Empirical Bayesian Change Point Detection

Ulrich Paquet UPAQUET@SUN.AC.ZA
Department of Process Engineering
University of Stellenbosch
South Africa

Editor: xxxx

Abstract

This paper explores a Bayesian method for the detection of sudden changes in the generative parameters of a data series. The problem is phrased as a hidden Markov model, where change point locations correspond to unobserved states, which grow in number with the number of observations. Our interest lies in the marginal change point posterior density. Rather than optimize a likelihood function of model parameters, we adapt the Baum-Welch algorithm to maximize a bound on the log marginal likelihood with respect to prior hyperparameters. This empirical Bayesian approach allows scale-invariance, and can be viewed as an expectation maximization algorithm for hyperparameter optimization in conjugate exponential models with latent variables. The expectation and maximization steps make respective use of variational and concave-convex inner loops. A judicious choice of change point prior allows for fast recursive computations on a graphical model. Results are shown on a number of real-world data sets.

Keywords: product partition model, hidden Markov model, variational Bayes, empirical Bayes, expectation maximization

1. Introduction

The task of determining whether a sudden change occurred in the generative parameters of a time series finds application in many areas. In industrial control, one may want to detect when there was a shift in the parameters generating certain quality measurements, whereas in finance one may be interested in whether there are change points in the volatilities of stock returns. Similar problems arise in for example disease mapping, robotics, and econometrics.

A useful approach to modeling change points is through the product partition model (Barry & Hartigan, 1992)—which assumes that time-series data can be partitioned into independent and identically distributed (i.i.d.) partitions, demarcated by the points where the data's generative parameters change—and variants thereof. The model, with its hidden partition variables, naturally adapts to inference through Gibbs sampling (Loschi & Cruz, 2002; Robert & Casella, 2004, chapter 11), and can be treated as a Hidden Markov Model (HMM) for the same reason (Chib, 1998). In the discussion following Lai (1995), S. L. Lauritzen mentioned the idea of treating change point problems through recursive computations on graphical models (Lauritzen & Spiegelhalter, 1998), and is the approach adopted here.

This paper builds on the online algorithm of Adams & MacKay (2006), who cast the product partition model into a Bayesian graphical model. Their model is equivalent to a

HMM with a possibly infinite number of hidden states, as there can be as many change points as data observations. The setting has the advantage over other HMM-based models as the number of change points does not have to be pre-specified (see e.g. Chib (1998)), although reversible jump Markov chain Monte Carlo methods (Green, 1995) can always be implemented in such models.

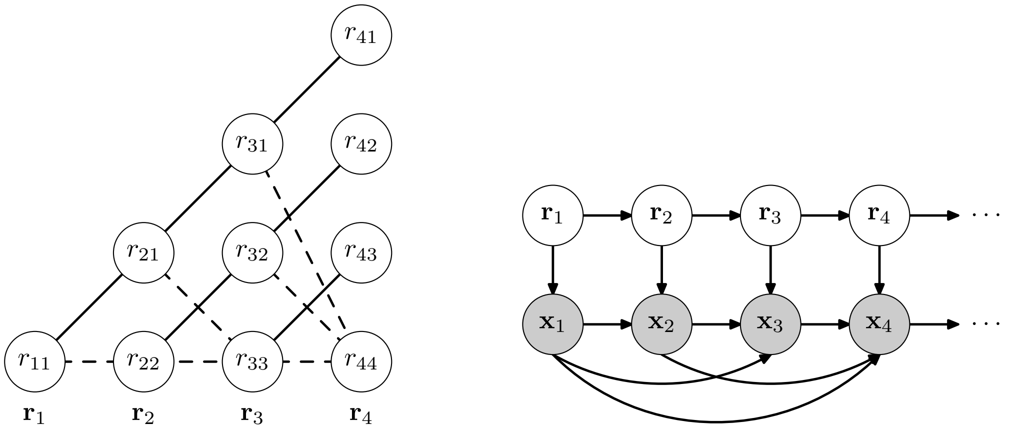
This paper's main contribution is an empirical Bayesian treatment of the prior hyperparameters (MacKay, 1992; Carlin & Louis, 2000), which relies on the standard addition of a "backwards loop" to Adams & MacKay (2006)'s algorithm. Our choice of hyperparameters is crucial in the successful detection of change points, as the posterior over possible data partitions is a *marginal* density, coming from an average over all possible generative parameters. This is illustrated in section 5.1. We can construct a lower bound on the marginal density, and this gives us a practical handle on the log marginal likelihood. We derive an expectation maximization (EM) algorithm to maximize the log marginal likelihood, and it can be seen as a generalization of the Baum-Welch algorithm for maximizing the volume under a function. Both steps in the EM algorithm require subloops, and we use a variational treatment of latent variables in the E-step, and introduce a concave-convex procedure for the M-step. The forward-backward HMM scheme, as well as the M-step, require the restriction of generative probabilities to the exponential family of models with conjugate priors.

We proceed by describing the details of the change point model in section 2, including the change point prior which gives a sparse structure to the HMM. In section 3 we present familiar HMM algorithms for recursive computations on a graphical model. In section 4 we present a novel and effective scheme to maximize the log marginal likelihood with respect to the prior hyperparameters. Section 5 contains results from various fields, and we conclude the paper with a discussion of related ideas in section 6.

2. Change point model

This paper revolves around detecting possible multiple change points in a time series $\mathcal{D} = \{\mathbf{x}_t\}_{t=1}^T$, based upon a rephrasing of the *product partition model* of Barry & Hartigan (1993). We assume that the data can be partitioned into contiguous sequences of *runs*, such that all the data in the sequence starting at time i were generated by parameters $\boldsymbol{\theta}_i \sim p(\boldsymbol{\theta}_i|\boldsymbol{\eta})$, with $\boldsymbol{\eta}$ being hyperparameters shared between all the priors.

This generative model can be described by the states of a (possibly infinite) hidden Markov model (HMM) in the following way: $\mathbf{x}_t$ can potentially be a member of a sequence starting at any time $i = 1, \ldots, t$, and therefore we let the latent *run* state $\mathbf{r}_t \in \{0, 1\}^t$, with $\sum_{i=1}^t r_{ti} = 1$, be a multinomial variable specifying which parameter setting $\boldsymbol{\theta}_i$ was responsible for generating $\mathbf{x}_t$, or of which run sequence it is a part. By constraining the transition probabilities of the unobserved states to allow only movement to the same state i, or a change to an entirely new state t, the latent states naturally partition the data into contiguous runs. Figure 1(a) illustrates the structure of the latent variables; a data point $\mathbf{x}_t$ could be part of any existing run starting at a time $i = 1, \ldots, t-1$, with $r_{ti} = 1$, or *a change point could have occurred* and $\mathbf{x}_t$ is the first example that we see after the change, and hence $r_{tt} = 1$. Therefore change points occur when we move from one run or data partition to the next, implying that the parameters responsible for generating the data have changed



(a) The structure of the change point prior. A run length can either increase, or be reset to zero. Notice that $r_{ti} = 1$ here means that we are at time t, and that the present data point comes from state i—which is the same as saying that it *still* comes from the stream of data that started with a first $\mathbf{x}_i$ at time i.

(b) With the parameters $\mathbf{\Theta}$ integrated out, the data points are coupled in the graphical model. The use of conjugate exponential models allows us to still do analytically tractable inference in this model.

Figure 1: Graphical models for the change point prior, and the HMM used in section 3. A Bayesian network that includes $\mathbf{\Theta}$, $\boldsymbol{\eta}$, and H is illustrated in figure 2.

abruptly. This model assumes that data $\mathcal{D}_{i:t} = \{\mathbf{x}_\tau\}_{\tau=i}^t$ are i.i.d. *given* the state $r_{ti} = 1$,

$$p(\mathcal{D}_{i:t}|\mathbf{\Theta}, r_{ti} = 1) = \prod_{\tau=i}^{t} p(\mathbf{x}_\tau|\boldsymbol{\theta}_i),$$

and is independent of all other parameters $\tau \neq i$ in $\mathbf{\Theta} = \{\boldsymbol{\theta}_\tau\}_{\tau=1}^T$. This is an alternative way of defining "non-overlapping product partitions", or stating that the data points are i.i.d. *within* a specific partition.

