Bayesian Segmentation and Rhythm Tracking
Part II: Segmentation With Priors

Harvey Thornburg

Nov. 9 2001
Forms of the Prior

- We begin with a fundamental question: how to represent a distribution over segment times. In general we cannot assume a fixed number of segments, so we specify a distribution over the indicator sequence $d_{1:N}$, where

$$d_t = \begin{cases} 
1, & \text{change at time } t \\
0, & \text{otherwise.}
\end{cases}$$

- $P(d_{1:N})$ may be decomposed over a mixture of distributions concerning a fixed number of changes:

$$P(d_{1:N}) = \sum_{m=0}^{N} P(d_{1:N}, M = m)$$

$$= \sum_{m=0}^{N} P(M = m)P(d_{1:N} | M = m)$$

$$= \sum_{m=0}^{N} P(M = m)P_m(d_{1:N})$$

$$= \sum_{m=0}^{N} P(M = m)P_m(T_{1:m})$$

(1)

Here $M = m$ represents the number of changes and $P_m(d_{1:N})$ is concentrated on the set of indicator sequences having $m$ 1's. The list of change times $T_{1:m}$ becomes equivalent to enumerating the places where $d_t = 1$. 
• In practice we operate using local windows, and the problem may be specified so we can assume $P(M \neq 1) < \epsilon$. The implications are as follows:

• Consider two distributions, $P_1(d_{1:N})$ and $P_2(d_{1:N})$ which satisfy $P(M \neq 1) < \epsilon$ and their marginals agree: $P_2(d_t) = P_1(d_t)$. It’s easy to show the maximum difference between the masses assigned to each singleton is no more than $2\epsilon$, i.e.

$$\sup_{d_{1:N} \{P(d_{1:N}): P(M \neq 1) < \epsilon\}} |P_1(d_{1:N}) - P_2(d_{2:N})| < 2\epsilon \quad (2)$$

So if exactly one change occurs with probability $1 - \epsilon$, we don’t really care about the exact form of the prior as long as the marginals agree. For instance, we can exchange a mutually exclusive prior for an independent prior. Often we’re forced to make additional assumptions because we can’t access the entire joint distribution, we know only some partial information about the distribution of change points.

Partial Information

• The most popular queries are the marginal $P(d_t)$ and the hazard rate $P(d_t | d_{1:t-1} = 0_{1:t-1})$. The marginal is useful as prior for offline detection, where we test whether a change occurs at a specific time in the signal.

• As we see, the hazard rate is accessed in the online (sequential) detection.
• As an example, we compute both quantities for the exclusive prior for exactly one change. The probability that the change time is $t$ is $P_1(T_1 = t) = p_t$.

  - **Marginal**: $P(d_t = 1) = P_1(T_1 = t) = p_t$.
  
  - **Hazard Rate**:
    
    $$P(d_t = 1|d_{1:t-1} = 0_{1:t-1}) = P_1(T_1 = t|T_1 > t - 1) \ (3)$$

    According to Bayes’ rule:

    $$P_1(T_1 = t|T_1 > t - 1) = \frac{P_1(T_1 = t)P(T_1 > t - 1|T_1 = t)}{P_1(T_1 > t - 1)} \ (4)$$

    $$= \frac{p_t}{1 - \sum_{s=0}^{t-1} p_s} \ (5)$$
Online Methods

• First, we describe the general problem setting for the online segmentation of an audio signal.

  – Signal Model Assume stationary Gaussian AR model for each segment. The segment time(s) are unknown. We wish to find them as early as possible.

  – Unknown Model Parameters As well as the change times, the signal model parameters before and after change are unknown. These types of unknown parameters are called nuisance parameters in the statistical literature, because it is not the main purpose to identify them.

  – To make things easier, we pursue methods where the model before change is known, but after change is unknown. The rationale is, if change times are usually far apart, we expect there to be enough samples to identify the model before change. However, we pay for samples used to identify the model after change in terms of detection delay.

• In the statistical literature, there are two standard approaches for dealing with such nuisance parameters in hypotheses tests.

  – Marginalized Approach When there exists a prior distribution (say \( p(\theta_1) \)) concerning the model parameter after change, we may recover a tangible, known model after change by integrating out \( \theta_1 \) w.r.t. \( p(\theta_1) \). In other words, let \( P(y|\theta_1) \) give the conditional law after change; the unconditional law \( P_1(y) \) is obtained:

\[
P_1(y) = \int_{\Theta} P(y|\theta_1) d\theta_1
\]
Now the model after change, $P_1(y)$, is completely specified. There are no longer any nuisance parameters.

- **Generalized Approach** We substitute the maximum likelihood estimate of $\theta_1$ for $\theta_1$. The generalized approach is useful when we either don’t have or don’t want to assume a prior. However, we can mimic this behavior in the marginalized approach by using a noninformative, or maximum-entropy prior.

