SEQUENTIAL INFERENCE FOR FACTORIAL CHANGEPOINT MODELS

A. Taylan Cemgil

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

ABSTRACT

Conditional Gaussian changepoint models are an interesting subclass of jump-Markov dynamic linear systems, in which, unlike the majority of such intractable hybrid models, exact inference is achievable in polynomial time. However, many applications of interest involve several simultaneously unfolding processes with occasional regime switches and shared observations. In such scenarios, a factorial model, where each process is modelled by a changepoint model is more natural. In this paper, we derive a sequential Monte Carlo algorithm, reminiscent to the Mixture Kalman filter (MKF) [1]. However, unlike MKF, the factorial structure of our model prohibits the computation of the posterior filtering density (the optimal proposal distribution). Even evaluating the likelihood conditioned on a few switch configurations can be time consuming. Therefore, we derive a propagation algorithm (upward-downward) that exploits the factorial structure of the model and facilitates computing Kalman filtering recursions in information form without the need for inverting large matrices. To motivate the utility of the model, we illustrate our approach on a large model for polyphonic pitch tracking.

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, [2]. 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.

The simplest change point model can be defined by the following hierarchical probabilistic model[1]

$$\begin{aligned} r_k &\sim p(r_k|r_{k-1}) \\ \theta_k &= [r_k = \text{reg}]\theta_k^{\text{reg}} + [r_k = \text{new}]\theta_k^{\text{new}} \\ y_k &\sim p(y_k|r_k, \theta_k) \end{aligned}$$

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 change indicator for time k where $r_k \in \{\text{new}, \text{reg}\}$.

The model is completed by defining the transition model f and reinitialization model π:

$$\begin{aligned} \theta_k^{\text{reg}} &\sim f(\theta_k|\theta_{k-1}) \\ \theta_k^{\text{new}} &\sim \pi(\theta_k) \end{aligned}$$

When $r_k = \text{new}$, the process switches to a new regime and a new state θ_k vector is drawn from the reinitialization model π; otherwise the state variable obeys its regular dynamics given by f. Each hidden configuration $r_{1:K}$ for some fixed K specifies a certain segmentation hypothesis: a possible model structure 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 is given by

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

The difficulty of this optimisation problem stems from the fact that a potentially intractable integral 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[2]. Lacking any special structure that can be exploited by local message passing algorithms such as Dynamic programming, 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 $r_{1:k}$.

Perhaps surprisingly, there are special nontrivial tractable cases where the optimum can be computed in polynomial

This research is funded by the EPSRC.

[1] We use the notation ["text"] to denote an indicator function that evaluates to 1 (or 0) when the proposition "text" is true (or false).

[2] This is in fact the manifestation of the rather obvious fact that integration and maximisation do not commute.

time/space [5, 6]: One such case is when the transition, reinitialization and observation models are conditionally Gaussian

$$\begin{aligned} f(\theta_k|\theta_{k-1}) &= \mathcal{N}(A_k\theta_{k-1}, Q_k) \\ \pi(\theta_k) &= \mathcal{N}(\theta_k; m_k, V_k) \\ p(y_k|\theta_k) &= \mathcal{N}(y_k; C_k\theta_k, R_k) \end{aligned}$$

where A_k, Q_k denote the transition matrix and noise covariance, C_k, R_k denote observation matrix and noise covariance and m_k, V_k are reinitialization mean and noise covariance. We assume these are known given r_k, i.e. we have $A_k = A(r_k)$, e.t.c. Intuitively, the model is tractable because the integral can be computed analytically *and* when a changepoint occurs, all the past memory in the state variable θ is "forgotten". In this case, it can be shown that one can introduce a deterministic pruning schema that reduces the number of exponentially many filtering potentials $\phi(\theta_k|r_{1:k})$ to a polynomial order and meanwhile guarantees that we will never eliminate the prefix of the MMAP configuration $r^*_{1:k}$ (e.g. see [6]). This exact pruning method hinges on the factorisation of the posterior for the assignment of variables $r_k = \text{new}$ that breaks the direct link between θ_k and θ_{k-1}.

