t}\right|o_{1:t}) = p\left(y_{1:t}\right|o_{1:t})\, dy_{1:t}$ of the form

$$\widehat{P}_N\left(dy_{1:t}\right|o_{1:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{y_{1:t}^{(i)}}\left(dy_{1:t}\right),$$

then $p\left(\beta_{0:t}\right|o_{1:t})$ can be approximated via

$$\begin{aligned}
\widehat{p}_N\left(\beta_{0:t}\right|o_{1:t}) &= \int p\left(\beta_{0:t}\right|x_{1:t}, y_{1:t}) \sum_{i=1}^{N} w_t^{(i)} \delta_{y_{1:t}^{(i)}}\left(y_{1:t}\right) \\
&= \sum_{i=1}^{N} w_t^{(i)} p\left(\beta_{0:t}\middle| x_{1:t}, y_{1:t}^{(i)}\right),
\end{aligned}$$

that is a mixture of Gaussians. From this approximation, one can estimate $\mathbb{E}\left(\beta_t\right|x_{1:t}, y_{1:t})$ and $\mathbb{E}\left(\beta_{t-L}\right|x_{1:t}, y_{1:t})$. For example, an estimate of $\mathbb{E}\left(\beta_t\right|x_{1:t}, y_{1:t})$ is given by

$$\begin{aligned}
\widehat{\mathbb{E}}_N\left(\beta_t\right|o_{1:t}) &= \int \beta_t \widehat{p}_N\left(\beta_{0:t}\right|o_{1:t})\, d\beta_{0:t} \\
&= \sum_{i=1}^{N} w_t^{(i)} \mathbb{E}\left\{\beta_t\middle| x_{1:t}, y_{1:t}^{(i)}\right\},
\end{aligned} \tag{4}$$

where $\mathbb{E}\left\{\beta_t\middle| x_{1:t}, y_{1:t}^{(i)}\right\}$ is computed through the Kalman filter associated to the linear Gaussian state-space model defined by (2)-(3). Similarly, an estimate of the predictive distribution is given by

$$\begin{aligned}
\widehat{\Pr}_N\left(z_{t+1} = 1\right|o_{1:t}, x_{t+1}) &= \int \Pr\left(z_{t+1} = 1\right|y_{t+1})\, \widehat{P}_N\left(dy_{1:t+1}\right|o_{1:t}, x_{t+1}) \\
&= \sum_{i=1}^{N} w_t^{(i)} \Pr\left(z_{t+1} = 1\middle| y_{t+1}^{(i)}\right) \\
&= \sum_{i=1}^{N} w_t^{(i)} \mathbb{I}_{(0,+\infty)}\left(y_{t+1}^{(i)}\right),
\end{aligned} \tag{5}$$

where $y_{t+1}^{(i)} \sim p\left(y_{t+1}\middle| x_{1:t+1}, y_{1:t}^{(i)}\right)$. This shows that we can restrict ourselves to the estimation of $p\left(y_{1:t}\right|o_{1:t})$ to perform inference.

In the SMC framework, we must estimate the "target" density $p\left(y_{1:t}\right|o_{1:t})$ pointwise up to a normalizing constant. By standard factorisation, one has

$$p\left(y_{1:t}\right|o_{1:t}) \propto \prod_{k=1}^{t} \Pr\left(z_k\right|y_k)\, p\left(y_k\right|x_{1:k}, y_{1:k-1}),$$

where $p(y_1|\, y_{1:0}, x_{1:0}) \triangleq p(y_1|\, x_1)$. Since $\Pr(z_k|\, y_k)$ is known, we only need to estimate $p(y_k|\, x_{1:k}, y_{1:k-1})$ up to a normalizing constant. This predictive density can be computed using the Kalman filter. Given $(x_{1:k}, y_{1:k-1})$, the Kalman filter equations are the following. Set $\beta_{0|0} = m_0$ and $\Sigma_{0|0} = \Sigma_0$, then for $t = 1, ..., k-1$ compute