2.1 Inference in the model

Our main interest is in determining when change points occurred, irrespective of the specific parameters $\mathbf{\Theta}$, although they can also be inferred. If $\mathbf{R} = \{\mathbf{r}_t\}_{t=1}^T$, we are going to determine the marginals $p(\mathbf{r}_t|\mathcal{D})$ from the posterior

$$p(\mathbf{R}|\mathcal{D}, \boldsymbol{\eta}, H) = \frac{p(\mathcal{D}|\mathbf{R}, \boldsymbol{\eta})p(\mathbf{R}|H)}{p(\mathcal{D}|\boldsymbol{\eta}, H)} \ , \tag{1}$$

where $\boldsymbol{\eta}$ and H are prior hyperparameters for the model. Their role is illustrated in the Bayesian network in figure 2. The likelihood is independent of $\mathbf{\Theta}$ and comes from a higher-level average

$$p(\mathcal{D}|\mathbf{R}, \boldsymbol{\eta}) = \int p(\mathcal{D}|\mathbf{\Theta}, \mathbf{R})p(\mathbf{\Theta}|\boldsymbol{\eta})\, d\mathbf{\Theta} \ ,$$

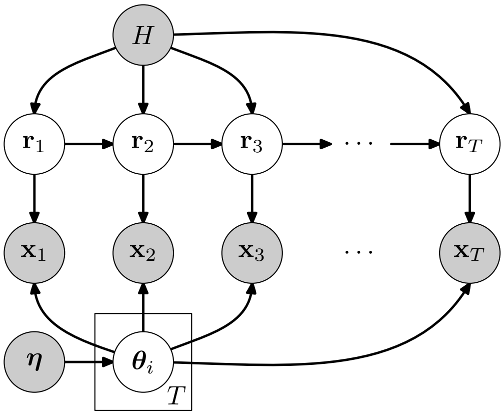


Figure 2: A Bayesian network for the joint density in (3). The *plate* (square) indicates T i.i.d. replicates of $\boldsymbol{\theta}_i$. Shaded and unshaded nodes are used to respectively indicate observed and hidden random variables. The hyperparameters H and $\boldsymbol{\eta}$ are set to maximize the log marginal likelihood in (2).

where $p(\boldsymbol{\Theta}|\boldsymbol{\eta})$ is used as shorthand for $\prod_{i=1}^{T} p(\boldsymbol{\theta}_i|\boldsymbol{\eta})$. Notice that $\boldsymbol{\eta}$ is a prior hyperparameter *shared* between all the $\boldsymbol{\theta}_i$s. Figure 1(b) illustrates a graphical model for the joint density $p(\mathbf{R}, \mathcal{D}|\boldsymbol{\eta}, H)$ after averaging over $\boldsymbol{\Theta}$; notice now that the data couples in the HMM. To treat this in a tractable way, we follow Adams & MacKay (2006) by using conjugate prior distributions and likelihood functions. This allows us to implement standard HMM and message passing algorithms to do inference, as data can effectively be incorporated into various posteriors $p(\boldsymbol{\theta}_i|\mathcal{D})$ through natural parameter updates.

The successful identification of change points depends on the hyperparameters $\boldsymbol{\eta}$—which will have to be on the same scale as the data, for example—and H, and can be set by maximizing the *log marginal likelihood* or log numerator

$$\log p(\mathcal{D}|\boldsymbol{\eta}, H) = \log \left[\sum_{\mathbf{R}} p(\mathbf{R}|H) \int p(\mathcal{D}|\boldsymbol{\Theta}, \mathbf{R}) p(\boldsymbol{\Theta}|\boldsymbol{\eta}) \, d\boldsymbol{\Theta} \right] \tag{2}$$

in (1). This is an *empirical Bayes* approach (Carlin & Louis, 2000), also known as maximizing the evidence (MacKay, 1992), or type II maximum likelihood, and is based on estimating the prior hyperparameters to best explain the observed data. There is no closed-form solution to maximize this quantity with respect to the hyperparameters, and we resort to an expectation maximization (EM) scheme to firstly lower bound it, and then maximize the bound. The EM scheme has been used before, for example in the context of tracking, by Vasconcelos & Lippman (2001).

We provide an outline of the empirical Bayes algorithm to both infer the marginals $p(\mathbf{r}_t|\mathcal{D})$ in (1), and maximize a bound on (2), in algorithm 1. It is presented merely as a framework to show where each of the coming sections play a role.

Algorithm 1 Empirical Bayesian change point detection

```
initialize: $\boldsymbol{\eta}^{(0)}$, $H^{(0)}$, $k = 0$.
repeat
    $k \leftarrow k + 1$.
    section 3: given $\boldsymbol{\eta}^{(k-1)}$ and $H^{(k-1)}$, determine $p(r_{ti} = 1|\mathcal{D})$ and $p(r_{ti} = 1, r_{t+1,j} = 1|\mathcal{D})$ for all $t$.
    section 4.1: given $p(r_{ti} = 1|\mathcal{D})$ and $p(r_{ti} = 1, r_{t+1,j} = 1|\mathcal{D})$, determine $q^{(k)}(\mathbf{R})$.
    section 4.2: given $q^{(k)}(\mathbf{R})$, determine $H^{(k)}$ and $\boldsymbol{\eta}^{(k)}$.
until convergence
```

2.2 The necessary distributions

Given a latent variable allocation $\mathbf{R}$, the likelihood function of observing $\mathcal{D}$ completely factorizes, resulting in the joint distribution $p(\mathcal{D}, \mathbf{R}, \boldsymbol{\Theta}|\boldsymbol{\eta}, H)$ over all the random variables being

$$p(\mathcal{D}|\mathbf{R}, \boldsymbol{\Theta})p(\mathbf{R}|H)p(\boldsymbol{\Theta}|\boldsymbol{\eta}) = \prod_{t=1}^{T} \left(\prod_{i=1}^{t} p(\mathbf{x}_t|\boldsymbol{\theta}_i)^{r_{ti}} \right) \times p(\mathbf{r}_1|H) \prod_{t=2}^{T} p(\mathbf{r}_t|\mathbf{r}_{t-1}, H) \times \prod_{i=1}^{T} p(\boldsymbol{\theta}_i|\boldsymbol{\eta}) \ . \tag{3}$$

The change point prior is crucial, and defines message passing on the lattice presented in figure 1(a). In this paper we conveniently choose the run length prior to follow a geometric sequence,

$$p(\mathbf{r}_t|r_{t-1,i} = 1, H) = \begin{cases} 1 - H & \text{if } r_{ti} = 1 \\ H & \text{if } r_{tt} = 1 \\ 0 & \text{otherwise} \end{cases}$$

so that from state $r_{t-1,i} = 1$ we can either stay in the same state with probability $1 - H$, or a change point can occur with probability H. For the sake of later maximizing a lower bound on (2) with respect to H, we will rewrite the prior as[1]

$$p(\mathbf{r}_t|\mathbf{r}_{t-1}, H) = H^{r_{tt}} \prod_{i=1}^{t-1} \left[r_{t-1,i}(1 - H) \right]^{r_{ti}} \ .$$

The prior is first-order Markov but memoryless given a previous state; Adams & MacKay (2006) also discuss memory-based priors that are a function of the number of observations already associated with state i. Without loss of generality we assume that observations start in the first state with $p(r_{11} = 1)$—and therefore $p(\mathbf{r}_1) = 1$ as there is only one state—giving $p(\mathbf{R}|H) = \prod_{t=2}^{T} p(\mathbf{r}_t|\mathbf{r}_{t-1}, H)$.

2.2.1 Conjugate exponential models

To tractably compute the marginals $p(\mathbf{r}_t|\mathcal{D})$, we assume that the data is generated from an exponential-family likelihood with a conjugate prior. In other words, we can write a prior in terms of its natural parameters $\boldsymbol{\eta}$ and its sufficient statistics $\boldsymbol{\phi}(\boldsymbol{\theta})$ as

$$p(\boldsymbol{\theta}|\boldsymbol{\eta}) = \exp\left\{ \boldsymbol{\eta}^{\top} \boldsymbol{\phi}(\boldsymbol{\theta}) - \log Z(\boldsymbol{\eta}) \right\} \quad \text{with} \quad Z(\boldsymbol{\eta}) = \int \exp\left\{ \boldsymbol{\eta}^{\top} \boldsymbol{\phi}(\boldsymbol{\theta}) \right\} d\boldsymbol{\theta} \ .$$

[1] Seeing that $r \log r \to 0$ as $r \to 0$, this formulation of the prior uses the convention that $0^0 = 1$.

In certain cases it may be necessary to multiply further constraints $h(\boldsymbol{\theta})$ into this formulation, say when $\boldsymbol{\theta}$ is required to be nonnegative or sum to one. A likelihood of the form

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{t=1}^{T} p(\mathbf{x}_t|\boldsymbol{\theta}) = \prod_{t=1}^{T} \exp\left\{\mathbf{u}(\mathbf{x}_t)^\top \boldsymbol{\phi}(\boldsymbol{\theta}) - \log \widetilde{Z}(\mathbf{u}(\mathbf{x}_t))\right\}$$

will cause the posterior $p(\boldsymbol{\theta}|\mathcal{D}) = p(\boldsymbol{\theta}|\boldsymbol{\eta}_{\mathsf{post}})$ to be of the same distribution as the prior, as $p(\boldsymbol{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\eta})$ allows the natural parameter to be updated with a simple update rule, $\boldsymbol{\eta}_{\mathsf{post}} = \boldsymbol{\eta} + \sum_{t=1}^{T} \mathbf{u}(\mathbf{x}_t)$.