However, for many real-world applications, that involve independent and several simultaneously occurring events, a single changepoint model may not be sufficient. For example, in video surveillance independent objects can enter and exit the scene and we may be interested in infering their trajectory or visual features, given only video data. Another application is in sound analysis and music transcription where independent sound sources occur simultaneously and these need to be separated and segmented jointly [6]. For such scenarios, it is convenient to model each object/source by an independent changepoint model. We call such models *factorial changepoint models.*

2. FACTORIAL CHANGEPOINT MODEL

The factorial changepoint model consists of $\nu = 1 \dots W$ changepoint models with a "common" observation. More precisely,

$$\begin{aligned} r_{k,\nu} &\sim p(r_{k,\nu}|r_{k-1,\nu}) \\ \theta_{k,\nu} &= [r_{k,\nu} = \text{reg}]\theta^{\text{reg}}_{k,\nu} + [r_{k,\nu} = \text{new}]\theta^{\text{new}}_{k,\nu} \end{aligned}$$

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

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

where $\mathbf{r}_k \equiv r_{k,1:W}$ and $\boldsymbol{\theta}_k \equiv \theta_{k,1:W}$. Unfortunately, calculation of MMAP in this model is no longer tractable, since the model degenerates (for the conditional Gaussian case) into a switching Kalman filter (Mixture Kalman filter, Jump Markov Linear Dynamical System) with a rather large latent state space.

2.1. Sequential Inference

One theoretically justifiable inference method for the MMAP problem is simulated annealing (SA). However, SA is an inherently batch method requiring simulation of a Markov Chain with a logarithmic cooling schedule [7]. For switching Kalman filter models, the analytical structure due to conditional Gaussianity can be exploited to design an efficient Rao- Blackwellized MCMC sampler, where it is sufficient to sample from the latent switches r and integrate out the continuous state variables θ analytically [8].

The sequential counterpart of this algorithm is Rao- Blackwellized particle filter (RBPF) [1, 9], that for the conditional Gaussian case is known as the mixture Kalman filter. While RBPF is not directly relevant for computing the MMAP but rather approximating the posterior by a weighted set of samples, we have found empirically that for a given computational cost the solution quality can be significantly better than a naive SA implementation (e.g. see [10]). Moreover, many real-time applications require sequential inference only.

A generic Rao- Blackwellized particle filter approximates the conditional filtering potential by a collection of Gaussian kernels

$$p(\boldsymbol{\theta}_k, \mathbf{r}_{1:k}) \approx \sum_{i=1}^{N} \phi^{(i)}(\boldsymbol{\theta}_k; \mathbf{r}^{(i)}_{1:k})$$

where each kernel is of form $Z^{(i)}_k \mathcal{N}(\boldsymbol{\theta}_k; \mu^{(i)}_k, \Sigma^{(i)}_k)$ with mean $\mu^{(i)}_k$, covariance $\Sigma^{(i)}_k$ and $Z^{(i)}_k = \int d\boldsymbol{\theta}_k \phi_i$.

1. Generate new samples $\mathbf{r}^{(i)}_t$ from $q(\mathbf{r}_t|y_t, \mathbf{r}^{(i)}_{t-1})$.
2. Calculate weights $w^{(i)}_t$ and the normalised weights $\tilde{w}^{(i)}_t$.
3. (Optional resampling:) Randomly select N samples $\mathbf{r}^{(j)}_{new}$ from $\mathbf{r}^{(i)}_t$. Each sample $\mathbf{r}^{(i)}_t$ is selected with probability equal to its normalised weight. The new samples are used further $\mathbf{r}^{(i)}_t \leftarrow \mathbf{r}^{(j)}_{new}$ with weights $w^{(i)}_t = 1$.

However, the factorial structure of the model prohibits the computation of the optimal proposal distribution (the filtering distribution) for the (i)'th particle $q = p(\mathbf{r}_k|y_{1:k}, \mathbf{r}^{(i)}_{k-1})$ for even a single time slice since the joint state space of the indicators $r_{k,1:W}$ scales exponentially with W. Indeed, when W is large, we need to solve at each time slice a variable selection problem [11].

3. APPROXIMATING THE FILTERING DISTRIBUTION

Suppose at time slice $k-1$, we have a particle $\phi^{(i)}(\boldsymbol{\theta}_{k-1}; \mathbf{r}^{(i)}_{1:k-1})$, and we wish to evaluate one step ahead filtering density

$$\begin{aligned} p(\mathbf{r}_k|y_{1:k}, \mathbf{r}^{(i)}_{1:k-1}) \propto\ & p(\mathbf{r}_k|\mathbf{r}^{(i)}_{k-1}) \int d\boldsymbol{\theta}_k d\boldsymbol{\theta}_{k-1} p(y_k|\boldsymbol{\theta}_k) \\ & p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{k-1}, \mathbf{r}_k)\phi^{(i)} \end{aligned}$$

