STRATEGIES FOR SEQUENTIAL INFERENCE IN FACTORIAL SWITCHING STATE SPACE MODELS

A. Taylan Cemgil

Signal Processing and Communications Lab. Dept. of Engineering, University of Cambridge, UK
atc27@cam.ac.uk

ABSTRACT

Factorial switching state space models are large hybrid time series models in which inference is intractable even in a single time slice. For the conditional Gaussian case, we derive a message propagation algorithm (upward-downward) that exploits the factorial structure of the model and facilitates computing messages without the need for inverting large matrices. Using the propagation algorithm as a subroutine, we develop a Rao-Blackwellized Gibbs sampler and a variational approximation of structured mean field type to compute an approximate proposal density. These proposal are useful for both filtering or for marginal maximum a-posteriori estimates. We illustrate the utility of our approach on a large factorial state space model for polyphonic music transcription.

Index Terms— Time series, Variational Bayes, Monte Carlo, Multi Hypothesis Tracker

1. INTRODUCTION

Time series models with switching regimes are useful in various areas of applied sciences, such as control, econometrics, signal processing and machine learning, see, e.g, [1]. In these disciplines, many phenomena of interest can be naturally described as a sequence of regimes, where, conditioned on the latent regime label, observed data is thought of as a realization from a (simple) model.

One simple switching state space model can be defined by the following hierarchical probabilistic model

$$\begin{aligned} r_0 &\sim p(r_0) \qquad \theta_0 \sim \mathcal{N}(m, V) \\ r_k &\sim p(r_k|r_{k-1}) \\ \theta_k &\sim p(\theta_k|\theta_{k-1}, r_k) = \mathcal{N}(A_k\theta_{k-1}, Q_k) \\ y_k &\sim p(y_k|r_k, \theta_k) = \mathcal{N}(y_k; C_k\theta_k, R_k) \end{aligned}$$

where the index $k = 0, 1, \ldots$ denotes the time, θ_k is a hidden state vector and y_k is the observation. The discrete switch variable r_k is a regime indicator with $|r|$ states. It selects the state transition model the latent process $\{\theta_k\}$ will take and the observation model active at time k. If transition and observation models are conditionally Gaussian, we have the switching Kalman filter model [2] where where A_k, Q_k denote the transition matrix and noise covariance, C_k, R_k denote observation matrix and noise covariance and m, V are initialisation mean and noise covariance. We assume these are known given r_k, i.e. we have $A_k = A(r_k)$, e.t.c.

1.1. Inference

Often, given observations $y_{1:K} \equiv y_1 \ldots y_K$, we are interested into various marginals of the posterior, such as $p(r_k, \theta_k|y_{1:k})$ (known as the filtering density) or $p(r_{1:K}|y_{1:K})$. In the latter case, each hidden configuration $r_{1:K} \equiv \{r_1, \ldots, r_K\}$ specifies a possible segmentation to explain the data up to time K. We are naturally interested into the most likely segmentation

$$r^*_{1:K} = \operatorname*{argmax}_{r_{1:K}} p(r_{1:K}|y_{1:K})$$

where the posterior marginal is given as

$$p(r_{1:K}|y_{1:K}) \propto p(r_{1:K}) \int d\theta_{1:K} p(y_{1:K}|\theta_{1:K}, r_{1:K}) p(\theta_{1:K}|r_{1:K})$$

The difficulty of this optimisation problem stems from the fact that integrand needs to be evaluated for each of the exponentially many configurations $r_{1:K}$. Such "hybrid" inference problems, also known as MMAP (Marginal Maximum a-posteriori [3]) are significantly harder than computing expectations and marginals (which only involves integration) or optimisation (which only involves maximisation) [4]. This is due to the fact that the "inner" integration over a subset of the variables renders the remaining variables fully coupled destroying the Markovian structure which in turn renders the "outer" optimisation problem a hard joint combinatorial optimisation problem. Apart from a few special cases, where an exact polynomial time algorithm is known [5, 6, 7], in general the only known exact solution is exhaustive search: which is in a sequential setting equivalent to carrying forward a conditional filtering potential $\phi(\theta_k|r_{1:k})$ for each of the exponentially many configurations of $r_{1:k}$.