$$\begin{aligned}
\beta_{t|t-1} &= A_t \beta_{t-1|t-1} \\
\Sigma_{t|t-1} &= A_t \Sigma_{t-1|t-1} A_t^{\mathsf{T}} + B_t B_t^{\mathsf{T}} \\
S_t &= \Psi^{\mathsf{T}}(x_t)\, \Sigma_{t|t-1} \Psi(x_t) + 1 \\
y_{t|t-1} &= \Psi^{\mathsf{T}}(x_t)\, \beta_{t|t-1} \\
\beta_{t|t} &= \beta_{t|t-1} + \Sigma_{t|t-1} \Psi(x_t)\, S_t^{-1} \left(y_t - y_{t|t-1}\right) \\
\Sigma_{t|t} &= \Sigma_{t|t-1} - \Sigma_{t|t-1} \Psi(x_t)\, S_t^{-1} \Psi^{\mathsf{T}}(x_t)\, \Sigma_{t|t-1},
\end{aligned} \tag{6}$$

where $\beta_{t|t-1} \triangleq \mathbb{E}(\beta_t|\, x_{1:t-1}, y_{1:t-1})$, $\beta_{t|t} \triangleq \mathbb{E}(\beta_t|\, x_{1:t}, y_{1:t})$, $y_{t|t-1} \triangleq \mathbb{E}(y_t|\, x_{1:t}, y_{1:t-1})$, $\Sigma_{t|t-1} \triangleq cov(\beta_t|\, x_{1:t-1}, y_{1:t-1})$, $\Sigma_{t|t} \triangleq cov(\beta_t|\, x_{1:t}, y_{1:t})$ and $S_t \triangleq cov(y_t|\, x_{1:t}, y_{1:t-1})$. One obtains

$$p(y_k|\, x_{1:k}, y_{1:k-1}) = \mathcal{N}\left(y_k; y_{k|k-1}, S_k\right). \tag{7}$$

3.2 Sampling Algorithm

Assume that at time $t-1$ we have N paths (particles) $\left\{y_{1:t-1}^{(i)}\right\}_{i=1}^{N}$ distributed according to $P(dy_{1:t-1}|\, o_{1:t-1})$ from which one can get the following empirical distribution approximation

$$P_N(dy_{1:t-1}|\, o_{1:t-1}) = \frac{1}{N} \sum_{i=1}^{N} \delta_{y_{1:t-1}^{(i)}}(dy_{1:t-1}).$$

Various SMC methods can be used to obtain N new paths $\left\{y_{1:t}^{(i)}\right\}_{i=1}^{N}$ distributed approximately according to $P(dy_{1:t}|\, o_{1:t})$. The most successful of these methods typically combine importance sampling and a selection scheme. Their asymptotic convergence ($N \to \infty$) is satisfied under mild conditions; see (Crisan, Del Moral and Lyons 1999, Crisan and Doucet 2000) for details.

3.2.1 Importance sampling step

The distributions $P(dy_{1:t-1}|\, o_{1:t-1})$ and $P(dy_{1:t}|\, o_{1:t})$ are of different dimension. We first modify and extend the current paths $y_{1:t-1}^{(i)}$ to obtain new paths $\widetilde{y}_{1:t}^{(i)}$. As our goal is to design a sequential procedure, we set $\widetilde{y}_{1:t}^{(i)} = \left(y_{1:t-1}^{(i)}, \widetilde{y}_t^{(i)}\right)$ where

$$\widetilde{y}_t^{(i)} \sim Q_t\left(dy_t|\, o_{1:t}, \widetilde{y}_{1:t-1}^{(i)}\right),$$

$Q_t(dy_t|\, o_{1:t}, y_{1:t-1}) (= q_t(y_t|\, o_{1:t}, y_{1:t-1})\, dy_t)$ is a conditional distribution. The aim of this distribution is to obtain new paths $\widetilde{y}_{1:t}^{(i)}$ whose density $q_t(y_{1:t})$ is as "close" as possible to $p(y_{1:t}|\, o_{1:t})$. In practice, it is not possible to have $q_t(y_{1:t}) = p(y_{1:t}|\, o_{1:t})$, so we have to weight the new paths to obtain consistent estimates. We perform this "correction" through importance sampling. Assume that the importance weight is well-defined,

$$w(y_{1:t}) = \frac{p(y_{1:t}|\, o_{1:t})}{q_t(y_{1:t})} = \frac{p(y_{1:t-1}|\, o_{1:t})}{p(y_{1:t-1}|\, o_{1:t-1})} \frac{p(y_t|\, o_{1:t}, y_{1:t-1})}{q_t(y_t|\, o_{1:t}, y_{1:t-1})} \tag{8}$$
$$\propto \frac{\Pr(z_t|\, y_t)\, p(y_t|\, x_{1:t}, y_{1:t-1})}{q_t(y_t|\, o_{1:t}, y_{1:t-1})}.$$