Example. In this paper we model $\mathbf{x} \in \mathbb{R}^D$ as a Gaussian random variable with mean $\boldsymbol{\mu}$ and precision matrix (inverse covariance) $\boldsymbol{\Lambda}$, so that the likelihood $p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda})$ of observing a data point is

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) = \exp\left\{-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} + \mathbf{x}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} - \frac{1}{2}\operatorname{tr}[\mathbf{x}\mathbf{x}^\top \boldsymbol{\Lambda}] + \frac{1}{2}\log|\boldsymbol{\Lambda}| - \frac{D}{2}\log(2\pi)\right\} .$$

A conjugate prior on the parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \boldsymbol{\Lambda}\}$ is the Normal-Wishart distribution, $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}) = \mathcal{N}(\boldsymbol{\mu}|\mathbf{m}, (v\boldsymbol{\Lambda})^{-1})\mathcal{W}(\boldsymbol{\Lambda}|a, \mathbf{B})$, also written as

$$\begin{aligned}\mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathbf{m}, v, a, \mathbf{B}) = \exp\Bigg\{&-\frac{1}{2}v\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} + v\mathbf{m}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} - \operatorname{tr}\left[\left(\mathbf{B} + \frac{1}{2}v\mathbf{m}\mathbf{m}^\top\right)\boldsymbol{\Lambda}\right] \\ &+ \left(a - \frac{D}{2}\right)\log|\boldsymbol{\Lambda}| - \log Z_{\mathcal{NW}}(\boldsymbol{\eta})\Bigg\} .\end{aligned}$$

The sufficient statistics $\boldsymbol{\phi}(\boldsymbol{\theta})$ are therefore $-\frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu}$, $\boldsymbol{\Lambda}\boldsymbol{\mu}$, $-\boldsymbol{\Lambda}$, and $\log|\boldsymbol{\Lambda}|$. In this form the natural parameters $\boldsymbol{\eta}$ are v, $v\mathbf{m}$, $\mathbf{B} + \frac{1}{2}v\mathbf{m}\mathbf{m}^\top$, and $a - \frac{D}{2}$, with the normalizer or partition function defined as

$$Z_{\mathcal{NW}}(\boldsymbol{\eta}) = \left(\frac{2\pi}{v}\right)^{D/2} \pi^{D(D-1)/4} \, |\mathbf{B}|^{-a} \prod_{d=1}^{D} \Gamma\left(a + \frac{1-d}{2}\right) .$$

Note that we implicitly assume constraints $a > (D-1)/2$, $v > 0$, and $\mathbf{B}$ being positive definite, to hold.

The natural parameters of the posterior are found by updating the prior parameters $\boldsymbol{\eta}$,

$$v_{\mathsf{post}} = v + T \qquad v_{\mathsf{post}}\mathbf{m}_{\mathsf{post}} = v\mathbf{m} + \sum_{t=1}^{T} \mathbf{x}_t$$

$$\mathbf{B}_{\mathsf{post}} + \frac{1}{2}v_{\mathsf{post}}\mathbf{m}_{\mathsf{post}}\mathbf{m}_{\mathsf{post}}^\top = \mathbf{B} + \frac{1}{2}v\mathbf{m}\mathbf{m}^\top + \frac{1}{2}\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t^\top \qquad a_{\mathsf{post}} - \frac{D}{2} = a - \frac{D}{2} + \frac{T}{2} .$$

Solving for the posterior parameters from the above natural parameters gives $p(\boldsymbol{\mu}, \boldsymbol{\Lambda}|\mathcal{D}) = \mathcal{NW}(\boldsymbol{\mu}, \boldsymbol{\Lambda} \,|\, \mathbf{m}_{\mathsf{post}}, v_{\mathsf{post}}, a_{\mathsf{post}}, \mathbf{B}_{\mathsf{post}})$, and the elegance of conjugate exponential models become clear: we only need to track the natural parameters as data arrive.

2.2.2 Predictive distributions on the lattice

The predictive distribution $p(\mathbf{x}_t | r_{ti} = 1, \mathcal{D}_{1:t-1})$ plays a key role in the standard HMM algorithms that follow in section 3. Given the hidden state, $\mathbf{x}_t$ is only dependent on data associated with the sequence starting at time i, and therefore simplifies as

$$p(\mathbf{x}_t | r_{ti} = 1, \mathcal{D}_{1:t-1}) = p(\mathbf{x}_t | \mathcal{D}_{i:t-1}) = \int p(\mathbf{x}_t | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i | \mathcal{D}_{i:t-1}) \, d\boldsymbol{\theta}_i = \int p(\mathbf{x}_t | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i | \boldsymbol{\eta}_i^{(t)}) \, d\boldsymbol{\theta}_i \ .$$

Instead of $\boldsymbol{\eta}_{\mathsf{post}}$, we use $\boldsymbol{\eta}_i^{(t)}$ to indicate $p(\boldsymbol{\theta}_i | \mathcal{D}_{i:t-1})$'s parameters for $i = 1, \ldots, t$. Therefore $\boldsymbol{\eta}_1^{(t)}$ includes data $\mathcal{D}_{1:t-1}$ into the posterior, $\boldsymbol{\eta}_2^{(t)}$ includes data $\mathcal{D}_{2:t-1}$, and so forth, until $\boldsymbol{\eta}_t^{(t)}$ is equal to the prior $\boldsymbol{\eta}$. As the predictive density depends on the hyperparameters $\boldsymbol{\eta}_i^{(t)}$, we shall use the shorthand

$$p(\mathbf{x}_t | r_{ti} = 1, \mathcal{D}_{1:t-1}) = p(\mathbf{x}_t | \boldsymbol{\eta}_i^{(t)}) \ .$$

Example. Continuing the initial example, the posterior distribution $p(\boldsymbol{\theta}_i | \boldsymbol{\eta}_i^{(t)})$ of the mean vector $\boldsymbol{\mu}_i$ and precision matrix $\boldsymbol{\Lambda}_i$ will be Normal-Wishart, and therefore the predictive density follows a student-t distribution:

$$p(\mathbf{x}_t | \boldsymbol{\eta}_i^{(t)}) = \int p(\mathbf{x}_t | \boldsymbol{\theta}_i) p(\boldsymbol{\theta}_i | \boldsymbol{\eta}_i^{(t)}) \, d\boldsymbol{\theta}_i = \mathcal{T}\left(\mathbf{x}_t \,\middle|\, \mathbf{m}_i^{(t)}, \frac{v_i^{(t)} + 1}{v_i^{(t)}} \frac{2\mathbf{B}_i^{(t)}}{2a_i^{(t)} - D + 1}, 2a_i^{(t)} - D + 1 \right) \ ,$$

where the multivariate t-distribution is given by

$$\begin{aligned} \mathcal{T}(\mathbf{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu) &= \frac{1}{Z_{\mathcal{T}}} \exp\left\{ -\frac{\nu + D}{2} \log\left(1 + \frac{1}{\nu} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right) \right\} \\ Z_{\mathcal{T}}(\boldsymbol{\Sigma}, \nu) &= (2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2} \left(\frac{\nu}{2} \right)^{D/2} \frac{\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu + D}{2})} \ . \end{aligned}$$

3. Forward-backward algorithms

In this section we determine the marginal and pairwise marginal densities, $p(\mathbf{r}_t | \mathcal{D})$ and $p(\mathbf{r}_{t-1}, \mathbf{r}_t | \mathcal{D})$, using standard forward-backward algorithms (Baum & Egon, 1967; Rabiner, 1989). These algorithms can also be interpreted as message passing routines on the Bayesian network shown in figure 1(b). In this setting Adams & MacKay (2006)'s online algorithm is equivalent to the forward pass presented in section 3.1.1.

3.1 The marginal densities

The marginal densities $p(\mathbf{r}_t | \mathcal{D})$ can be written in a standard form,

$$p(\mathbf{r}_t | \mathcal{D}) = \frac{p(\mathcal{D} | \mathbf{r}_t) p(\mathbf{r}_t)}{p(\mathcal{D})} = \frac{p(\mathcal{D}_{t+1:T} | \mathcal{D}_{1:t}, \mathbf{r}_t) p(\mathcal{D}_{1:t} | \mathbf{r}_t) p(\mathbf{r}_t)}{p(\mathcal{D})} = \frac{p(\mathbf{r}_t, \mathcal{D}_{1:t}) p(\mathcal{D}_{t+1:T} | \mathcal{D}_{1:t}, \mathbf{r}_t)}{p(\mathcal{D})} \ ,$$

such that

$$p(r_{ti} = 1 | \mathcal{D}) = \gamma_i^{(t)} = \frac{\alpha_i^{(t)} \beta_i^{(t)}}{p(\mathcal{D})} \ , \quad \text{where} \quad \alpha_i^{(t)} = p(r_{ti} = 1, \mathcal{D}_{1:t})$$

$$\text{and} \quad \beta_i^{(t)} = p(\mathcal{D}_{t+1:T} | \mathcal{D}_{1:t}, r_{ti} = 1) \ .$$

This defines the setup for the *forward-backward algorithm* for inferring the of hidden states. These equations assume that the parameters $\mathbf{\Theta}$ are already marginalized away, coupling all the data points: given $\mathbf{R}$, the likelihood does not factorize (as is true for HMMs).