1.2. Factorial switching state space model

In many applications such as vision, audio signal processing (source separation, spectral analysis and polyphonic transcription [6]) and monitoring [8] one often faces with simultaneously unfolding processes which collectively describe the observed phenomena. In such scenarios, a factorial model, where individual processes are modelled by a switching state space model is useful.

The factorial switching state space model consists of $\nu = 1 \ldots W$ models with a shared observation. More precisely,

$$\begin{aligned} r_{k,\nu} &\sim p(r_{k,\nu}|r_{k-1,\nu}) \\ \theta_{k,\nu} &\sim p(\theta_{k,\nu}|\theta_{k-1,\nu}, r_{k,\nu}) \end{aligned}$$

and the observation is given (in the conditionally Gaussian case)

$$y_k \sim p(y_k|\boldsymbol{\theta}_k) = \mathcal{N}(y_k; \sum_{i=\nu}^{W} C_\nu \theta_{k,\nu}, R)$$

where $\boldsymbol{\theta}_k \equiv \theta_{k,1:W}$. Here, to simplify notation, we have assumed $p(y_k|\mathbf{r}_k, \boldsymbol{\theta}_k) = p(y_k|\boldsymbol{\theta}_k)$ where $\mathbf{r}_k \equiv r_{k,1:W}$; the algorithms can be

This research is funded by the EPSRC.

easily modified when this is not the case. Unfortunately, the factorial structure of the model prohibits the computation of the posterior filtering density, even when conditioned on all random variables in the previous time slice.

2. INFERENCE

In the batch case, when all observations are available, we can compute marginals of the posterior (e.g., smoothed estimates) by efficient Rao- Blackwellized [9] Gibbs sampling methods, that only sample from switches r and integrate out the latent continuous state variables [10]. Similarly, the MMAP problem can be attacked by simulated annealing (SA) using a logarithmic cooling schedule [11].

If sequential inference is desirable due to real-time requirements, a Rao- Blackwellized particle filter (RBPF) [12] can be used, that for the conditional Gaussian case is known as the mixture Kalman filter (MKF) [13]. Similarly, a suboptimal breadth first search algorithm can be used for computing MMAP, that is similar to the well known multi hypothesis tracker (MHT).

MKF approximates the conditional filtering density by a collection of Gaussian kernels $p(\boldsymbol{\theta}_k, \mathbf{r}_{1:k}) \approx \sum_{i=1}^{N} \phi^{(i)}(\boldsymbol{\theta}_k; \mathbf{r}_{1:k}^{(i)})$ where each kernel is of form $Z_k^{(i)} \mathcal{N}(\boldsymbol{\theta}_k; \mu_k^{(i)}, \Sigma_k^{(i)})$ with mean $\mu_k^{(i)}$, covariance $\Sigma_k^{(i)}$ and $Z_k^{(i)} = \int d\boldsymbol{\theta}_k \phi^{(i)}$. An alternative representation is the canonical form $\phi^{(i)}(\boldsymbol{\theta}) \equiv \exp\left(-\frac{1}{2}\boldsymbol{\theta}^\top K^{(i)} \boldsymbol{\theta} + \boldsymbol{\theta}^\top h^{(i)} + g^{(i)}\right)$ where K is the precision matrix, h is a vector and g is a scalar. Similarly, a MHT keeps track of a set of trajectories $\mathbf{r}_{1:k}^{(i)}$ (hypotheses) that correspond to likely switch configurations. Omitting technical details, starting with a set of particles/hypotheses $\{\phi_{k-1}^{(i)}\}_{i=1\ldots N}$ where $\phi_{k-1}^{(i)} \equiv \phi^{(i)}(\boldsymbol{\theta}_{k-1}; \mathbf{r}_{1:k-1}^{(i)})$, the following steps are iterated:

Propose: Generate $j = 1 \ldots J$ new particles from each particle

$$\mathbf{r}_k^{(j|i)} \sim q(\mathbf{r}_k | y_k, \mathbf{r}_{k-1}^{(i)})$$

Extend: Conditioned on $\mathbf{r}_k = \mathbf{r}_k^{(j|i)}$, compute new particles and their weights

$$\begin{aligned} \phi_k^{(j|i)} &= p(\mathbf{r}_k | \mathbf{r}_{k-1}^{(i)}) \int d\boldsymbol{\theta}_{k-1} p(y_k|\boldsymbol{\theta}_k) p(\boldsymbol{\theta}_k | \boldsymbol{\theta}_{k-1}, \mathbf{r}_k) \phi_{k-1}^{(i)} \\ w_k^{(j|i)} &= \int d\boldsymbol{\theta}_k \phi_k^{(j|i)} \end{aligned}$$

Prune/Resample: (Optional) Select $\{\phi_k^{(i)}\}_{i=1\ldots N}$ from $\{\phi_k^{(j|i)}\}_{i=1\ldots N}^{j=1\ldots J}$ according to the weights $w_k^{(j|i)}$

Typically, for MKF, the selection schema is by sampling from a categorical distribution proportional to the weights, but other asymptotically consistent schemata are possible. In MHT, we typically retain N particles with the highest weights, but other heuristics, which try to introduce diversity may be used.

In both algorithms, the success of the sequential schema hinges on the quality of the proposal. Intuitively, we would like to make use of the recent observation y_k hence for each particle, we wish to propose from $q = p(\mathbf{r}_k | y_{1:k}, \mathbf{r}_{k-1}^{(i)})$. It turns out for MKF, this choice is indeed optimal in terms of reducing variance of importance weights [12] which is equivalent to choosing the nearest distribution to the exact posterior in the sense of a certain divergence measure. Similarly, for MHT we wish to apply a greedy search mechanism and for each particle select the most likely switch configuration $\mathbf{r}_k^{(i)*} \equiv \arg\max_{\mathbf{r}_k} p(\mathbf{r}_k |\, y_{1:k}, \mathbf{r}_{k-1}^{(i)})$. However, the factorial structure of the model prohibits sampling from this optimal proposal distribution or computing its mode for even a single time slice since the joint state space of the indicators $r_{k,1:W}$ scales exponentially with W.

2.1. Approximating the proposal distribution

The optimal proposal $p(\mathbf{r}_k | y_{1:k}, \cdot)$ is proportional to

$$\int d\theta_{1:W} d\bar{\theta}_{1:W} \phi^{(i)}(\bar{\theta}_{1:W}) \left(\prod_\nu p^{(i)}(r_\nu) p(\theta_\nu | \bar{\theta}_\nu, r_\nu) \right) p(y_k | \theta_{1:W})$$

where we drop the time index k when referring to r_k and use the notation $p^{(i)}(r_\nu) = p(r_{k,\nu} | r_{k-1,\nu}^{(i)})$. Moreover we define $\theta \equiv \theta_k$ and $\bar{\theta} \equiv \theta_{k-1}$. Conditioned on a particular configuration $r_{1:W}$, this integral can be computed in various orders, analogous to forward-backward message passing in HMM's or two filter smoother formulation of Kalman Filter Models. Note that, the propagation is across factors ν rather than time index k. To highlight this distinction, we define *upward* order when we integrate out variables $\theta_{k-1,\nu}$ in the order $\nu = 1, 2, \ldots, W$ and *downward* $\nu = W, W-1, \ldots, 1$. The idea is to exploit the factorial structure of the transition model; our derivation is exactly analogous to the junction-tree algorithm specialised to the factorial hidden Markov model (FHMM) of [14]. However, unlike the FHMM, the messages are tractable because space requirements scale quadratically in contrast to exponentially. We define the following messages:

- **upward:** $\alpha_\nu \equiv p(\hat{y}_{1:k-1}, \theta_{k-1,\nu+1:W}, \theta_{k,1:\nu}, \hat{r}_{1:\nu})$

$$\begin{aligned} \alpha_0 &\equiv \phi^{(i)}(\bar{\theta}_{1:W}) \\ \alpha_\nu &= \int d\bar{\theta}_\nu p^{(i)}(r_\nu) p(\theta_\nu | \bar{\theta}_\nu, r_\nu) \alpha_{\nu-1} \end{aligned}$$