to compute the proposal distribution. In general, we have to evaluate this integral for each of the 2^W configurations of $\mathbf{r}_k$. However, in practice, changepoints are rare and it will be relatively unlikely that two or more factors will have changepoints exactly at the same timeslice. Hence, we can "bias" our sampling towards zero or one changepoint configurations $\hat{\mathcal{R}}_1$ defined as $\hat{\mathcal{R}}_1 \equiv \{\mathbf{r}_k : \sum_{i=1}^W [r_{k,i} = \text{new}] \leq 1\}$. We will denote the zero changepoint configuration as $\mathbf{r}^{\text{reg}}$; hence $\hat{\mathcal{R}}_1$ is the 1-neighbourhood of $\mathbf{r}^{\text{reg}}$ (in terms of Hamming distance). In principle, we could evaluate the filtering density pointwise for each of the $W+1$ configurations in $\hat{\mathcal{R}}_1$ separately. However, this is still time consuming and potentially numerically unstable when W, the number of changepoint models, is large. The numerical instability becomes more pronounced when we need to execute the Kalman recursions in information form; i.e. when particles are represented by their canonical parameters (i.e. inverse covariance matrix etc.).

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 [12]. However, unlike the FHMM, the intermediate calculations are tractable because space requirements scale quadratically in contrast to exponentially. We can evaluate the conditional likelihood of all configuration $\mathbf{r}_k \in \hat{\mathcal{R}}_1$ in one "upward-downward" pass as explained below.

The transition equation has the following factorial form:

$$p(\boldsymbol{\theta}_k, \mathbf{r}_k | \boldsymbol{\theta}_{k-1}, \mathbf{r}_{k-1}) = \prod_{\nu=1}^{W} p(\theta_{k,\nu} | \theta_{k-1,\nu}, r_{k,\nu}) p(r_{k,\nu} | r_{k-1,\nu})$$

Hence the optimal proposal density is proportional to

$$\int d\boldsymbol{\theta}_k d\boldsymbol{\theta}_{k-1} p(y_k | \boldsymbol{\theta}_k) \left(\prod_\nu p^{(i)}(r_\nu) p(\theta_{k,\nu} | \theta_{k-1,\nu}, r_\nu) \right) \phi^{(i)}(\boldsymbol{\theta}_{k-1})$$

where we drop the time index k when refering to r_k and use the notation $p^{(i)}(r_\nu) = p(r_{k,\nu} | r^{(i)}_{k-1,\nu})$. Conditioned on a particular configuration $r_{1:W}$, this integral can be computed in various orders. We call the upward (analogous to forward) pass when we integrate out variables in the order $\theta_{k-1,1}, \theta_{k-1,2}, \ldots$. Alternatively, since the expression is entirely symmetric, we also define a downward pass (analogous to backward) where we integrate out in the order $\theta_{k,W}, \theta_{k,W-1}, \ldots$.

For the upward-downward pass, we define the intermediate potentials (where we implicitly condition or $\mathbf{r} = \mathbf{r}^{\text{reg}}$) for $\nu = 1 \ldots W$

- **Upward message** $\alpha_\nu \equiv p(y_{1:k-1}, \theta_{k,1:\nu}, \theta_{k-1,\nu+1:W})$

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

- **Downward message** $\beta_\nu \equiv p(y_k | \theta_{k,1:\nu}, \theta_{k-1,\nu+1:W})$

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

Hence, the conditional likelihood of a configuration is

$$p(r_\nu | r_{\neg\boldsymbol{\nu}}, y_{1:k}) \propto \int \theta_{k,1:\nu}, \theta_{k-1,\nu+1:W} \alpha_\nu \beta_\nu \qquad (1)$$

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

The algorithm is as follows:

- Compute and store β_ν for $\nu = W \ldots 1$ where $\mathbf{r} = \mathbf{r}^{\text{reg}}$
- for $\nu = 1 \ldots W$
  - Compute $\alpha_\nu(r_\nu = \text{"reg"})$ and $\alpha^{\text{new}}_\nu(r_\nu = \text{"new"})$
  - Compute $p(r_\nu = \text{"new"} | r_{\neg\boldsymbol{\nu}} = \text{"reg"}, y_{1:k})$ using α^{new}_ν in Eq.1
- Compute $p(r_\nu = \text{"reg"} | r_{\neg\boldsymbol{\nu}} = \text{"reg"}, y_{1:k})$ with $\nu = W$ using Eq.1

The advantage of this organisation is that the proposal can be evaluated for every configuration $\mathbf{r}_k \in \hat{\mathcal{R}}_1$ directly. To see the other advantage, consider the Kalman prediction in information form

$$K^{[\nu]} = \Lambda - F^\top S^{-1} F$$

where $\Lambda = \mathbf{blkdiag}\{K_{22}^{[\nu-1]}, Q_\nu^{-1}\}$ and

$$\begin{aligned} F &= \begin{pmatrix} K_{12}^{[\nu-1]} & -\mathbf{A}_\nu^\top Q_\nu^{-1} \end{pmatrix} \\ S &= K_{11}^{[\nu-1]} + \mathbf{A}_\nu^\top Q_\nu^{-1} \mathbf{A}_\nu \end{aligned}$$