3.1.1 The forward pass

Computing the elements of vector $\boldsymbol{\alpha}^{(t)}$, i.e. $\alpha_i^{(t)}$ for $i = 1, \ldots, t$, is based on a message passing routine that uses the factorization

$$p(\mathbf{r}_t, \mathcal{D}_{1:t}) = p(\mathbf{x}_t|\mathbf{r}_t, \mathcal{D}_{1:t-1})p(\mathbf{r}_t, \mathcal{D}_{1:t-1}) = p(\mathbf{x}_t|\mathbf{r}_t, \mathcal{D}_{1:t-1})\sum_{\mathbf{r}_{t-1}} p(\mathbf{r}_t|\mathbf{r}_{t-1})p(\mathbf{r}_{t-1}, \mathcal{D}_{1:t-1}) \ .$$

Marginalization over $\mathbf{r}_{t-1}$ implies a sum over all binary settings of $\mathbf{r}_{t-1}$. Due to the choice of change point prior there is only one non-zero change point probability in this sum when $i = 1, \cdots, t-1$, and therefore

$$\begin{aligned} \alpha_i^{(t)} = p(r_{ti} = 1, \mathcal{D}_{1:t}) &= p(\mathbf{x}_t|r_{ti} = 1, \mathcal{D}_{1:t-1})\sum_{j=1}^{t-1} p(r_{ti} = 1|r_{t-1,j} = 1)p(r_{t-1,j}, \mathcal{D}_{1:t-1}) \\ &= p(\mathbf{x}_t|\boldsymbol{\eta}_i^{(t)})\,(1-H)\,\alpha_i^{(t-1)} \quad \text{for } i = 1, \ldots, t-1 \ . \end{aligned}$$

When $i = t$ a change point occurs, and there is a nonzero probability of setting $r_{tt} = 1$ from *any* previous state (recall that $\boldsymbol{\eta}_t^{(t)}$ is equal to the prior hyperparameters $\boldsymbol{\eta}$):

$$\alpha_t^{(t)} = p(\mathbf{x}_t|\boldsymbol{\eta})\sum_{j=1}^{t-1} H\,\alpha_j^{(t-1)} \ .$$

As $p(r_{ti} = 1|\mathcal{D})$ comes from renormalizing a vector, $\boldsymbol{\alpha}^{(t)}$ can be stored in a numerically stabler normalized form representing $p(r_{ti} = 1|\mathcal{D}_{1:t})$.

3.1.2 The backward pass

The elements of $\boldsymbol{\beta}^{(t)}$ can be found by a recursive formula similar to the forward pass, by writing $\beta_i^{(t)} = p(\mathcal{D}_{t+1:T}|\mathcal{D}_{1:t}, r_{ti} = 1)$ in terms of $\beta_j^{(t+1)} = p(\mathcal{D}_{t+2:T}|\mathcal{D}_{1:t+1}, r_{t+1,j} = 1)$:

$$p(\mathcal{D}_{t+1:T}|\mathcal{D}_{1:t}, \mathbf{r}_t) = \sum_{\mathbf{r}_{t+1}} p(\mathcal{D}_{t+2:T}|\mathcal{D}_{1:t+1}, \mathbf{r}_{t+1})\,p(\mathbf{x}_{t+1}|\mathcal{D}_{1:t}, \mathbf{r}_{t+1})\,p(\mathbf{r}_{t+1}|\mathbf{r}_t) \ .$$

A recursive formula for $\beta_i^{(t)}$ for $i = 1, \ldots, t$, is therefore determined by

$$\begin{aligned} \beta_i^{(t)} &= \sum_{j=1}^{t+1} p(\mathcal{D}_{t+2:T}|\mathcal{D}_{1:t+1}, r_{t+1,j} = 1)\,p(\mathbf{x}_{t+1}|\mathcal{D}_{1:t}, r_{t+1,j} = 1)\,p(r_{t+1,j} = 1|r_{ti} = 1) \\ &= \beta_i^{(t+1)}\,p(\mathbf{x}_{t+1}|\boldsymbol{\eta}_i^{(t+1)})\,(1-H) + \beta_{t+1}^{(t+1)}\,p(\mathbf{x}_{t+1}|\boldsymbol{\eta})\,H \ . \end{aligned}$$

The last line is a result of there being only two forward links from r_{ti} in the lattice (see figure 1(a)), as either $j = i$ and the run length increases and we stay in the same state, or $j = t+1$ and a new run starts and we move to a new hidden state. For $\beta_i^{(T-1)}$ to be correctly defined, we set $\beta_j^{(T)} = 1$ for all $j = 1, \cdots, T$. $\boldsymbol{\beta}^{(t)}$ can also be stored in a numerically stabler normalized form.

3.2 Pairwise marginals

The EM algorithm in section 4 makes use of the pairwise marginal probabilities $p(\mathbf{r}_{t+1}, \mathbf{r}_t|\mathcal{D})$, which will be defined by a matrix $\boldsymbol{\xi}^{(t,t+1)}$ with entries

$$\xi_{ij}^{(t,t+1)} = p(r_{ti} = 1, r_{t+1,j} = 1|\mathcal{D}) \ .$$

This is a $t \times (t+1)$ matrix, although the change point prior allows a sparse representation as it only has $2t$ non-zero entries. The pairwise marginals can be expressed as

$$p(\mathbf{r}_t, \mathbf{r}_{t+1}|\mathcal{D}) = p(\mathbf{r}_t, \mathbf{r}_{t+1}, \mathcal{D})\, p(\mathbf{x}_{t+1}|\mathbf{r}_{t+1}, \mathcal{D}_{1:t})\, p(\mathbf{r}_{t+1}|\mathbf{r}_t)\, p(\mathcal{D}_{t+2:T}|\mathbf{r}_{t+1}, \mathcal{D}_{1:t+1}) \Big/ p(\mathcal{D}) \ ,$$

from which we get

$$\xi_{ij}^{(t,t+1)} = \alpha_i^{(t)}\, p(\mathbf{x}_{t+1}|r_{t+1,j} = 1, \mathcal{D}_{1:t})\, p(r_{t+1,j} = 1|r_{ti} = 1)\, \beta_j^{(t+1)} \Big/ p(\mathcal{D}) \ .$$

A sparse representation of $\boldsymbol{\xi}^{(t,t+1)}$ arises as $\xi_{ij}^{(t,t+1)}$ is non-zero only for the following (i, j) pairs: (i, i) and $(i, t+1)$,

$$\begin{aligned}\xi_{ii}^{(t,t+1)} &\propto \alpha_i^{(t)}\, p(\mathbf{x}_{t+1}|\boldsymbol{\eta}_i^{(t+1)})\, (1 - H)\, \beta_i^{(t+1)} \\ \xi_{i,t+1}^{(t,t+1)} &\propto \alpha_i^{(t)}\, p(\mathbf{x}_{t+1}|\boldsymbol{\eta}_{t+1}^{(t+1)})\, H\, \beta_{t+1}^{(t+1)} \ .\end{aligned}$$

4. Expectation maximization for empirical Bayes

As $p(\mathbf{r}_t|\mathcal{D})$ summarizes an integral over $\boldsymbol{\Theta}$, a correctly scaled prior hyperparameter setting plays a key role in successfully detecting change points—this is illustrated in section 5.1. In this section we derive a Baum-Welch-like algorithm (Baum et al., 1970) for maximizing the *log marginal likelihood* with respect to the hyperparameters. This method can be interpreted as a volume-based version of EM (Dempster et al., 1977) where, instead of finding parameters to maximize the likelihood function, we find hyperparameters that maximize the volume under the likelihood-prior product. Unlike the EM algorithm, both the E- and M-steps require inner loops: for the E-step a variational minimization is performed, while the M-step uses a simple iterative algorithm based on the concave-convex procedure (Yuille, 2002).

The log marginal likelihood from (2) can be lower-bounded in a form amenable to EM through the introduction of an additional distribution $q(\mathbf{R})$:

$$\begin{aligned}\log p(\mathcal{D}|\boldsymbol{\eta}, H) &= \log \sum_{\mathbf{R}} q(\mathbf{R}) \frac{p(\mathcal{D}, \mathbf{R}|\boldsymbol{\eta}, H)}{q(\mathbf{R})} \\ &\geq \mathcal{F}(q; H, \boldsymbol{\eta}) = \sum_{\mathbf{R}} q(\mathbf{R}) \log \frac{p(\mathcal{D}, \mathbf{R}|\boldsymbol{\eta}, H)}{q(\mathbf{R})} \\ &= \Big\langle \log p(\mathcal{D}, \mathbf{R}|H, \boldsymbol{\eta}) \Big\rangle_{q(\mathbf{R})} + \mathcal{H}[q(\mathbf{R})] \ ,\end{aligned}$$

where $\mathcal{H}[q(\mathbf{R})]$ is the entropy of $q(\mathbf{R})$. $\mathcal{F}$ can then be maximized by an EM algorithm by iterating the following two steps:

E-step. Optimize $\mathcal{F}(q; H, \boldsymbol{\eta})$ with respect to the distribution of the hidden variables, holding the parameters fixed:

$$q^{(k)}(\mathbf{R}) \leftarrow \arg\max_{q(\mathbf{R})} \mathcal{F}(q(\mathbf{R}); H^{(k-1)}, \boldsymbol{\eta}^{(k-1)}) \ .$$

M-step. Maximize $\mathcal{F}(q; H, \boldsymbol{\eta})$ with respect to the parameters while holding the hidden distribution fixed:

$$H^{(k)}, \boldsymbol{\eta}^{(k)} \leftarrow \arg\max_{H,\boldsymbol{\eta}} \mathcal{F}(q^{(k)}(\mathbf{R}); H, \boldsymbol{\eta}) = \arg\max_{H,\boldsymbol{\eta}} \Big\langle \log p(\mathcal{D}, \mathbf{R}|H, \boldsymbol{\eta}) \Big\rangle_{q^{(k)}(\mathbf{R})} \ .$$

For tractability we choose a fully factorized form for $q(\mathbf{R}) = \prod_{t=1}^{T} q(\mathbf{r}_t) = \prod_{t=1}^{T}(\prod_{i=1}^{t} q_{ti}^{r_{ti}})$, where each $q(\mathbf{r}_t)$ is a multinomial distribution, with $\sum_{i=1}^{t} q_{ti} = 1$ and each $q_{ti} \geq 0$.