Hence, using the importance sampling identity

$$p(y_{1:t}|\, o_{1:t}) = \frac{w(y_{1:t})\, q_t(y_{1:t})}{\int w(y_{1:t})\, q_t(y_{1:t})\, dy_{1:t}},$$

a Monte Carlo approximation of $p(dy_{1:t}|\, o_{1:t})$ is given by

$$\widetilde{P}_N(dy_{1:t}|\, o_{1:t}) = \sum_{i=1}^{N} w_t^{(i)} \delta_{\widetilde{y}_{1:t}^{(i)}}(y_{1:t}), \tag{9}$$

where $w_t^{(i)} \propto w\left(\widetilde{y}_{1:t}^{(i)}\right)$ and $\sum_{i=1}^{N} w_t^{(i)} = 1$.

A standard choice for the importance density is $q_t(y_t|\, o_{1:t}, y_{1:t-1}) = p(y_t|\, x_{1:t}, y_{1:t-1})$ (Gordon et al. 1993, Kitagawa 1996, Isard and Blake 1996). According to (7), the density $p(y_t|\, x_{1:t}, y_{1:t-1})$ is Gaussian and, therefore, it is easy to sample from it. Using this proposal, the importance weights become $w(y_{1:t}) \propto \Pr(z_t|\, y_t)$. However, it is well-known that this strategy is inefficient if the new observation z_t is very informative (Doucet et al. 2001, Pitt and Shephard 1999). This situation arises often when treating binary time series with a vague prior. It is, therefore, necessary to engineer better sampling densities.

Here, we implement an *optimal* importance distribution, that is one that minimizes the variance of the importance weights $w(y_{1:t})$ *conditional upon* not modifying the path $y_{1:t-1}$. This distribution, which makes use of the latest observation, is given by

$$q_t(y_t|\, o_{1:t}, y_{1:t-1}) = p(y_t|\, x_{1:t}, y_{1:t-1}, z_t).$$

Hence, $w(y_{1:t}) = \Pr(z_t|\, x_{1:t}, y_{1:t-1})$ (Doucet et al. 1995, Doucet et al. 2000). In particular, we have

$$p(y_t|\, x_{1:t}, y_{1:t-1}, z_t) \propto \begin{cases} p(y_t|\, x_{1:t}, y_{1:t-1})\, \mathbb{I}_{[0,+\infty)}(y_t) & \text{if } z_t = 1 \\ p(y_t|\, x_{1:t}, y_{1:t-1})\, \mathbb{I}_{(-\infty,0)}(y_t) & \text{if } z_t = 0 \end{cases},$$

which is a truncated Gaussian distribution and consequently

$$w(y_{1:t}) \propto \Pr(z_t|\, x_{1:t}, y_{1:t-1}) = \left(1 - \Phi\left(-\frac{y_{t|t-1}}{\sqrt{S_t}}\right)\right)^{z_t} \Phi\left(-\frac{y_{t|t-1}}{\sqrt{S_t}}\right)^{1-z_t}.$$

3.2.2 Selection step

The aim of the selection is to obtain an "unweighted" approximate empirical distribution $P_N(dy_{1:t}|\, o_{1:t})$ of the "weighted" distribution $\widetilde{P}_N(dy_{1:t}|\, o_{1:t})$. The basic idea is to discard the samples with small weights and multiply those with large weights. This enables us to track moving distributions. The introduction of this key step in led to the first operational SMC method (Gordon et al. 1993).