For the online segmentation with Gaussian AR models, either approach has problems.

- The marginalized approach is generally intractable; the integral cannot be performed without resorting to Monte Carlo methods.
- With the generalized approach, we incur detection delay while estimating $\theta_1$.

**Multiple hypothesis codebooks**

- The proposed solution, inspired by a very different application in (Nikiforov 2001) is to sample the alternative model space for $\theta_1$ according to shells of increasing “detectability”, measured by the Kullback-Leibler distance $K(\theta_1, \theta_0)$.

- As a detectability measure, K-L distance is justified by Lorden’s lower bound.

$$E^*(T) > \frac{\log \alpha}{K(\theta_1, \theta_0)}$$ (6)
Here $\alpha$ is the mean false alarm time and $E^*(T)$ is the essential supremum of the detection delay, over all values for the change time and trajectories of the signal before change.

- **Practical justification:** There is usually some minimum detectability (beyond which the effect of the change cannot be heard) and some maximum (beyond which we don’t care if the algorithm is optimal, or in an extreme case where the detection delay $< 1$ sample.)

- The general idea is to choose a finite number of shells, ranging from minimum to maximum detectability. We can’t choose an infinity of models on the surface of each shell, so we approximate by sampling along principal axes, which represent the principal directions of change.

- **Principal axis sampling** is convenient for iid Gaussian models (constant covariance matrix, piecewise-constant mean) because the surfaces of constant K-L distance are nothing but scaled versions of the covariance ellipsoids.

- **Principal deviation sampling.** Principal axis sampling does not work out so well for the AR models, however. In fact for $p \geq 3$, it can be shown the surfaces of constant $K(\theta_1, \theta_0)$ are not guaranteed to be convex in $\theta_1$. We form the codebook of alternative models, as follows:

  - For each complex pole pair, four models corresponding to perturbations in the directions of increased and decreased frequency; increased and decreased damping.
  - For each real pole, two models corresponding to positive and negative perturbations of this pole.
– For the innovations variance, two models corresponding to increased and decreased variance.

The size of each perturbation is chosen to achieve the shell’s associated K-L distance $K(\theta_1, \theta_0)$.

• For AR(2) model, we show surfaces of constant K-L distance. As expected, the distance tends to infinity on the contours of the stability triangle.

Figure 1: Level surfaces of constant Kullback’s information
Empirical investigation of the multimodel sampling

• Before discussing the multihypothesis sequential test, we present some empirical results for different sampling strategies. We do this for the Gaussian iid case, where the marginalized approach is tractable. Vs. the marginalized approach, we compare:
  – Principal axis sampling, one shell
  – 2x oversampling, one shell
  – Principal axis sampling, two shells

• Only marginal improvements are realized by increasing the number of shells or the number of points per shell. All results compare favorably with the marginalized approach. Instead of the essential supremum of detection delay, we show the percentage of delay which exceeds an acceptable threshold (30 samples) vs. false alarm rate.

• Similar results were found concerning the introduction of multiple shells in the Gaussian AR case. Unfortunately, we are not able to compare with the marginalized case (thanks to intractability.)
Figure 2: Comparative performance for multimodel sampling strategies (one shell; principal axis sampling vs. oversampling) vs. marginalized approach
Figure 3: Comparative performance for multimodel sampling strategies (one and two shells; principal axis sampling) vs. marginalized approach
Binary hypothesis sequential test

- The optimal Bayes sequential test for online change detection is based in the theory of *optimal stopping* (Shiryayev 1973, Speyer 1984). To give a flavor of this theory without adding too much complexity, we show the development for the binary hypothesis case. The multihypothesis case (Malladi and Speyer 1999) is handled in the writeup.

- The general idea is that we have a prior distribution over the time of transition from hypothesis $\theta_0$ to hypothesis $\theta_1$. Both $\theta_0$ and $\theta_1$ are known, which means the conditional distributions before and after change are known: $P(y_t|\theta_0, y_{1:t-1})$ and $P(y_t|\theta_1, y_{1:t-1})$.

- We process the signal until some time $T$, at which point we decide that a change has occurred. Since $T$ is causally determined as a function of past and present observations, it fits the probabilistic requirements of a *stopping time*. There are two situations we wish to avoid:
  - *False Alarm*: At the stopping time of the test, the true signal model is $\theta_0$.
  - *Detection Delay*: The change to $\theta_1$ occurred in the past, at time $T - \delta$, incurring delay of $\delta$.

We assign specific costs to each situation. The total cost is the sum of false alarm cost and detection delay times per-sample detection delay cost. Since scaling doesn’t matter, we might as well assign a detection delay cost of 1, and fix the false alarm cost relative to the delay cost.
• The goal of the Bayesian sequential test is to minimize expected total cost. We may now compute this expected cost because the distribution of the change time is specified.

