SEQUENTIAL MONTE CARLO FOR MODEL SELECTION AND ESTIMATION OF NEURAL NETWORKS

Nando de Freitas

Computer Science Division
387 Soda Hall
University of California, Berkeley
Berkeley, CA 94720-1776 USA
jfgf@cs.berkeley.edu

Christophe Andrieu

Cambridge University
Engineering Department
Trumpington Street
Cambridge, CB2 1PZ, UK
ca226@eng.cam.ac.uk

ABSTRACT

We address the complex problem of sequential Bayesian learning and model selection for neural networks. This problem does not usually admit any type of closed-form analytical solution and, as a result, one has to resort to numerical methods. We propose here an original sequential simulation-based strategy to perform the necessary computations. It combines sequential importance sampling, a selection procedure, variance reduction techniques and reversible jump Markov chain Monte Carlo (MCMC) moves. We demonstrate the effectiveness of the method by applying it to radial basis function networks. The approach can be easily extended to other interesting on-line model selection problems.

1. INTRODUCTION

Model selection is a fundamental data analysis task that arises in many fields of science and engineering. Research in this area indicates that the Bayesian paradigm is a natural and powerful approach for dealing with this problem. However, even within this framework, model selection is a difficult problem as it typically involves the evaluation of posterior model probabilities, which, in turn, require the computation of high-dimensional integrals with no closed-form analytical expression.

In this paper, we develop an original sequential Monte Carlo (SMC) method to perform Bayesian model selection: see [6] for an introduction[1]. This on-line simulation-based estimation scheme allows us to run a single algorithm which concentrates itself on the model orders of interest. We apply this method to the problem of estimating, on-line, the parameters, noise variance, and number of neurons of a hybrid model consisting of a linear model and a radial basis function (RBF) network. We propose a full Bayesian prior that allows us to integrate the basis coefficients and linear regression parameters. As a result of this, we obtain an efficient algorithm that computes the coefficients and linear regression parameters using a stochastic bank of Kalman filters.

[1] A more detailed version of this paper [2] and the software are available at `http://www-svr.eng.cam.ac.uk/~jfgf`.

2. PROBLEM STATEMENT

Many physical processes may be described by the following nonlinear, multivariate input-output mapping:

$$\mathbf{y}_t = \mathbf{f}(\mathbf{x}_t) + \mathbf{n}_t$$

where $\mathbf{x}_t \in \mathbb{R}^d$ corresponds to a group of input variables, $\mathbf{y}_t \in \mathbb{R}^c$ to the output variables and $\mathbf{n}_t \in \mathbb{R}^c$ to an unknown noise process. In this context, the learning problem involves computing an approximation to the function $\mathbf{f}$ and estimating the characteristics of the noise process each time new data $\mathbf{d}_t \triangleq \{\mathbf{x}_t, \mathbf{y}_t\}$ arrive.

When the exact nonlinear structure of the multivariate function $\mathbf{f}$ cannot be established *a priori*, it may be synthesised as a combination of parametrised basis functions. For the purposes of this paper, we adopt an approximation scheme consisting of a mixture of k radial basis functions (RBFs) and a linear regression term [11]. However, the work can be straight-forwardly extended to other regression models. More precisely, our approximation model for the observations at time t is:

$$\mathbf{y}_t = \boldsymbol{b}_t + \boldsymbol{\beta}'_t\mathbf{x}_t + \mathbf{n}_t \quad k_t = 0$$

$$\mathbf{y}_t = \sum_{j=1}^{k_t} \mathbf{a}_{j,t}\phi_t(\|\mathbf{x}_t - \boldsymbol{\mu}_{j,t}\|) + \boldsymbol{b}_t + \boldsymbol{\beta}'_t\mathbf{x}_t + \mathbf{n}_t \quad k_t \geq 1$$

where $\|\cdot\|$ denotes a distance metric (usually Euclidean or Mahalanobis), $\boldsymbol{\mu}_{j,t} \in \mathbb{R}^d$ denotes the j-th RBF centre for a model with k RBFs, $\mathbf{a}_{j,t} \in \mathbb{R}^c$ the j-th RBF amplitude and $\boldsymbol{b}_t \in \mathbb{R}^c$ and $\boldsymbol{\beta}_t \in \mathbb{R}^d \times \mathbb{R}^c$ the linear regression parameters. The noise sequence $\mathbf{n}_t \in \mathbb{R}^c$ is assumed to be zero-mean Gaussian; its variance changes over time.