A selection procedure associates with each particle $\widetilde{y}_{1:t}^{(i)}$ a number of offspring $N_t^{(i)} \in \mathbb{N}$, such that $\sum_{i=1}^{N} N_t^{(i)} = N$, to obtain N new particles $y_{1:t}^{(i)}$ with associated empirical distribution $\overline{\pi}_t^N$, where

$$P_N\left(dy_{1:t}|\, o_{1:t}\right) = N^{-1} \sum_{i=1}^{N} N_t^{(i)} \delta_{\widetilde{y}_{1:t}^{(i)}}\left(dy_{1:t}\right) = N^{-1} \sum_{i=1}^{N} \delta_{y_{1:t}^{(i)}}\left(dy_{1:t}\right). \tag{10}$$

For the surviving particles $\widetilde{y}_{1:t}^{(i)}$, the weights are reset to $1/N$, and these particles are distributed approximately according to $P_N\left(dy_{1:t}|\, o_{1:t}\right)$.

Many schemes have been presented in the literature to perform selection. These include the multinomial sampling originally proposed in (Gordon et al. 1993) and stratified resampling (Kitagawa 1996). All these algorithms ensure that $\mathbb{E}\left\{N_t^{(i)}\right\} = Nw_t^{(i)}$ while they differ in terms of $var\left\{N_t^{(i)}\right\}$. In practice, one must favour algorithms with small Monte Carlo variations. The selection step is done according to the stratified sampling scheme (Kitagawa 1996) as it can be shown that it minimizes $var\left\{N_t^{(i)}\right\}$ subject to $\mathbb{E}\left\{N_t^{(i)}\right\} = Nw_t^{(i)}$.

3.2.3 Algorithm

In practice, one does not want to estimate the joint density $p\left(y_{1:t}|\, o_{1:t}\right)$ as the memory requirements would increase over time. We are typically only interested in characteristics of the filtering density $p\left(\beta_t|\, o_{1:t}\right)$ such as $\mathbb{E}\left(\beta_t|\, o_{1:t}\right)$ or $cov\left(\beta_t|\, o_{1:t}\right)$. Moreover, we adopt importance densities $q_t\left(y_t|\, o_{1:t}, y_{1:t-1}\right)$ that only depend on $y_{1:t-1}$ via the set of low-dimensional sufficient statistics $\left(\beta_{t|t-1}, \Sigma_{t|t-1}\right)$, and we write, with some abuse of notation, $q_t\left(y_t|\, d_t, \beta_{t|t-1}, \Sigma_{t|t-1}\right) \triangleq q_t\left(y_t|\, o_{1:t}, y_{1:t-1}\right)$. This applies to both the prior and optimal importance densities. In particular, since

$$p\left(y_t|\, \beta_{t|t-1}, \Sigma_{t|t-1}\right) \triangleq p\left(y_t|\, x_{1:t}, y_{1:t-1}\right) = \mathcal{N}\left(y_t; y_{t|t-1}, S_t\right),$$

$y_{t|t-1}$ and S_t being deterministic functions of $\beta_{t|t-1}$ and $\Sigma_{t|t-1}$, it is only necessary to keep in memory $\left\{y_t^{(i)}, \beta_{t|t-1}^{(i)}\right\}_{i=1}^{N}$ and $\Sigma_{t|t-1}$ instead of $\left\{y_{1:t}^{(i)}\right\}_{i=1}^{N}$. Also note that $\Sigma_{t+1|t}^{(i)} = \Sigma_{t+1|t}^{(j)} = \Sigma_{t+1|t}$ for any $(i,j) \in \{1, \ldots, N\}$. In other words, the computationally demanding step of the Kalman filter, namely the covariance update, is outside the main particles loop. This results in substantial computational savings.

Given N particles $\left\{y_{t-1}^{(i)}\right\}_{i=1}^{N}$ at time $t-1$ distributed approximately according to $p\left(y_{t-1}|\, z_{1:t-1}\right)$ and the associated sufficient statistics $\left\{\beta_{t|t-1}^{(i)}\right\}_{i=1}^{N}$ and $\Sigma_{t|t-1}$, the particle filter with importance density $q_t\left(y_t|\, o_{1:t}, \beta_{t|t-1}, \Sigma_{t|t-1}\right)$ proceeds as shown in Figure 1 at time t. The computational complexity of this algorithm at each time step is $O(N)$. We remark again that one does not have to compute N parallel Kalman filters as most of the calculations required need only to be done once. For example, $\Sigma_{t|t-1}^{(i)} = \Sigma_{t|t-1}^{(j)}$, $\Sigma_{t|t}^{(i)} = \Sigma_{t,t}^{(j)}$ and $S_{t|t}^{(i)} = S_{t|t}^{(j)}$ for any $(i,j) \in \{1, \ldots, N\}$.