- **downward:** $\beta_\nu \equiv p(\hat{y}_k | \theta_{k-1,\nu+1:W}, \theta_{k,1:\nu}, \hat{r}_{\nu+1:W})$

$$\begin{aligned} \beta_W &\equiv p(y_k | \theta_{1:W}) \\ \beta_{\nu-1} &= \int d\theta_\nu p^{(i)}(r_\nu) p(\theta_\nu | \bar{\theta}_\nu, r_\nu) \beta_\nu \end{aligned}$$

Hence, the conditional distribution is given by

$$p(r_\nu | r_{\neg\boldsymbol{\nu}}, y_{1:k}) \propto \int d\bar{\theta}_{\nu:W}, d\theta_{1:\nu} p^{(i)}(r_\nu) p(\theta_\nu | \bar{\theta}_\nu, r_\nu) \alpha_{\nu-1} \beta_\nu \quad (1)$$

where $\neg\boldsymbol{\nu} \equiv \{1, \ldots, W\} - \{\nu\}$.

2.2. Approximate Inference

In this paper, we investigate two approximate inference methods to approximate the optimal proposal

- A Rao Blackwellized Gibbs sampler [9] (remisicent to [10])
- A Variational approximation of Structured Mean Field type ([15, 16])

The Gibbs sampler relies on constructing a Markov chain where we sample iteratively from full conditional densities

$$r_\nu \sim p(r_\nu | r_1^{(t+1)}, r_2^{(t+1)}, \ldots, r_{\nu-1}^{(t+1)}, r_{\nu+1}^{(t)}, \ldots, r_W^{(t)}, y_{1:k})$$

where t is the iteration index. It is easy to see that this density can be calculated using 1.

Variational Bayes is an alternative approximation method based on deterministic fixed point iterations [17, 16] and have direct links

Algorithm 1 Rao Blackwellized Gibbs sampler/Variational Bayes

for $\nu = 1 \dots W$ **do**
if Method = Gibbs **then**
Sample from transition prior of indicators
$\hat{r}_\nu^{(i,0)} \sim p^{(i)}(r_\nu)$
else
Set the variational approximation on switches
$q_\nu^{(i,0)} \leftarrow p^{(i)}(r_\nu)$
end if
end for
for $\tau = 1$ to MAXEPOCH **do**
Downward Pass: compute and store β messages
$\beta_W^{(\tau)} \leftarrow p(y_k|\boldsymbol{\theta})$
for $\nu = W \dots 1$ **do**
if Method = Gibbs **then**
$\beta_{\nu-1}^{(\tau)} \leftarrow \int d\theta_\nu p^{(i)}(r_\nu = \hat{r}^{(i,\tau-1)}) p(\theta_\nu|\bar{\theta}_\nu, r_\nu = \hat{r}^{(i,\tau-1)}) \beta_\nu^{(\tau)}$
else
$\beta_{\nu-1}^{(\tau)} \leftarrow \int d\theta_\nu \exp\left(\left\langle \log p^{(i)}(r_\nu) p(\theta_\nu|\bar{\theta}_\nu, r_\nu)\right\rangle_{q_\nu^{(i,\tau-1)}}\right) \beta_\nu^{(\tau)}$
end if
end for
Upward Pass
$\alpha_0^{(\tau)} \leftarrow \phi^{(i)}(\bar{\boldsymbol{\theta}})$
for $\nu = 1 \dots W$ **do**
Evaluate the full conditional of the marginal filtering density $p(r_\nu|\hat{r}_{\neg\boldsymbol{\nu}}^{(i,\tau)}, y_{1:t})$
for $c = 1 \dots |r_\nu|$ **do**
$\pi_\nu^{(\tau)}(c) \leftarrow \int d\bar{\theta}_{\nu:W}, d\theta_{1:\nu} p^{(i)}(r_\nu = c) p(\theta_\nu|\bar{\theta}_\nu, r_\nu = c) \alpha_{\nu-1}\beta_\nu$
end for
if Method = Gibbs **then**
Sample the indicator from proposal computed with annealing parameter ρ_τ