4.1 E-step

We can optimize $\mathcal{F}(q; H, \boldsymbol{\eta})$ with respect to $q(\mathbf{R})$ by sequentially optimizing each of the factors $q(\mathbf{r}_t)$ until convergence, using a variational method. We can interpret the E-step as a variational message passing scheme (Winn & Bishop, 2005) defined on figure 1(b)'s Bayesian network. The key to a variational maximization over this network is the observation that the joint probability simplifies as a function of $\mathbf{r}_t$,

$$p(\{\mathbf{r}_i\}_{i=1}^{T}|\mathcal{D}) = p(\mathbf{r}_t|\{\mathbf{r}_i\}_{i\neq t}, \mathcal{D})p(\{\mathbf{r}_i\}_{i\neq t}|\mathcal{D}) = p(\mathbf{r}_t|\mathbf{r}_{t-1}, \mathbf{r}_{t+1}, \mathcal{D})p(\{\mathbf{r}_i\}_{i\neq t}|\mathcal{D}) \ ,$$

using the Markov boundary of $\mathbf{r}_t$. The functional derivatives of $\mathcal{F}$ can then be taken with respect to *one* $q(\mathbf{r}_t)$; setting the derivative to zero (with Lagrange multipliers to enforce normalization) gives

$$q(\mathbf{r}_t) \propto \exp\left\{ \sum_{\mathbf{r}_{t+1}} \sum_{\mathbf{r}_{t-1}} q(\mathbf{r}_{t+1})\, q(\mathbf{r}_{t-1})\ \log p(\mathbf{r}_t|\mathbf{r}_{t-1}, \mathbf{r}_{t+1}, \mathcal{D}) \right\} \ .$$

This is the well-known "variational free-form maximization" method, and interested readers are referred to Jordan et al. (1999), Winn & Bishop (2005), and Bishop (2006). The log posterior $\log p(\mathbf{r}_t|\mathbf{r}_{t-1}, \mathbf{r}_{t+1}, \mathcal{D})$ can be rewritten in terms of random variable $\mathbf{r}_t$, such that the above equation equals

$$q(\mathbf{r}_t) \propto \exp\left\{ \sum_{\mathbf{r}_{t+1}} q(\mathbf{r}_{t+1}) \log p(\mathbf{r}_{t+1}|\mathbf{r}_t, \mathcal{D}) + \sum_{\mathbf{r}_{t-1}} q(\mathbf{r}_{t-1}) \log p(\mathbf{r}_{t-1}|\mathbf{r}_t, \mathcal{D}) \right\} p(\mathbf{r}_t|\mathcal{D}) \ . \tag{4}$$

The VB loop updates in (4) therefore require the pairwise and marginal probabilities, $\xi_{ij}^{(t,t+1)}$ and $\gamma_i^{(t)}$ Using the shorthand

$$\begin{aligned} \pi_{jk}^{(t,t+1)} &= p(r_{t+1,k} = 1|r_{tj} = 1, \mathcal{D}) = \xi_{jk}^{(t,t+1)}/\gamma_j^{(t)} \\ \text{and} \quad \chi_{ji}^{(t-1,t)} &= p(r_{t-1,i} = 1|r_{tj} = 1, \mathcal{D}) = \xi_{ij}^{(t-1,t)}/\gamma_j^{(t)} \end{aligned}$$

for the “forward and backward probabilities from r_{ti}”, we see that $q(r_{t,j} = 1)$ is a function of incoming links $i \to j$ and outgoing links $j \to k$ in figure 1(a),

$$q(r_{t,j} = 1) = \frac{\exp\left\{\sum_{k=1}^{t+1} q_{t+1,k} \log \pi_{jk}^{(t,t+1)} + \sum_{i=1}^{t-1} q_{t-1,i} \log \chi_{ji}^{(t-1,t)}\right\} \gamma_j^{(t)}}{\sum_{j'=1}^{t} \exp\left\{\sum_{k=1}^{t+1} q_{t+1,k} \log \pi_{j'k}^{(t,t+1)} + \sum_{i=1}^{t-1} q_{t-1,i} \log \chi_{j'i}^{(t-1,t)}\right\} \gamma_{j'}^{(t)}} \ . \tag{5}$$

The change point prior allows another simplification: as the “outgoing probability” $\pi_{jk}^{(t,t+1)}$ has only two nonzero values, at $k = j$ and $k = t + 1$ (as the run length can only increase or drop down to one), we can use

$$\sum_{k=1}^{t+1} q_{t+1,k} \log \pi_{jk}^{(t,t+1)} = q_{t+1,j} \log \pi_{jj}^{(t,t+1)} + q_{t+1,t+1} \log \pi_{j,t+1}^{(t,t+1)}$$

in (5). Similarly, when $j < t$, the “incoming probability” is non-zero only when the run length increases, and $i = j$. In that case the second sum simplifies as

$$\sum_{i=1}^{t-1} q_{t-1,i} \log \chi_{ji}^{(t-1,t)} = q_{t-1,j} \log \chi_{jj}^{(t-1,t)} \ .$$

When $j = t$ the second sum does not simplify.

A sensible initialization of $q(\mathbf{R})$ for the VB loop would be to set $q_{ti} = \gamma_i^{(t)}$. (The approach taken by Vasconcelos & Lippman (2001) in a similar setting was to guess $q_{ti} \approx \gamma_i^{(t)}$, and not implement a variational scheme at all.) When the latent state variables are *not* coupled in any way, for example in a mixture or factor analysis model, a variational loop would not be necessary.

4.2 M-step

A direct consequence of using conjugate exponential models is that a simple algorithm can be derived to maximize a *lower bound* on $\mathcal{F}$ with respect to hyperparameters $\boldsymbol{\eta}$. $\mathcal{F}$ can be directly maximized over H. From (3) the likelihood, averaged over $\mathbf{\Theta}$, is tractable,

$$p(\mathcal{D}|\mathbf{R}, \boldsymbol{\eta}) = \prod_{i=1}^{T} \exp\left\{ -\sum_{t=i}^{T} r_{ti} \log \widetilde{Z}(\mathbf{u}(\mathbf{x}_t)) - \log Z(\boldsymbol{\eta}) + \log Z\left(\sum_{t=i}^{T} r_{ti}\mathbf{u}(\mathbf{x}_t) + \boldsymbol{\eta}\right)\right\} \ . \tag{6}$$

By using Jensen’s inequality, we obtain $\langle \log p(\mathcal{D}, \mathbf{R}|H, \boldsymbol{\eta})\rangle_{q^{(k)}(\mathbf{R})} \geq \mathcal{L}(\boldsymbol{\eta}, H)$, where

$$\begin{aligned}\mathcal{L}(\boldsymbol{\eta}, H) = \sum_{t=2}^{T} &\left[q_{tt} \log H + \sum_{i=1}^{t-1} q_{ti} \log(1 - H) + \sum_{i=1}^{t-1} \left\langle q_{ti} \log q_{t-1,i}\right\rangle_{q^{(k)}(\mathbf{R})}\right] \\ &+ \sum_{i=1}^{T}\left[-\sum_{t=i}^{T} q_{ti} \log \widetilde{Z}(\mathbf{u}(\mathbf{x}_t)) - \log Z(\boldsymbol{\eta}) + \log Z\left(\sum_{t=i}^{T} q_{ti}\mathbf{u}(\mathbf{x}_t) + \boldsymbol{\eta}\right)\right] \ .\end{aligned}$$

The inequality was applied to the last $\log Z$ in (6), which is a convex function of r_{ti}.

The stationary point $\partial \mathcal{L}(H)/\partial H = 0$ is at

$$H = \frac{\sum_{t=2}^{T} q_{tt}}{\sum_{t=2}^{T} \sum_{i=1}^{t} q_{ti}} \ ,$$

which has an intuitive meaning: here H is the (geometric) prior probability of starting a new run, and it is set to the probability of remaining at run length one for each data point, normalized by the total probability. (A result of assuming that we start in state r_{11} at $t = 1$, the above sums are from $t = 2$.)