Sequential importance sampling step

- For $i=1,...,N$, set $\widetilde{\beta}_{t|t-1}^{(i)} \triangleq \beta_{t|t-1}^{(i)}$ and sample $\widetilde{y}_t^{(i)} \sim Q_t\left(dy_t|\, o_{1:t}, \widetilde{\beta}_{t|t-1}^{(i)}, \Sigma_{t|t-1}\right)$.
- For $i=1,...,N$, evaluate and normalize the importance weights

$$w_t^{(i)} \propto \frac{\Pr\left(z_t|\, \widetilde{y}_t^{(i)}\right) p\left(\widetilde{y}_t^{(i)}\middle|\, \widetilde{\beta}_{t|t-1}^{(i)}, \Sigma_{t|t-1}\right)}{q_t\left(\widetilde{y}_t^{(i)}\middle|\, d_t, \widetilde{\beta}_{t|t-1}^{(i)}, \Sigma_{t|t-1}\right)}. \tag{11}$$

Selection step

- Multiply/Discard particles $\left\{\widetilde{y}_t^{(i)}, \widetilde{\beta}_{t|t-1}^{(i)}\right\}_{i=1}^{N}$ with respect to high/low importance weights $w_t^{(i)}$ to obtain N particles $\left\{y_t^{(i)}, \beta_{t|t-1}^{(i)}\right\}_{i=1}^{N}$.

Updating step

- Compute $\Sigma_{t+1|t}$ given $\Sigma_{t|t-1}$.
- For $i=1,...,N$, use one step of the Kalman recursion (6) to compute $\left\{\widetilde{\beta}_{t+1|t}^{(i)}\right\}$ given $\left\{y_t^{(i)}, \widetilde{\beta}_{t|t-1}^{(i)}\right\}$ and $\Sigma_{t|t-1}$.

Figure 1: Particle filtering for semiparametric binary classification. In our particular case, the sampling step can be placed after the selection step to improve performance.

Remark 1 *When we use $q_t(y_t|\, o_{1:t}, y_{1:t-1}) = p(y_t|\, o_{1:t}, y_{1:t-1})$ the importance weight $w_t \propto \Pr(z_t|\, x_{1:t}, y_{1:t-1})$ does not depend on y_t. It is thus possible to carry out the selection step before the sampling step. The algorithm is then similar to the auxiliary variable particle filter of (Pitt and Shephard 1999). This modification to the original algorithm has important implications. It enables us to search for more likely regions of the posterior at time $t-1$ using the information at time t to generate better samples at time t.*

3.2.4 Convergence results

For any n, let $B(\mathbb{R}^n)$ be the space of bounded, Borel measurable functions on $\mathbb{R}^n$, and $f \in B(\mathbb{R}^n)$: $\|f\| \triangleq \sup_{z\in\mathbb{R}^n} |f(z)|$. One can show that if the importance weight (8) is bounded above, then the following theorem holds (this theorem is an extension of previous results obtained in (Crisan et al. 1999))

Theorem 1 *(Crisan and Doucet 2000) For all $t \geq 0$, there exists c_t independent of N such that for any $f_t \in B\left((\mathbb{R}^{n_x})^{t+1}\right)$*

$$\mathbb{E}\left[\left(\int f_t(y_{1:t})\left(\widetilde{P}_N(dy_{1:t}|\, o_{1:t}) - P(dy_{1:t}|\, o_{1:t})\right)\right)^2\right] \leq c_t \frac{\|f_t\|^2}{N}. \tag{12}$$

where the expectation is taken with respect to the particle filter realisations. As a corollary, one obtains easily that $\widehat{\Pr}_N\left(z_{t+1}=1|\, o_{1:t}, x_{t+1}\right)$, given by (5), satisfies

$$\mathbb{E}\left[\left(\widehat{\Pr}_N\left(z_{t+1}=1|\, o_{1:t}, x_{t+1}\right)-\Pr\left(z_{t+1}=1|\, o_{1:t}, x_{t+1}\right)\right)^2\right] \leq \frac{c_t}{N}$$

This result establishes convergence of our particle filtering method.