Depending on our *a priori* knowledge about the smoothness of the mapping, we can choose different types of basis functions. The most common choices are linear, cubic, thin plate spline, Gaussian and multiquadric basis functions [1, 11]. For convenience, we express the approximation model in vector-matrix form[2] $\mathbf{y}_t = \mathbf{D}(\boldsymbol{\mu}_{1:k_t,t}, \mathbf{x}_t)\boldsymbol{\alpha}_{1:1+d+k_t,t} + \mathbf{n}_t$,

[2] We adopt the notation $\mathbf{y}_{1:c,t} \triangleq (\mathbf{y}_{1,t}, \mathbf{y}_{2,t}, \ldots, \mathbf{y}_{c,t})'$ to denote all the output observations at time t. To simplify this notation, $\mathbf{y}_t$ is equivalent to $\mathbf{y}_{1:c,t}$. That is, if one index does not appear, it is implied that we are referring to all of its possible

where $\mathbf{D}$ is the matrix of regressors and basis functions and $\boldsymbol{\alpha}_{1:1+d+k_t,t}$ is a matrix of linear and RBF coefficients.

The number k_t of basis functions and their parameters $\boldsymbol{\theta}_{k_t,t} \triangleq \{\boldsymbol{\alpha}_{1:c,1:m_t,t}, \boldsymbol{\mu}_{1:d,1:k_t,t}, \log(\boldsymbol{\sigma}^2_{1:c,t})\}$, with $m_t = 1 + d + k_t$, are assumed to be unknown and need to be estimated sequentially.

3. BAYESIAN MODEL

We adopt a state space representation to describe the network's evolution over time:

$$\begin{aligned}
\mathbf{y}_t &= \mathbf{D}(\boldsymbol{\mu}_{1:k_t,t}, \mathbf{x}_t)\boldsymbol{\alpha}_{1:m_t,t} + \mathbf{n}_t & (1)\\
k_{t+1} &= k_t + \epsilon_k & (2)\\
\boldsymbol{\mu}_{1:k_{t+1},t+1} &= \boldsymbol{\mu}_{1:k_{t+1},t} + \epsilon_{\boldsymbol{\mu}} & (3)\\
\boldsymbol{\alpha}_{1:m_{t+1},t+1} &= \boldsymbol{\alpha}_{1:m_{t+1},t} + \epsilon_{\boldsymbol{\alpha}} & (4)\\
\log(\boldsymbol{\sigma}^2_{t+1}) &= \log(\boldsymbol{\sigma}^2_t) + \epsilon_{\boldsymbol{\sigma}} & (5)
\end{aligned}$$

where the diffusion process ϵ_k is sampled from a uniform discrete distribution: $\epsilon_k \sim \mathcal{UD}(0, \delta_k^2)$. That is, it corresponds to an integer value in the interval $[-\delta_k, \ldots, 0, \ldots, \delta_k]$. The other processes are sampled from normal distributions:

$$\epsilon_{\boldsymbol{\mu}} \sim \mathcal{N}(0, \delta^2_{\boldsymbol{\mu}}\mathbf{I}_d) \qquad \epsilon_{\boldsymbol{\alpha}} \sim \mathcal{N}(0, \delta^2_{\boldsymbol{\alpha}}\mathbf{I}_{m_t}) \qquad \epsilon_{\boldsymbol{\sigma}} \sim \mathcal{N}(0, \delta^2_{\boldsymbol{\sigma}}\mathbf{I}_c)$$

where $m_t = k_t + d + 1$. Note that if k_t changes, the dimensions of $\boldsymbol{\alpha}_{1:m_t,t}$ and $\boldsymbol{\mu}_{1:k_t,t}$ also change accordingly.

We adopt multivariate normal distributions to represent the priors $p(\boldsymbol{\mu}_{k_0,0})$ and $p(\boldsymbol{\alpha}_{k_0,0})$, a uniform distribution for $p(\boldsymbol{\sigma}^2_{k_0,0})$ and a discrete uniform distribution for $p(k_0)$. The parameters of these priors should reflect our degree of belief on the initial values of the various quantities of interest. Figure 1 shows the directed acyclic graphical model representation of the joint distribution for the first four time steps.