Here[3], $K_{11}^{[\nu-1]}$ is the block in the precision matrix K of $\alpha_{\nu-1}$ that corresponds to $\theta_{k-1,\nu}$ where $\alpha_{\nu-1} = \exp(-\frac{1}{2}\boldsymbol{\theta}^\top K \boldsymbol{\theta} + h^\top \boldsymbol{\theta} + g)$. $K_{22}^{[\nu-1]}$ is the partition corresponding to the remaining variables and $K_{12}^{[\nu-1]}$ correspond cross terms. Since S is "small" and F is "thin", we can proceed by low-rank downdates. Hence, with careful programming, α and β can be computed rather efficiently.

3.1. 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 changepoint process models the sound generation mechanism of a pitch with fundamental (angular) frequency ω_ν. The discrete indicators $r_{k,\nu}$ denote onset events. 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 $r_\nu = \text{"reg"}$) has a block diagonal structure as

$$A_\nu \equiv \mathbf{blkdiag}\{\rho_1 B(\omega_\nu)^\top, \ldots, \rho_H B(H\omega_\nu)^\top\}^N$$

[3] Other canonical parameters are given as $h^{[\nu]} = \begin{pmatrix} h_2^{[\nu-1]} \\ 0 \end{pmatrix} - F^\top S^{-1} h_1^{[\nu-1]}$ and $g^{[\nu]} = g^{[\nu-1]} - (1/2)\log|2\pi Q_\nu|$

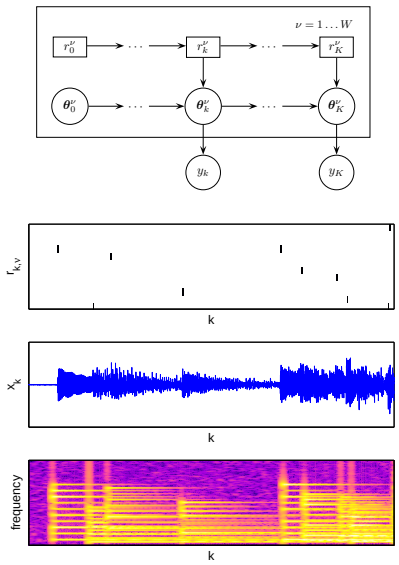

Fig. 1. (Top) The graphical model of the factorial changepoint model, the rectangle denotes a plate, W copies of the nodes inside, (Bottom) A typical sample generated from the model –from top to bottom– piano-roll (indicators $r_{k,\nu}$ where black="new" (onset)), the acoustic signal $\mathbf{x}_k$ and its spectrogram. The task is to find $r^*_{1:K}$ given $y_{1:K}$.

where B is a rotation matrix[4] 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 changepoint mechanism controls the transition noise variance Q_k. When there is no regime switch, $Q_k = Q(r_{k,\nu} =$ "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 $\boldsymbol{\theta}$) is plucked by injecting some unknown amount of energy.

Due to space limitations, 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/nsspw`.

[4] $B(\omega) \equiv \begin{pmatrix} \cos(\omega) & -\sin(\omega) \\ \sin(\omega) & \cos(\omega) \end{pmatrix}$

4. DISCUSSION

In this paper, we have described a time efficient method to evaluate the filtering density for a factorial changepoint model. The advantage of the method is that the likelihood of a configuration and its 1-neighbours can be computed efficiently and numerically stable – 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.

Clearly, the upward-downward propagation can be used to compute the likelihood of all configurations with Hamming distance one to an arbitrary configuration, not only the zero changepoint configurations $\mathbf{r}^{\text{reg}}$. This could be used to design an off-line Rao-Blackwellized Metropolis algorithm.

One other alternative approach, that we have not addressed here is to compute a variational approximation to the exact filtering density by variational methods such as mean field or expectation propagation [13]. The former of these approaches requires the propagation equations to be in information form, hence upward-downward algorithm is useful here, too.

5. REFERENCES

[1] R. Chen and J. S. Liu, "Mixture Kalman filters," *J. R. Statist. Soc.*, vol. 10, 2000.

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

[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] 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.

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

[9] 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.

[10] A. T. Cemgil and H. J. Kappen, "Monte Carlo methods for Tempo Tracking and Rhythm Quantization," *Journal of Artificial Intelligence Research*, vol. 18, pp. 45–81, 2003.

[11] S. J. Godsill, "On the relationship between Markov chain Monte Carlo methods for model uncertainty," *J. Comp. Graph. Stats*, vol. 10, no. 2, pp. 230–248, 2001.

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

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