4 Simulations

To compare our model, using the Rao-Blackwellised particle filtering (RBPF) algorithm, to standard logistic and probit classification with PF, we generated data from clusters that change with time as shown in Figure 2. This data set captures the characteristics of a fault detection problem that we are currently studying. (For some results of applying PF to fault detection in marine diesel engines, please refer to (Højen-Sørensen et al. 2000). More results will become available once permission is granted.) This data cannot be easily separated with an algorithm based on a time-invariant model.

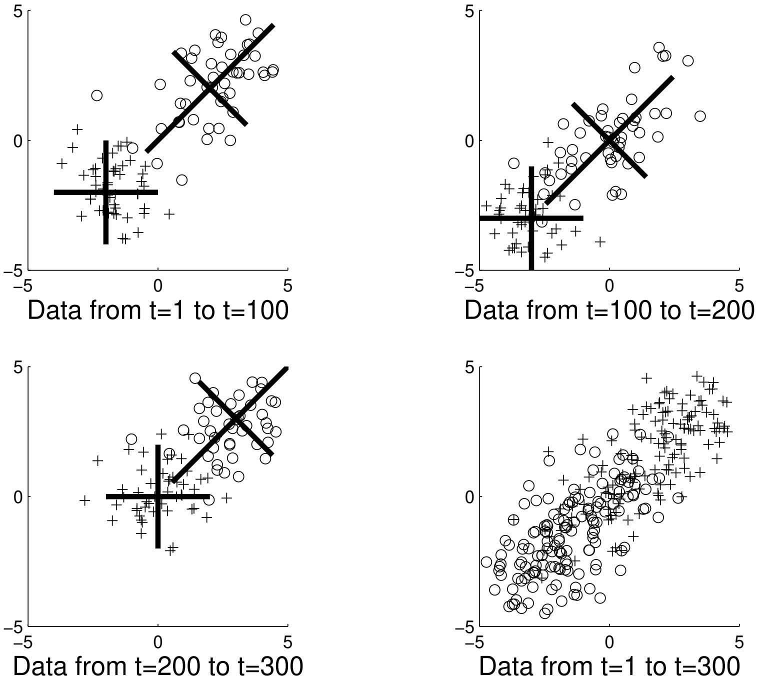

Figure 2: Time-varying data.

For the results presented here, we set the initial distributions to: $\beta_0 \sim \mathcal{N}(\mathbf{0}, 5\mathbf{I})$ and $y_0 \sim \mathcal{N}(\mathbf{0}, 5\mathbf{I})$. The process matrices were set to $A = I$ and $B = \delta I$, where $\delta^2 = 0.1$ is a smoothing parameter. The number of bases (cubic splines with random locations) was set to 10. (It is of course possible, when we have some data already, to initialise the bases locations so that they correspond to the input data. This trick for efficient classification in high dimensional input spaces is used in the support vector machines setting (Vapnik

1995).) The experiment was repeated with the number of particles varying between 10 and 400.

Figure 3 shows the "value for money" summary plot. The new algorithm has a lower computational cost and shows a significant reduction in estimation variance. Note that the computation of the RBPF stays consistently low even for small numbers of particles. This has enabled us to apply the technique to large models consisting of hundreds of Bases using a suitable regulariser. Another advantage of PF algorithms for classification is that they yield entire probability estimates of class membership as shown in Figure 4.