4.2.1 Concave-convex procedure

The key to solving for $\partial \mathcal{L}(\boldsymbol{\eta})/\partial \boldsymbol{\eta} = \mathbf{0}$ is to notice that the log partition function $\log Z(\boldsymbol{\eta})$ is a convex function of $\boldsymbol{\eta}$, and can be written as a function of a convex $\smile$ and concave $\frown$ term,

$$\mathcal{L}(\boldsymbol{\eta}) = -T \log Z(\boldsymbol{\eta}) + \sum_{i=1}^{T} \log Z \left(\sum_{t=i}^{T} q_{ti} \mathbf{u}(\mathbf{x}_t) + \boldsymbol{\eta} \right) = \mathcal{L}_{\frown}(\boldsymbol{\eta}) + \mathcal{L}_{\smile}(\boldsymbol{\eta}) \ .$$

We can therefore write $\partial \mathcal{L}_{\frown}(\boldsymbol{\eta})/\partial \boldsymbol{\eta} = -\partial \mathcal{L}_{\smile}(\boldsymbol{\eta})/\partial \boldsymbol{\eta}$ at a stationary point, and the concave-convex procedure (CCCP) iteratively updates $\boldsymbol{\eta}$ until guaranteed convergence with

$$\nabla \mathcal{L}_{\frown}(\boldsymbol{\eta}^{(k+1)}) \leftarrow -\nabla \mathcal{L}_{\smile}(\boldsymbol{\eta}^{(k)}) \ .$$

As $\log Z(\boldsymbol{\eta})$ can be used as generating function for the sufficient statistics of $\boldsymbol{\phi}(\boldsymbol{\theta})$, this rule updates $\boldsymbol{\eta}^{(k+1)}$ from $\boldsymbol{\eta}^{(k)}$ with

$$\big\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \big\rangle_{p(\boldsymbol{\theta} | \boldsymbol{\eta}^{(k+1)})} \leftarrow \frac{1}{T} \sum_{i=1}^{T} \big\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \big\rangle_{p(\boldsymbol{\theta} | \sum_{t=i}^{T} q_{ti} \mathbf{u}(\mathbf{x}_t) + \boldsymbol{\eta}^{(k)})} \ . \tag{7}$$

Equation (7) has a very pleasing interpretation: to maximize a log marginal likelihood, the *moments* of the prior has to match the moments of the posterior—in this case the average moments of all run length posteriors.

Example. Continuing with the Gaussian example, we need the posterior parameters $\boldsymbol{\eta}_{\mathsf{post}}^{(i)}$ of $i = 1, \ldots, T$ distributions $p(\boldsymbol{\theta} | \sum_{t=i}^{T} q_{ti} \mathbf{u}(\mathbf{x}_t) + \boldsymbol{\eta})$, each of which can be determined using

$$\begin{aligned} v_{\mathsf{post}}^{(i)} &= v + \sum_{t=i}^{T} q_{ti} & v_{\mathsf{post}}^{(i)} \mathbf{m}_{\mathsf{post}}^{(i)} &= v\mathbf{m} + \sum_{t=i}^{T} q_{ti} \mathbf{x}_t \\ \mathbf{B}_{\mathsf{post}}^{(i)} + \frac{1}{2} v_{\mathsf{post}}^{(i)} \mathbf{m}_{\mathsf{post}}^{(i)} \mathbf{m}_{\mathsf{post}}^{(i)\top} &= \mathbf{B} + \frac{1}{2} v\mathbf{m}\mathbf{m}^{\top} + \frac{1}{2} \sum_{t=i}^{T} q_{ti} \mathbf{x}_t \mathbf{x}_t^{\top} & a_{\mathsf{post}}^{(i)} - \frac{D}{2} &= a - \frac{D}{2} + \frac{1}{2} \sum_{t=i}^{T} q_{ti} \ . \end{aligned}$$

The above equations include data points $\mathbf{x}_t$ from time i (hence according to the ith diagonal of figure 1(a)), weighted by how likely it is that each $\mathbf{x}_t$ comes from a stream of data with generating parameters being reset at time i.

The expected sufficient statistics $\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle$ of $p(\boldsymbol{\theta}|\boldsymbol{\eta})$ are

$$\begin{aligned} \Big\langle \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} \Big\rangle &= \frac{1}{2}\left[\frac{D}{v} + \mathbf{m}^\top \langle \boldsymbol{\Lambda} \rangle \mathbf{m}\right] & \big\langle \boldsymbol{\Lambda} \big\rangle &= a\mathbf{B}^{-1} \\ \big\langle \log|\boldsymbol{\Lambda}| \big\rangle &= \sum_{d=1}^{D} \Psi\left(a + \frac{1-d}{2}\right) - \log|\mathbf{B}| & \big\langle \boldsymbol{\Lambda}\boldsymbol{\mu} \big\rangle &= \langle \boldsymbol{\Lambda} \rangle \mathbf{m} \ , \end{aligned} \qquad (8)$$

where $\Psi(\cdot)$ is the digamma function. Given the posterior parameters $\boldsymbol{\eta}_{\mathsf{post}}^{(i)}$, we can compute $\langle \phi(\boldsymbol{\theta}) \rangle_{\mathsf{post}}^{(i)}$ using expressions similar to (8) for $i = 1, \ldots, T$. The prior sufficient statistics in (8) should be updated to match the average of the posterior sufficient statistics $\langle \phi(\boldsymbol{\theta}) \rangle_{\mathsf{post}}^{(i)}$, and therefore we can *update* $\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle$ with

$$\begin{aligned} \Big\langle \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} \Big\rangle &= \frac{1}{T}\sum_{i=1}^{T} \Big\langle \frac{1}{2}\boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} \Big\rangle_{\mathsf{post}}^{(i)} & \big\langle \boldsymbol{\Lambda} \big\rangle &= \frac{1}{T}\sum_{i=1}^{T} \big\langle \boldsymbol{\Lambda} \big\rangle_{\mathsf{post}}^{(i)} \\ \big\langle \log|\boldsymbol{\Lambda}| \big\rangle &= \frac{1}{T}\sum_{i=1}^{T} \big\langle \log|\boldsymbol{\Lambda}| \big\rangle_{\mathsf{post}}^{(i)} & \big\langle \boldsymbol{\Lambda}\boldsymbol{\mu} \big\rangle &= \frac{1}{T}\sum_{i=1}^{T} \big\langle \boldsymbol{\Lambda}\boldsymbol{\mu} \big\rangle_{\mathsf{post}}^{(i)} \ . \end{aligned} \qquad (9)$$

The new prior hyperparameters $\mathbf{m}$, v, a, and $\mathbf{B}$ can be solved for by substituting $\langle \boldsymbol{\phi}(\boldsymbol{\theta}) \rangle$ from (9) into (8). Solving for $\mathbf{m}$ and v in (8) is straightforward with

$$\mathbf{m} = \langle \boldsymbol{\Lambda} \rangle^{-1} \langle \boldsymbol{\Lambda}\boldsymbol{\mu} \rangle \quad \text{and} \quad v = D\Big[\langle \boldsymbol{\mu}^\top \boldsymbol{\Lambda}\boldsymbol{\mu} \rangle - \mathbf{m}^\top \langle \boldsymbol{\Lambda} \rangle \mathbf{m}\Big]^{-1} \ .$$

Solving for a and $\mathbf{B}$ in (8) requires solving a nonlinear equation; we can use Newton's method to solve for a in[2]

$$\sum_{d=1}^{D} \Psi\left(a + \frac{1-d}{2}\right) - D\log a + \log\Big|\big\langle \boldsymbol{\Lambda} \big\rangle\Big| - \big\langle \log|\boldsymbol{\Lambda}| \big\rangle = 0 \ ,$$

and then substitute back with $\mathbf{B} = a\langle \boldsymbol{\Lambda} \rangle^{-1}$.

5. Results

The change point detection algorithm is illustrated on a number of examples, showing unsupervised learning of class labels, detection of change points of means and volatilities of stock returns, and a practical approach to process control. We commence with a toy example:

[2] The Newton-Raphson parameter updates can be cast into a multiplicative form, as $a > (D-1)/2$ is constrained:

$$a \leftarrow (a-k)\exp\left\{-\frac{\sum_{d=1}^{D} \Psi(a + (1-d)/2) - D\log a - c}{(a-k)\sum_{d=1}^{D} \Psi'(a + (1-d)/2) - D(a-k)/a}\right\} + k \ ,$$

where $k = (D-1)/2$, $c = \langle \log|\boldsymbol{\Lambda}| \rangle - \log|\langle \boldsymbol{\Lambda} \rangle|$, and $\Psi'(\cdot)$ is the trigramma function; see for example (Paquet, 2007).