$$q_\nu^{(\tau)} \leftarrow (\pi_\nu^{(\tau)})^{\rho_\tau} / \sum_c (\pi_\nu^{(\tau)}(c))^{\rho_\tau} \qquad \hat{r}_\nu^{(i,\tau)} \sim q_\nu^{(\tau)}$$

else
Rescale variational approximation with annealing parameter ρ_τ

$$q_\nu^{(i,\tau)} \leftarrow (\pi_\nu^{(\tau)})^{\rho_\tau} / \sum_c (\pi_\nu^{(\tau)}(c))^{\rho_\tau}$$

end if
if Method = Gibbs **then**
Compute the upward message

$$\alpha_\nu^{(i,\tau)} = \int d\bar{\theta}_\nu p^{(i)}(r_\nu = \hat{r}_\nu^{(i,\tau)}) p(\theta_\nu|\bar{\theta}_\nu, r_\nu = \hat{r}_\nu^{(i,\tau)}) \alpha_{\nu-1}^{(i,\tau)}$$

else
Compute the upward message using average canonical parameters

$$\alpha_\nu^{(i,\tau)} = \int d\bar{\theta}_\nu \exp\left(\left\langle \log p^{(i)}(r_\nu) p(\theta_\nu|\bar{\theta}_\nu, r_\nu)\right\rangle_{q_\nu^{(i,\tau)}}\right) \alpha_{\nu-1}^{(i,\tau)}$$

end if
end for
end for

with the well-known expectation-maximisation (EM) type of algorithms. In our case, mean field boils down to approximating a target posterior with a simple distribution $\mathcal{Q}$ in such a way that the integral becomes tractable. An intuitive interpretation of mean field method is minimising the KL divergence with respect to (the parameters of) $\mathcal{Q}$ where $KL(\mathcal{Q}||\mathcal{P}) = \langle \log \mathcal{Q} \rangle_{\mathcal{Q}} - \left\langle \log \frac{1}{Z_y} \phi_y \right\rangle_{\mathcal{Q}}$ Here, $\mathcal{P} = \phi_y / Z_y$ where the ϕ_y is the integrand in the integral that defines the optimal proposal and Z_y is the unknown normalisation constant. Here, the notation $\langle f(x) \rangle_{p(x)}$ denote the expectation of the function $f(x)$ under the distribution $p(x)$.

In our case, we choose $\mathcal{Q} \equiv q_\theta \prod_{\nu=1}^W q_\nu \equiv q_\theta q_{1:W}$ where q_ν are discrete distributions and $q_\theta = q_\theta(\boldsymbol{\theta}, \boldsymbol{\theta})$. The VB approach leads to the following fixed point equations that need to be iterated until convergence:

$$q_\nu \propto \exp\left(\langle \log \phi_y \rangle_{q_{\neg\boldsymbol{\nu}} q_\theta}\right) \qquad q_\theta \propto \exp\left(\langle \log \phi_y \rangle_{q_{1:W}}\right) \quad (2)$$

where $q_{\neg\boldsymbol{\nu}} \equiv \prod_v q_v / q_\nu$, that is the joint distribution of all factors excluding q_ν. It turns out that (2) can be computed using the same message passing schema, but using expected canonical parameters where expectations are taken w.r.t. $q_{1:W}$. While the starting principles differ, both methods are algorithmically very similar, as detailed in panel 1.