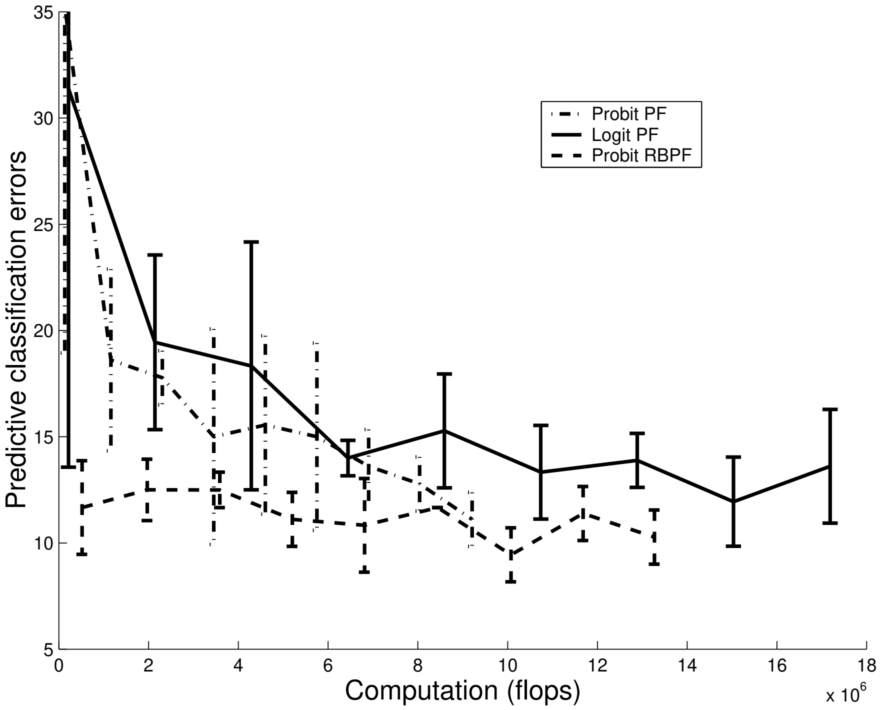

Figure 3: Number of classification errors as the number of particles varies between 10 and 400 (different computational costs). The algorithm with the augmentation trick (RBPF) is more efficient than standard PF algorithms.

5 Conclusions

In this paper, we proposed a dynamic Bayesian model for time-varying binary classification and a remarkably efficient particle filtering algorithm to perform the required computations. The efficiency of our algorithm is a result of an augmentation trick, Rao-Blackwellisation, adopting the optimal importance distribution, being able to swap the sampling and selection steps and only needing to update the Kalman filter means in the particles loop. This extends the realm of efficient particle filtering in high dimensions to the ubiquitous setting of Gaussian latent variables and binary observations. Extensions to n-ary observations, different link functions and estimation of the hyper-parameters can be carried out in the

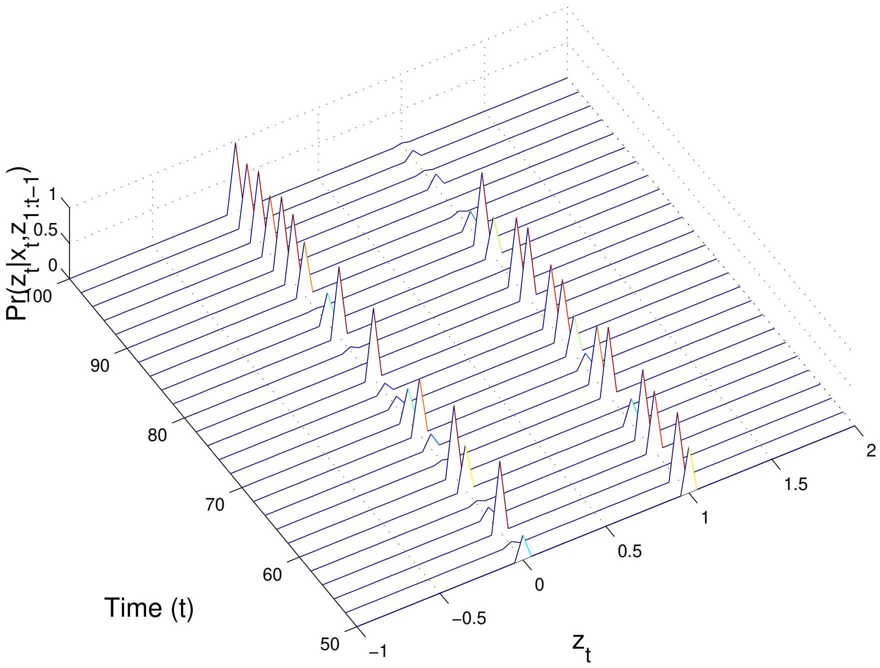

Figure 4: Predictive density.

same framework.

6 Acknowledgments

We would like to thank Torben Fog and Pedro Højen-Sørensen.

References

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data, *Journal of the American Statistical Association* **88**(422): 669–679.