5.1 Toy example: change in covariance, mean

We argued the importance of treating the prior hyperparameters through maximizing the log marginal likelihood, and illustrate the point here with a practical example. Figure 3(a) shows a change in the correlation structure and means of a data set, with fifty data points generated from a $\mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ density, followed by fifty more from a $\mathcal{N}(\mathbf{x}_t|\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$ density, where

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \ \boldsymbol{\mu}_2 = \begin{bmatrix} \frac{1}{2} \\ -\frac{1}{2} \end{bmatrix}, \ \boldsymbol{\Sigma}_1 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \text{ and } \boldsymbol{\Sigma}_2 = \begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}.$$

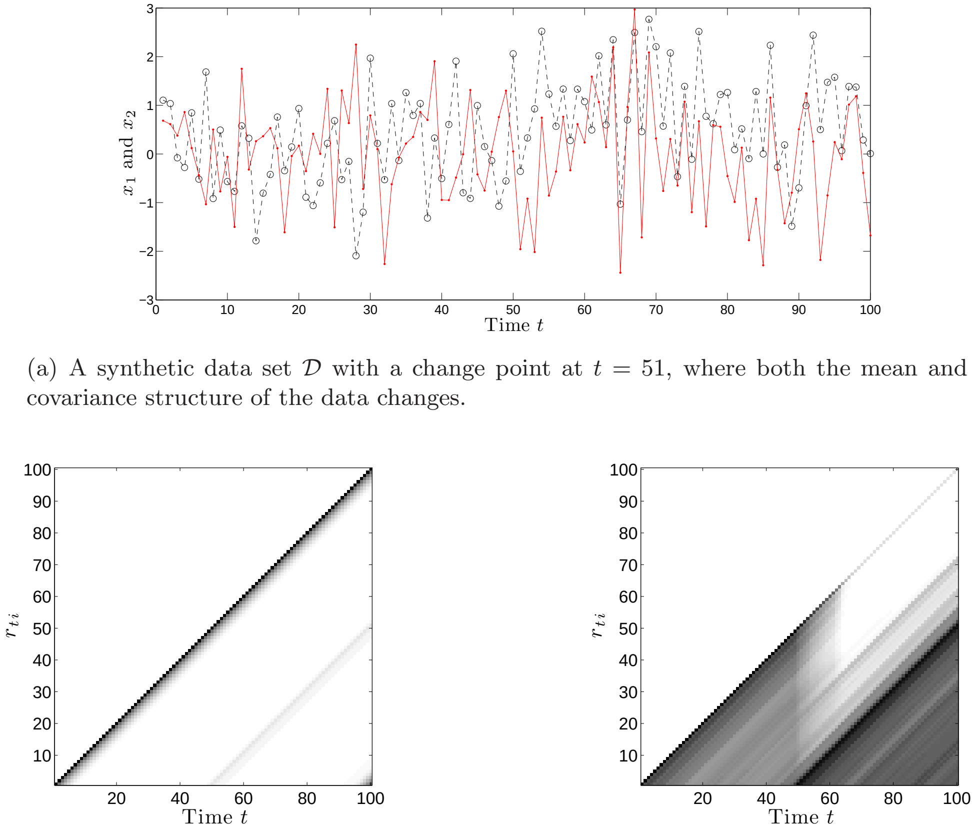
With initial hyperparameters, $\mathbf{m} = \mathbf{0}$, $v = 0.25$, $a = (D+1)/2$, $\mathbf{B} = [16,\ 1;\ 1,\ 16]$, and $H = 0.1$, we arrive at figure 3(b). The problem is that seeing a data point changes the posterior for the initial run enough, so that its predictive density always dominates over predictive densities based on a misscaled prior. Figure 3(c) shows $p(\mathbf{r}_t|\mathcal{D})$ *after* treating the hyperparameters with an EM algorithm, which gave $\mathbf{m} \approx [0.6,\ -0.1]$, $v \approx 13$, $a \approx 3$, $\mathbf{B} \approx [2.6,\ 1.2;\ 1.2,\ 3.3]$, and $H \approx 0.006$. The advantage of the EM hyperparameter scheme is clear, as the change point is correctly detected.

5.2 Iris data set

As a second example, we removed the class labels from the well known iris data set (Fisher, 1936), and show an *unsupervised* approach to inferring the number of classes, and the class labels. Our only assumption is that classes will appear in sequence over time. This assumption is relevant to process control (see section 5.4), where one may assume that one class models “in-control” quality control variables, which, after some failure, will be followed by a change in class and distribution. The iris data set contains 150 sets of measurements[3] coming from three types of flowers. There are 50 instances of each flower. Figure 4(a) shows the data projected onto its two principal components. We removed the class labels, but colored the data points in the order of observing them.

The “sequential” assumption allows us to correctly infer that there are three classes, as illustrated in figure 4(b). Using the maximum i in $p(r_{ti} = 1|\mathcal{D})$ for each example $\mathbf{x}_t$, the class labels can be correctly attached. This is true even for the slightly overlapping classes (iris-versicolor and iris-virginica), which separate at $t = 101$.

It is interesting to note that an unoptimized hyperparameter setting only gives two classes, in a fashion similar to figure 3(b). This relates to the following test: If we did not make the “sequential” assumption, and assumed the data to come from a K-component Gaussian mixture model, then a Variational Bayes (VB) lower bound to the log marginal likelihood (Attias, 2000), using a broad zero-mean prior (with $v = 10^{-6}$, for example) gives a $K = 2$ model to be around 76 times more likely than a $K = 3$ model. Running VB using (K times replicated) hyperparameters optimized with the change point model—i.e. those used in figure 4(b)—reverses the result, and presents a $K = 3$ model to be 41 times more likely than a $K = 2$ model.



(a) A synthetic data set $\mathcal{D}$ with a change point at $t = 51$, where both the mean and covariance structure of the data changes.

(b) Change point probabilities $p(\mathbf{r}_t|\mathcal{D})$ before the start of the EM loop, with prior hyperparameters roughly chosen on the correct scale; here $v = 0.25$, for example.

(c) After maximizing a bound on the log marginal likelihood; now $v \approx 13$.

Figure 3: An Illustration the importance of hyperparameter choice, given the synthetic data set of figure 3(a). A log-scale plot of $p(\mathbf{r}_t|\mathcal{D})$ is shown for an unoptimized and optimized choice of hyperparameters. Darker pixels correspond to higher probabilities. This is a probabilistic version of figure 1(a), with the numbering on the vertical axis indicating *run length*.

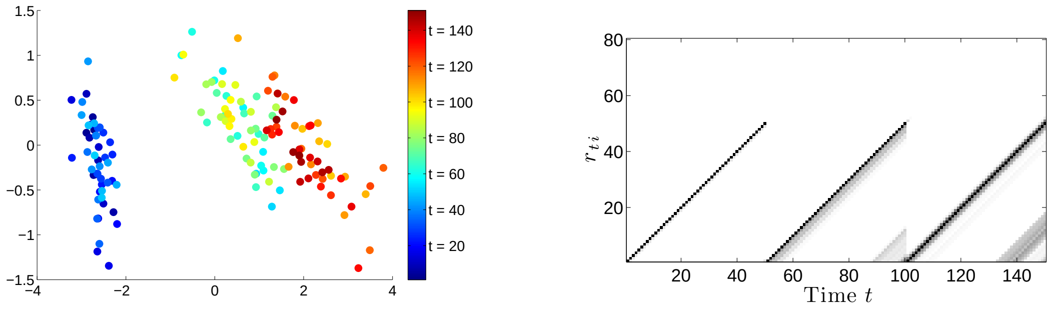
5.3 South African economic data

A plausible application of this change point model is in the detection of changes in the returns (expected means) and volatilities (variances) of stock market indices (Adams & MacKay, 2006; Loschi & Cruz, 2002). In figure 5 the weekly log-returns

$$x_t = \log(I_t/I_{t-1})$$

of the Johannesburg Stock Exchange all-share index (JSE-ALSI) were used, where I_t is the index at the end of week t. We used indices from August 1978 to July 2007, and from the

[3] Each measurement contains the sepal length, sepal width, petal length, and petal width of a flower.



(a) The iris data set (which was centered to obtain the figure) projected onto its two principal components. The color bar indicates the ordering of data $\mathbf{x}_t$.

(b) A log-scale plot of $p(\mathbf{r}_t|\mathcal{D})$, strongly indicating the presence of three classes, with crisp separation even for slightly overlapping classes.

Figure 4: The unlabeled iris data set, and resulting change points.

figure highlight a number of change points that could possibly have affected South Africa's economy. *October 1987:* Black Monday, the second largest one-day percentage decline in stock market history; *April 1994:* South Africa's first fully democratic elections, with Nelson Mandela coming to power; *July 1997:* start of the Asian financial crisis; *January 1999:* run-up to South Africa's second fully democratic elections, ending Mandela's term in office, and the Euro is introduced in non-physical form; *September 2001:* World Trade Center attacks in New York.

5.4 Gravel Process

A typical task in quality control of a process is the detection of changes in the "in-control" statistical distribution of measurements, as these changes are often caused by specific changes in the process itself. Control charts are usually used to estimate the time of change or the presence of multiple changes in multivariate control measurements.