2.3. Example

To motivate our approach, we illustrate the algorithm on a model for polyphonic music. This model is a slightly different version of a model described in [6]. In this model, each factor process models the sound generation mechanism of a pitch with fundamental (angular) frequency ω_ν. The discrete indicators denote onset events where $r_{k,\nu} \in \{$"new", "reg"$\}$. The state vector θ represents the state of an harmonic oscillator. The fundamental frequency of the oscillation is determined by the transition matrix (for the regular regime $r_\nu =$ "reg") has a block diagonal structure as

$$\begin{aligned} A_\nu &\equiv \mathbf{blkdiag}\{\rho_1 B(\omega_\nu)^\top, \dots, \rho_H B(H\omega_\nu)^\top\}^N \\ B(\omega) &\equiv \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix} \end{aligned}$$

where B is a rotation matrix and ρ_h are damping factors such that $0 < \rho_h < 1$ for $h = 1 \dots H$. The observation matrix has a block structure with $C = [C_1 \dots C_\nu \dots C_W]$ where each block C_ν is $N \times 2H$. In turn, each of the blocks C_ν consist of smaller blocks of size 1×2 where the block at $t+1$'th row and h'th double column is given by $\rho_h^t[\cos(h\omega_\nu t) \sin(h\omega_\nu t)]$. The observation noise is isotropic with diagonal covariance R.

The switches control the transition noise variance Q_k. In regular mode, $Q_k = Q(r_{k,\nu} =$ "reg") is small, meaning that the model undergoes its regular damped periodic dynamics. When an onset occurs, the transition noise is has large variance, $Q_k = Q(r_{k,\nu} =$ "new"), and the transition matrix is set to $A_k = 0$. This has the effect of forgetting the past and reinitialising the state vector $\theta_{k,\nu}$. Intuitively, this is a simplification of a physical model where a vibrating string (as represented by $\boldsymbol{\theta}$ in state space form) is plucked by injecting some unknown amount of energy.

In the first illustration, we have sampled from the model a single frame of length $N = 640$ samples that corresponds to 29 msec with sampling frequency $F_s = 22050$ Hz. Each factor is assumed to be a single frequency $H = 1$ distributed geometrically between roughly 100 and 300 Hz (corresponding to Midinotes $30 \dots 50$). In this case, the task is to find the number of frequency components.

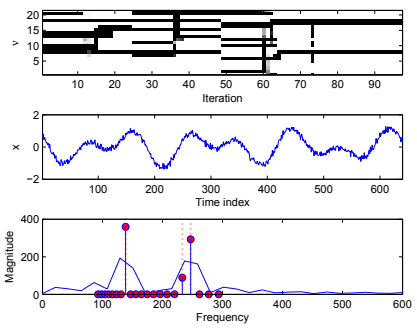

Fig. 1. A typical run of variational approximation. (Top) Variational approximation during iterations; the higher $q_\nu(r = \text{“}new''\text{)}$, the darker the corresponding cell (Middle) Time Domain signal, (Bottom) FFT modulus and true (dashed) and estimated (stem) frequency components.

As can be seen in fig. 1, bottom panel, the frequencies are to close to be resolved by FFT. The top panel shows the variational approximation q_ν, roughly shows the configurations that are visited by the VB method during iterations. Abrupt changes correspond to reinitialisations. In the second experiment, we have generated 100 independent cases from the model. In figure 2.a, we show the distribution of edit distance errors, where we count the number of mismatches between the true and estimated switch configuration and illustrate in 2.b a MHT algorithm that uses VB as a proposal.

3. DISCUSSION

In this paper, we have described an approximate inference method to evaluate the filtering density for a factorial switching state space model and described a stochastic (Gibbs sampling) and a deterministic (Mean field - VB) method. Both methods make use of a message passing schema, where only matrices of size equal to the bandwidth of the transition matrix need to be inverted. The disadvantage in contrast to the direct approach is increased storage requirement: the downward (or upward) messages need to be stored. Our simulations suggest that both methods are comparable, with MCMC slightly superior to VB in terms of quality. However, VB tends to converge in less iteration and with annealing it seems to be a viable and fast candidate. Due to space limitations, further simulation results for this model along with a longer technical note about the details of the algorithm will be made available on our web-site `http://www-sigproc.eng.cam.ac.uk/~atc27/icassp07`.