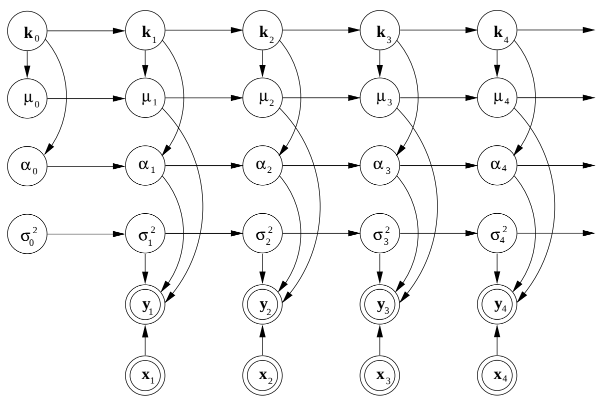

Figure 1: Dynamical directed acyclic graphical model. The circles represent unknowns, while the double circles correspond to the data. The hyper-parameters $\delta^2_{\boldsymbol{\mu}}$, $\delta^2_{\boldsymbol{\alpha}}$, $\delta^2_{\boldsymbol{\sigma}}$ and δ^2_k have been omitted for clarity.

values. We shall favour the shorter notation and only invoke the longer notation to avoid ambiguities and emphasise certain dependencies.

4. SEQUENTIAL MONTE CARLO

Many researchers in the the statistical and signal processing communities have proposed the use of sequential Monte Carlo algorithms [5, 2, 13]. Figure 2 illustrates the operation of a generic SMC method. Most existing SMC methods belong to a framework where the model order is assumed to be known. We present here a Bayesian importance sampling method to deal with the more general case of unknown model dimension. To accomplish this goal in an efficient way, we introduce reversible jump MCMC steps into the algorithm.

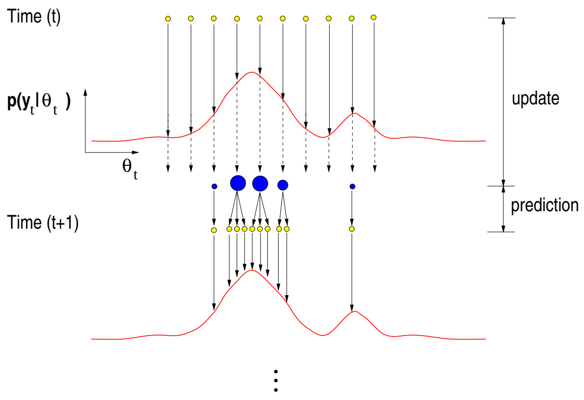

Figure 2: Update and prediction stages in a generic sequential Monte Carlo algorithm with the transition prior as proposal. The particles with higher importance weights are allowed to have more "children" as they provide a better weighted description of the posterior.

5. SMC APPLIED TO RBF NETWORKS

In this section, we present the algorithm's pseudo-code and then proceed to describe the various steps. Given N particles (samples) $\{(\boldsymbol{\theta}^{(i)}_{k_{0:t-1},0:t-1}, k^{(i)}_{0:t-1}); i \in \{1, \ldots, N\}\}$, at time $t-1$, approximately distributed according to the posterior $p(\boldsymbol{\theta}_{k_{0:t-1},0:t-1}, k_{0:t-1}|\mathbf{d}_{1:t-1})$ the algorithm allows us to compute N particles $(\boldsymbol{\theta}^{(i)}_{k_{0:t},0:t}, k^{(i)}_{0:t})$ approximately distributed according to $p(\boldsymbol{\theta}_{k_{0:t},0:t}, k_{0:t}|\mathbf{d}_{1:t})$, at time t.