Anderson, B. D. and Moore, J. B. (1979). *Optimal Filtering*, Prentice-Hall, New Jersey.

Andrieu, C., de Freitas, N. and Doucet, A. (1999a). Robust full Bayesian learning for radial basis networks, To appear in *Neural Computation*.

Andrieu, C., de Freitas, N. and Doucet, A. (1999b). Sequential Bayesian estimation and model selection applied to neural networks, *Technical Report CUED/F-INFENG/TR 341*, Cambridge University Engineering Department.

Bergman, N., Doucet, A. and Gordon, N. J. (2001). Optimal estimation and Cramér-Rao bounds for partial non-Gaussian state space models, *Annals of the Institute of Mathematical Statistics* **52**(1): 1–17.

Crisan, D. and Doucet, A. (2000). Convergence of sequential Monte Carlo methods, *Technical Report CUED/F-INFENG/TR 381*, Cambridge University Engineering Department.

Crisan, D., Del Moral, P. and Lyons, T. (1999). Discrete filtering using branching and interacting particle systems, *Markov Processes and Related Fields* **5**(3): 293–318.

Doucet, A., Barat, E. and Duvaut, P. (1995). Monte Carlo approach to recursive Bayesian state estimation, *Proceedings of the IEEE Signal Processing/Athos Workshop on Higher Order Statistics*, Girona, Spain, pp. 12–14.

Doucet, A., de Freitas, N. and Gordon, N. J. (eds) (2001). *Sequential Monte Carlo Methods in Practice*, Springer-Verlag.

Doucet, A., Godsill, S. and Andrieu, C. (2000). On sequential Monte Carlo sampling methods for Bayesian filtering, *Statistics and Computing* **10**(3): 197–208.

Gordon, N. J., Salmond, D. J. and Smith, A. F. M. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation, *IEE Proceedings-F* **140**(2): 107–113.

Gouriéroux, C. and Monfort, A. (1996). *Simulation-Based Econometric Methods*, Core Lectures, Oxford University Press.

Højen-Sørensen, P., de Freitas, N. and Fog, T. (2000). On-line probabilistic classification with particle filters, *IEEE Neural Networks for Signal Processing*, Sydney, Australia.

Holmes, C. C. and Denison, D. G. T. (2000). A Bayesian MARS classifier, *Technical Report TR-00-04*, Department of Mathematics, Imperial College of London.

Holmes, C. C. and Mallick, B. K. (2000). Bayesian wavelet networks for nonparametric regression, *IEEE Transactions on Neural Networks* **11**(1): 27–35.

Isard, M. and Blake, A. (1996). Contour tracking by stochastic propagation of conditional density, *European Conference on Computer Vision*, Cambridge, UK, pp. 343–356.

Kitagawa, G. (1996). Monte Carlo filter and smoother for non-Gaussian nonlinear state space models, *Journal of Computational and Graphical Statistics* **5**: 1–25.

Kitagawa, G. and Gersch, W. (1996). *Smoothness Priors Analysis of Time Series*, Vol. 116 of *Lecture Notes In Statistics*, Springer-Verlag.

McFadden, D. (1989). A method of simulated momemts for estimation of discrete response models without numerical integration, *Econometrica* **57**: 995–1026.

Melvin, D. G. (1996). A comparison of statistical and connectionist techniques for liver transplant monitoring, *Technical Report CUED/F-INFENG/TR 282*, Cambridge University Engineering Department.

Penny, W. D., Roberts, S. J., Curran, E. and Stokes, M. (1999). EEG-based communication: a pattern recognition approach, *IEEE Transactions on Rehabilitation Engineering* **20**(5): 101–102.

Pitt, M. K. and Shephard, N. (1999). Filtering via simulation: Auxiliary particle filters, *Journal of the American Statistical Association* **94**(446): 590–599.

Robert, C. P. and Casella, G. (1999). *Monte Carlo Statistical Methods*, Springer-Verlag, New York.

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

West, M. and Harrison, J. (1997). *Bayesian forecasting and dynamic models*, Springer Series in Statistics, second edn, Springer-Verlag, New York.

Wood, S. and Kohn, R. (1998). A Bayesian approach to robust binary nonparametric regression, *Journal of the American Statistical Association* **93**(441): 203–213.