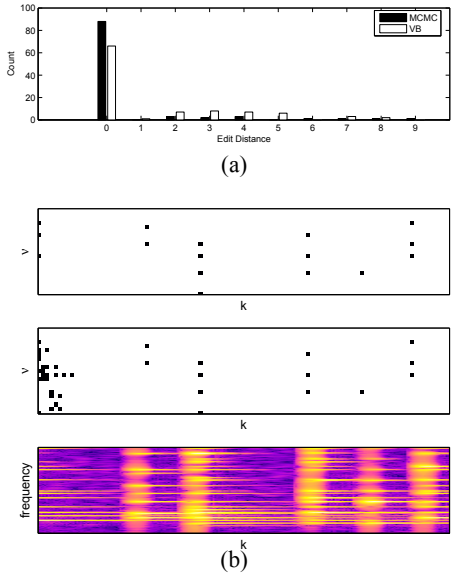

Fig. 2. (a) Comparison of Gibbs sampler and Variational approximation on 100 independent cases in terms of edit distance with $N = 320, H = 5$. In this case, the Gibbs sampler seems to be slightly superior. (b) Sequential inference results (MHT) with (Top) 10 particles, (Middle) 1 particle, (Bottom) STFT modulus of the generated signal. The true piano roll is identical to the one on the top panel.

4. REFERENCES

[1] Fredrik Gustafsson, *Adaptive filtering and change detection*, John Wiley and Sons, Ltd, 2000.

[2] K. P. Murphy, “Switching Kalman filters,” Tech. Rep., Dept. of Computer Science, University of California, Berkeley, 1998.

[3] A. Doucet, S. J. Godsill, and C. P. Robert, “Marginal maximum a posteriori estimation using MCMC,” *Statistics and Computing*, vol. 12, pp. 77–84, 2002.

[4] James D. Park and Adnan Darwiche, “Complexity Results and Approximation Strategies for MAP Explanations,” *Journal of Artificial Intelligence Research*, vol. 21, pp. 101–133, 2004.

[5] P. Fearnhead, “Exact and efficient bayesian inference for multiple changepoint problems,” Tech. Rep., Dept. of Math. and Stat., Lancaster University, 2003.

[6] A. T. Cemgil, H. J. Kappen, and D. Barber, “A Generative Model for Music Transcription,” *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, March 2006.

[7] A. T. Cemgil, “Sequential inference for Factorial Changepoint Models,” in *Nonlinear Statistical Signal Processing Workshop*, Cambridge, UK, 2006, IEEE.

[8] C. K. I. Williams, J. Quinn, and N. McIntosh, “Factorial switching kalman filters for condition monitoring in neonatal intensive care,” in *NIPS 18*. 2006, MIT Press.

[9] G. Casella and C. P. Robert, “Rao-Blackwellisation of sampling schemas,” *Biometrika*, vol. 83, pp. 81–94, 1996.

[10] C. K. Carter and R. Kohn, “Markov Chain Monte Carlo in conditionally Gaussian state space models,” *Biometrika*, vol. 83, no. 3, pp. 589–601, 1996.

[11] E. H. L. Aarts and P. J. M. van Laarhoven, “Statistical cooling: A general approach to combinatorial optimization problems,” *Philips Journal of Research*, vol. 40, no. 4, pp. 193–226, 1985.

[12] A. Doucet, S. Godsill, and C. Andrieu, “On sequential Monte Carlo sampling methods for Bayesian filtering,” *Statistics and Computing*, vol. 10, no. 3, pp. 197–208, 2000.

[13] R. Chen and J. S. Liu, “Mixture Kalman filters,” *J. R. Statist. Soc.*, vol. 10, 2000.

[14] Z. Ghahramani and M. I. Jordan, “Factorial hidden Markov models,” *Machine Learning*, , no. 29, pp. 245–273, 1997.

[15] W. Wiegerinck, “Variational approximations between mean field theory and the junction tree algorithm,” in *UAI (16-th conference)*, 2000, pp. 626–633.

[16] M. Wainwright and M. I. Jordan, “Graphical models, exponential families, and variational inference,” Tech. Rep. 649, Department of Statistics, UC Berkeley, September 2003.

[17] Z. Ghahramani and M. Beal, “Propagation algorithms for variational Bayesian learning,” in *Neural Information Processing Systems 13*, 2000.