In this example we take 56 observations from a European grit- or gravel-producing plant, first presented in the control literature by Holmes & Mergen (1993). Each data point contains the per cent of particles (by weight) that are of large and medium size, and is shown in figure 6(a). Under the assumption that "the nature of the process may suggest rational subgroups within which the quality measurements are relatively homogenous", Sullivan & Woodall (2000) used the data to construct control charts to detect changes in the mean vector and covariance matrix responsible for generating the data. The problem lends itself perfectly to this paper's empirical Bayesian change point algorithm, which comes with a number of bonuses: (a) the detection of multiple shifts (change points) don't need recursive splits of the data around most likely change points; (b) no control statistic has to be constructed, as full use is made of probability theory; (c) a probability distribution over change points is given; (d) hyperparameters and model assumptions are explicitly stated and treated in a principled way.

Figure 6(b) illustrates the change point posterior for this data set. Sullivan & Woodall (2000) gave shifts after $t = 24$ and $t = 43$, which can be observed in the figure.

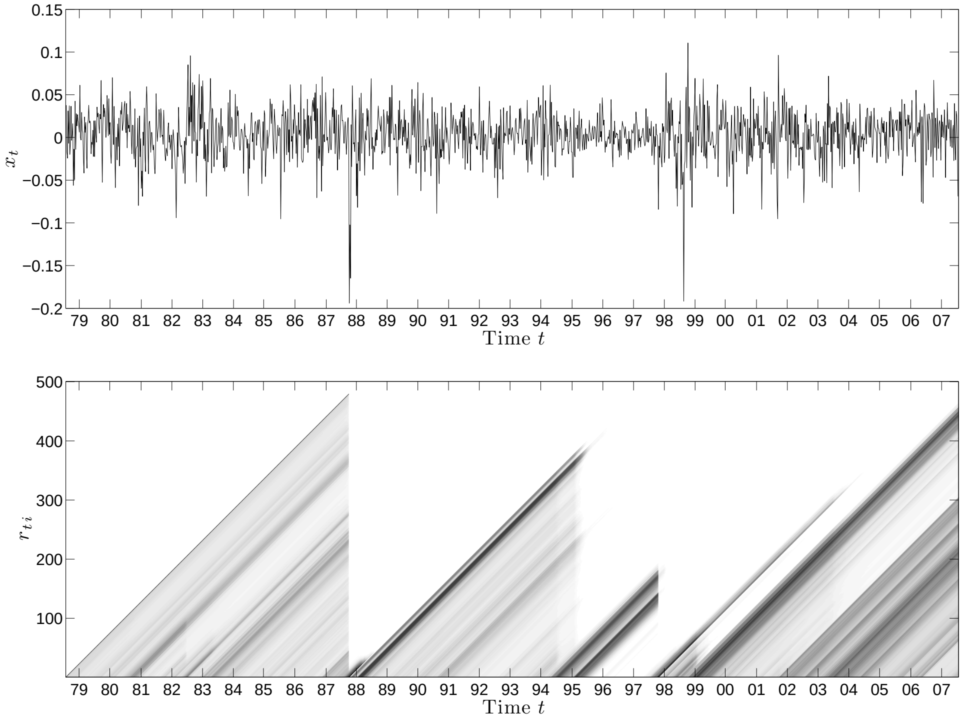


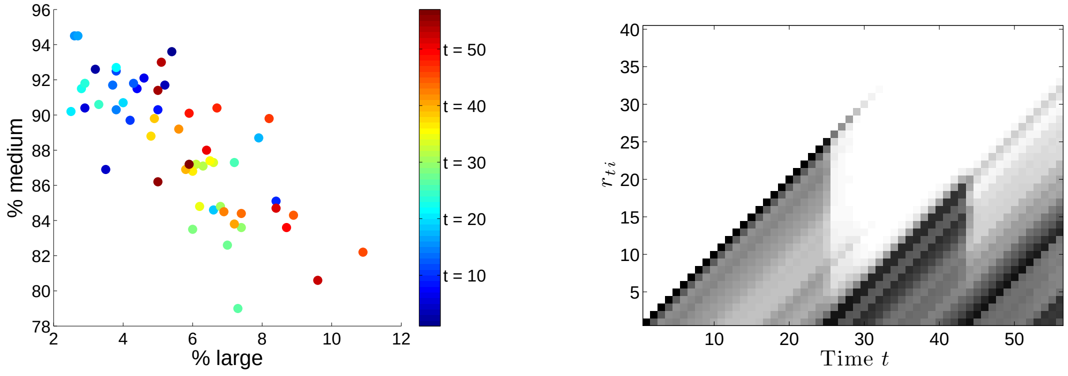
Figure 5: *Top:* Weekly returns, or log-differences x_t, of the Johannesburg Stock Exchange all-share index (JSE-ALSI), from August 1978 to July 2007. *Bottom:* Change point probabilities $p(\mathbf{r}_t|\mathcal{D})$ using the JSE-ALSI. The plot is shown on a log-scale, with darker pixels indicating higher probability; time t is indicated in years.

6. Conclusion

The product partition model is a useful tool in detection change points, but may not rise to its full potential due to misscaled choices of prior hyperparameter settings. This paper addressed the problem through an efficient EM algorithm to treat prior hyperparameters, opening the door to other interesting applications in models with latent variables:

Hyperparameter optimization is often associated with *automatic relevance determination* (ARD). In the context of mixture modeling, ARD is useful in determining “feature saliences”, i.e. how much a certain input dimension plays a role in clustering the data. This paper’s EM algorithm can be used in a similar way to infer the number of components in a mixture, without using fixed hyperparameters and a variational lower bound like Corduneanu & Bishop (2001). Rather, given a choice of K and a Dirichet mixing-weight prior, the hyperparameters can be optimized, and a scheme can be devised where either suppressed components (with small posterior mixing weights, etc.) are pruned, or components added sequentially until the bound decreases.

Latent variable problems like the one discussed above typically have multimodal posteriors due to permutation symmetries in the latent variable structure. Hyperparameter



(a) The gravel data set, colored according to the ordering of $\mathbf{x}_t$.

(b) A log-scale plot of $p(\mathbf{r}_t|\mathcal{D})$.

Figure 6: The gravel data set, and resulting change points.

optimization will force the prior to strengthen one such mode and give little mass to the others. In light of modes being permutation-equivalent, this is a promising approach to ARD in many models.

Acknowledgments

Thank you to Blaise Thomson, Amy Morris and Jacques Plaut for helpful discussions on the JSE data.

References

Adams, R. P. & MacKay, D. J. C. (2006). Bayesian online changepoint detection. Technical report, Cavendish Laboratory, University of Cambridge.

Attias, H. (2000). A variational Bayesian framework for graphical models. In Solla, S. A., Leen, T. K., & Müller, K.-R. (Eds.), *Advances in Neural Information Processing Systems 12*, (pp. 209–215). MIT Press.

Barry, D. & Hartigan, J. A. (1992). Product partition models for change point problems. *The Annals of Statistics*, *20*(1), 260–279.

Barry, D. & Hartigan, J. A. (1993). A Bayesian analysis for change point problems. *Journal of the American Statistical Association*, *88*(421), 309–319.

Baum, L. E. & Egon, J. A. (1967). An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bulletin of the American Mathematical Society*, *73*, 360–363.

Baum, L. E., Petrie, T., Soules, G., & Weiss, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *The Annals of Mathematical Statistics*, *41*, 164–171.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Carlin, B. P. & Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis* (Second ed.). Chapman Hall.

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

Corduneanu, A. & Bishop, C. M. (2001). Variational Bayesian model selection for mixture distributions. In Richardson, T. & Jaakkola, T. (Eds.), *Proceedings Eighth International Conference on Artificial Intelligence and Statistics*, (pp. 27–34). Morgan Kaufmann.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, *39*(1), 1–38.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*, 179–188.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, *82*, 711–732.

Holmes, D. S. & Mergen, A. E. (1993). Improving the performance of the T^2 control chart. *Quality Engineering*, *5*(4), 619–625.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., & Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, *37*(2), 183–233.

Lai, T. L. (1995). Sequential changepoint detection in quality control and dynamical systems. *Journal of the Royal Statistical Society B*, *57*(4), 613–658.

Lauritzen, S. L. & Spiegelhalter, D. J. (1998). Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society B*, *50*(2), 157–224.

Loschi, R. H. & Cruz, F. R. B. (2002). Applying the product partition model to the identification of multiple change points. *Advances in Complex Systems*, *5*(4), 371–387.

MacKay, D. J. C. (1992). The evidence framework applied to classification networks. *Neural Computation*, *4*(5), 698–714.

Paquet, U. (2007). *Bayesian inference for latent variable models*. PhD thesis, University of Cambridge.

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(22), 257–286.

Robert, C. P. & Casella, G. (2004). *Monte Carlo Statistical Methods* (Second ed.). Springer.

Sullivan, J. H. & Woodall, W. H. (2000). Change-point detection of mean vector or covariance matrix shifts using multivariate individual observations. *IIE Transactions*, *32*(6), 537–549.

Vasconcelos, N. & Lippman, A. (2001). Empirical Bayesian motion segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 217–221.

Winn, J. & Bishop, C. M. (2005). Variational message passing. *Journal of Machine Learning Research*, *6*, 661–694.

Yuille, A. L. (2002). CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, *14*, 1691–1722.