SMC Algorithm for RBFs Bayesian Model Selection

1. Sequential importance sampling step
   - For $i = 1, \ldots, N$, sample:

$$\begin{aligned}
\widehat{k}^{(i)}_t &\sim p(k_t|k^{(i)}_{t-1})\\
\log(\widehat{\boldsymbol{\sigma}}^2_t)^{(i)}| &\sim p(\log(\boldsymbol{\sigma}^2_t)|\log(\boldsymbol{\sigma}^2_{t-1})^{(i)})\\
\widehat{\boldsymbol{\mu}}^{(i)}_{k_t,t} &\sim p(\boldsymbol{\mu}_{1:k_t,t}|\boldsymbol{\mu}^{(i)}_{1:\widehat{k}_t,t-1}, \widehat{k}^{(i)}_t)
\end{aligned}$$

   and set:

$$(\widehat{\boldsymbol{\theta}}^{(i)}_{\widehat{k}_{0:t},0:t}, \widehat{k}^{(i)}_{0:t}) \triangleq ((\boldsymbol{\theta}^{(i)}_{k_{0:t-1},0:t-1}, k^{(i)}_{0:t-1}), (\widehat{\boldsymbol{\theta}}^{(i)}_{\widehat{k}_t,t}, \widehat{k}^{(i)}_t))$$

- For $i = 1, \ldots, N$, evaluate the importance weights up to a normalising constant:

$$w_t^{(i)} = p(\mathbf{y}_t | \widehat{\boldsymbol{\mu}}_{1:\widehat{k}_{0:t},0:t}^{(i)}, \log(\widehat{\boldsymbol{\sigma}}_{0:t}^2)^{(i)}, \widehat{k}_{0:t}^{(i)}, \mathbf{x}_t, \mathbf{d}_{1:t-1})$$

- For $i = 1, \ldots, N$, normalise the importance weights.

2. Selection step

   - Use a selection scheme to multiply (suppress) particles $(\widehat{\boldsymbol{\theta}}_{\widehat{k}_{0:t},0:t}^{(i)}, \widehat{k}_{0:t}^{(i)})$ with high (low) importance weights $\widetilde{w}_t^{(i)}$, respectively, to obtain N random samples $(\widetilde{\boldsymbol{\theta}}_{\widetilde{k}_{0:t},0:t}^{(i)}, \widetilde{k}_{0:t}^{(i)})$.

3. MCMC step

   - Apply a Markov transition kernel of invariant distribution $\prod_{i=1}^{N} p(\boldsymbol{\theta}_{k_{0:t},0:t}^{(i)}, k_{0:t}^{(i)} | \mathbf{d}_{1:t})$ to obtain the new samples $(\boldsymbol{\theta}_{k_{0:t},0:t}^{(i)}, k_{0:t}^{(i)})$.
   - Update the parameters $\boldsymbol{\alpha}_{1:m_t,t}^{(i)}$ using a bank of Kalman filters.

■

It is important to note that for the parameters $\boldsymbol{\alpha}_{1:m_t,t}$, we have a linear Gaussian state space model. Consequently, it is possible to integrate out $\boldsymbol{\alpha}_{1:m_t,t}$. The end result is a bank of Kalman filters for each network output [12]. It can be proved that this integration reduces the variance of the estimates [5]. The Bayesian importance sampling stage has been extensively documented in the literature [5, 4]. The selection scheme associates to each particle $(\widehat{\boldsymbol{\theta}}_{\widehat{k}_{0:t},0:t}^{(i)}, \widehat{k}_{0:t}^{(i)})$ a number of "children", say $N_i \in \mathbb{N}$, such that $\sum_{i=1}^{N} N_i = N$. Various efficient $\mathcal{O}(N)$ selection schemes have been proposed in the literature, for example multinomial resampling [9, 15] and residual resampling [13]. These schemes satisfy $\mathbb{E}[N_i] = N\widetilde{w}_t^{(i)}$ but their performance varies in terms of $var[N_i]$.

The selection step can increase the variance of the estimates at time t. A strategy for overcoming this problem is to introduce MCMC steps of invariant distribution $p(k_{0:t}, \boldsymbol{\theta}_{k_{0:t},0:t} | \mathbf{d}_{1:t})$ on each particle [3, 7, 8, 14]. The basic idea is that, by applying a Markov transition kernel, the total variation of the current distribution with respect to the invariant distribution can only decrease. Note, however, that we do not require this kernel to be ergodic. Convergence results for this type of algorithm are presented in [8]. In this paper, we apply reversible jump MCMC [10] steps to each particle. This allows us jump between subspaces of different dimensions, thereby performing model order selection. The algorithms for the various moves are presented in detail in [2].

6. EXPERIMENT

In this section, we demonstrate the algorithm with a simple and instructive example. The application of the algorithm to various real data sets is discussed in the technical report version of this paper [2]. The data was generated from the following univariate, nonlinear and time-varying function using 1000 covariate points uniformly on $[0, 1]$:

$$y_t = \begin{cases} 4x_t - 2 + \left(2 + \frac{t}{150}\right)e^{-15(x_t-0.3)^2} + 2e^{-15(x_t-0.7)^2} + n_t & 1 \leq t \leq 500 \\ 4x_t - 2 + 2e^{-15(x_t-0.5)^2} + n_t & 500 < t \leq 1000 \end{cases}$$

where $n_t \sim \mathcal{N}(0, \sigma^2)$ with $\sigma^2 = 0.05$. Note that the drift in the size of the Gaussian component at $x_t = 0.3$ is seriously affected by the noise corruption. We selected Gaussian radial basis functions with the same variance as the Gaussian signal components. We chose a uniform prior for $\boldsymbol{\sigma}_0^2$ between 0 and 0.35. Since the domain of the data is the interval $[0, 1]$, we chose a multivariate Gaussian prior for $\boldsymbol{\mu}_{1:k_0,0}$ with mean 0.5 and variance 0.05. The diffusion parameters $\delta_{\boldsymbol{\mu}}^2$, $\delta_{\boldsymbol{\alpha}}^2$ and $\delta_{\boldsymbol{\sigma}}^2$ were all set to 0.001. The parameters $k_{\max}$ and $\varsigma^\star$ were set to 20 and 0.05 respectively. We used an informative multivariate Gaussian prior for $\boldsymbol{\alpha}_{1:m_0,0}$ with mean 2 and variance 2. The extended means $\overline{\boldsymbol{\alpha}}_{1:m_0,0}$ of the bank of Kalman filters were drawn from the same prior.

The simulation results, using 1000 samples, are depicted in Figures 3 and 4. Figure 3 shows that the estimated model orders and noise variances (posterior means) converge to the right values. As a result, the output one-step-ahead predictions are very accurate. Figure 4 shows the estimates of the linear regression coefficients, basis centres and basis coefficients. These posterior mean estimates correspond to the centroids of the distributions computed using the main modes ($k = 2$ from $t = 1$ to $t = 500$ and $k = 1$ afterwards). The algorithm performs extremely well when computing the linear regression parameters. However, it is clear, that to obtain similar results for the other quantities, it requires more samples. It should also be mentioned, that given the large amount of noise, it is very difficult to track the drift in the basis function amplitude. Figure 5 shows the probabilities of the estimated model order.

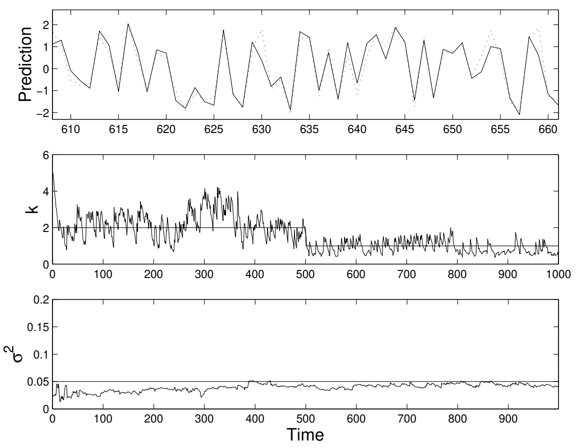

Figure 3: The top plot shows the one-step-ahead output predictions. The middle and bottom plots show the true values and estimates of the model order and noise variance respectively.

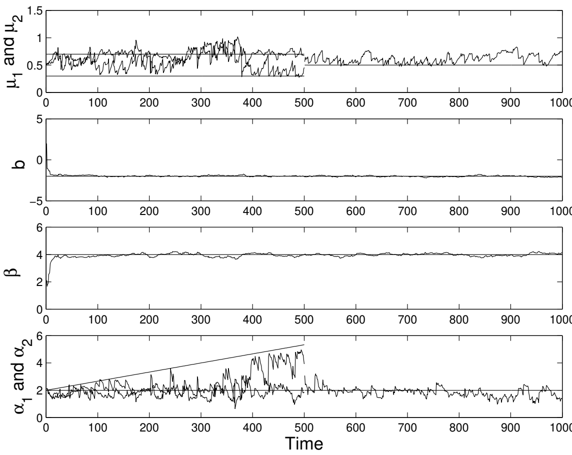

Figure 4: Estimated and true values of the basis centres, linear regression parameters and basis coefficients.

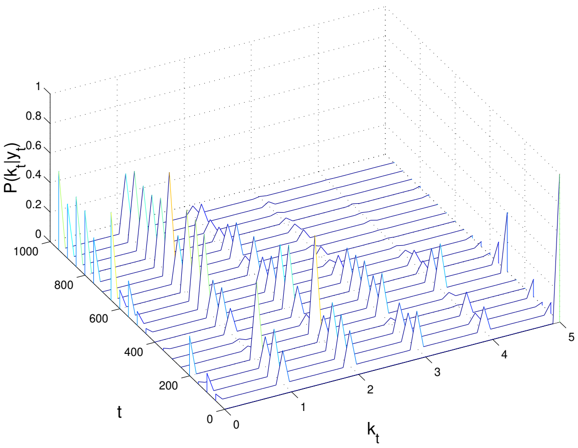

Figure 5: Probabilities of the estimated model order.

7. CONCLUSIONS

We proposed a simulation-based approach to perform sequential Bayesian model selection. This strategy combines sequential Monte Carlo methods and reversible jump MCMC methods. It also makes use of variance reduction techniques. Its only drawback is that it is computationally intensive, yet it can be straightforwardly implemented on parallel computers. We demonstrated the effectiveness of the method on various complex joint estimation and model selection problems. There are many avenues for future research. These include adopting MCMC schemes with forgetting factors to estimate the noise variance, thereby avoiding the need for the diffusion parameter δ_σ^2; allowing for models with different types of basis functions; performing input variable selection; experimenting with other noise distributions and testing the algorithms on additional problems.

REFERENCES

[1] C Andrieu, J F G de Freitas, and A Doucet. Robust full Bayesian learning for neural networks. Technical Report CUED/F-INFENG/TR 343, Cambridge University Engineering Department, April 1999.

[2] C Andrieu, J F G de Freitas, and A Doucet. Sequential Bayesian estimation and model selection applied to neural networks. Technical Report CUED/F-INFENG/TR 341, Cambridge University Engineering Department, May 1999.

[3] C Andrieu, J F G de Freitas, and A Doucet. Sequential MCMC for Bayesian model selection. In *IEEE Higher Order Statistics Workshop*, pages 130–134, Ceasarea, Israel, 1999.

[4] J F G de Freitas, M Niranjan, A H Gee, and A Doucet. Sequential Monte Carlo methods to train neural network models. To appear in *Neural Computation*, 1999.

[5] A Doucet. On sequential simulation-based methods for Bayesian filtering. To appear in *Statistics and Computing*, 1999.

[6] A Doucet, J F G de Freitas, and N J Gordon. *Sequential Monte Carlo Methods in Practice*. Springer-Verlag, 2000.

[7] A Doucet and N J Gordon. Simulation-based optimal filter for manoeuvring target tracking. In *SPIE Signal and Data Processing of Small Targets*, volume SPIE 3809, 1999. To appear.

[8] W R Gilks and C Berzuini. Monte Carlo inference for dynamic Bayesian models. Unpublished. Medical Research Council, Cambridge, UK, 1998.

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

[10] P J Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[11] C C Holmes and B K Mallick. Bayesian radial basis functions of variable dimension. *Neural Computation*, 10(5):1217–1233, 1998.

[12] A H Jazwinski. *Stochastic Processes and Filtering Theory*. Academic Press, 1970.

[13] J S Liu and R Chen. Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, 93:1032–1044, 1998.

[14] S N MacEachern, M Clyde, and J S Liu. Sequential importance sampling for nonparametric Bayes models: the next generation. To appear in the *Canadian Journal of Statistics*, 1999.

[15] B D Ripley. *Stochastic Simulation*. Wiley, New York, 1987.