• Recall that in the non-Bayesian case, we had to find the optimal tradeoff between the rate of false alarms and the detection delay. Since the optimal test is parameterized in terms of one or more thresholds, we would play with the thresholds to navigate this tradeoff. Bayesian methods always have this advantage in theory, that we can specify points on the tradeoff curve which correspond to physically meaningful values. We don’t have to constantly vary the threshold and observe the effect.

• Prior information about the change time enters in the form of a hazard rate: At time $t$, we access $P(T_1 = t|T_1 > t - 1)$ as the prior probability of transition from hypothesis $\theta_0$ to hypothesis $\theta_1$.

• An important quantity to propagate in the actual test is $P(T_1 = t|T_1 > t - 1, y_{1:t})$ the posterior probability of transition to $\theta_1$ at time $t$, given signal observations $y_{1:t}$. Before discussing the objectives of the test, we begin by developing recursions for these probabilities.

Recursions for the Posterior Probabilities of Transition

• Define:
  
  $\rho_t = P(T_1 = t|T_1 > t - 1)$, the hazard rate
  
  $F_{t,i} = P(T_i \leq t|y_{1:t})$, the filtered posterior probability that a change has not occurred before time $t$
\(- G_{t,i} = P(T_i \leq t + 1|y_{1:t})\), the predicted posterior probability

\(- f_{\theta(t)}(y_t|y_{1:t−1})\) be the conditional distribution of \(y_t\) given \(y_{1:t−1}\)

and some \(\theta \in \{\theta_0, \theta_1\}\).

**Lemma Time Update**

\[
G_{t,1} = F_{t,1} + \rho_t(1 - F_{t,1}) \\
G_{t,0} = 1 - G_{t,1}
\]

**Proof**

\[
G_{t,1} = P(T_1 \leq t + 1|y_{1:t}) \\
= P(T_1 \leq t|y_{1:t}) + P(T_1 = t + 1|y_{1:t}) \\
= F_{t,1} + P(T_1 = t + 1, T_1 > t|y_{1:t}) \\
= F_{t,1} + P(T_1 = t + 1|T_1 > t, y_{1:t})P(T_1 > t|y_{1:t}) \\
= F_{t,1} + P(T_i = t + 1|T_i > t)(1 - F_{t,1}) \\
= F_{t,1} + \rho_t(1 - F_{t,1})
\]

The conclusion \(P(T_1 = t + 1|T_1 > t, y_{1:t}) = P(T_1 = t + 1|T_1 > t)\)
is justified because \(\{T_1 = t + 1\} \equiv \{\theta(t + 1) = \theta_1\} \cup \{\theta(1 : t) = \theta_0\}\)and \(\theta(t + 1)\) and \(y_{1:t}\) are conditionally independent given \(\theta(1 : t)\).

**Lemma Measurement Update**

\[
F_{t+1,i} = \frac{\int_{\theta} f_{\theta_i}(y_{t+1}|y_{1:t})G_{t,i}}{\sum_{i=0}^{M} \int_{\theta} f_{\theta_i}(y_{t+1}|y_{1:t})G_{t,i}}, \forall i = 0, 1
\]
Proof

\[ F_{t+1,i} = \frac{P(T_i \leq t + 1|y_{1:t+1})}{P(T_i \leq t + 1, y_{t+1}|y_{1:t})} \]
\[ = \frac{P(T_i \leq t + 1, y_{t+1}|y_{1:t})}{P(y_{t+1}|y_{1:t})} \]
\[ = \frac{P(y_{t+1}|y_{1:t}, T_i \leq t + 1)P(T_i \leq t + 1|y_{1:t})}{P(y_{t+1}|y_{1:t})} \]
\[ = \frac{f_{\theta_i}(y_{t+1}|y_{1:t})G_{t,i}}{\sum_{i=0}^{M} f_{\theta_i}(y_{t+1}|y_{1:t})G_{t,i}} \]

• The process is initialized as follows: \( F_{0,i} \) is nothing but the prior probability for each hypothesis. Because with probability 1, we begin with \( \theta = \theta_0 \), it follows:

\[ F_{0,i} = 0, i = 1 \ldots M \]
\[ F_{0,0} = 1 \]

• Finally, the stopping rule may be expressed:

\[ T = \inf\{t \geq 1 : F_{t,1} \geq \gamma\} \]

where \( \gamma \) is a user-specified threshold, induced by the false alarm and detection delay penalties. Obviously if we lower the relative false alarm penalty, we end up with a less sensitive test (\( \gamma \) decreases).

Results

• Before verifying the stopping rule, we will show some results: Sample trials are illustrated below for cases of both correct and incorrect prior information. We adopt the exclusive prior: \( P(d_{1:N}) = \)
$P_1(T_1)$; i.e. exactly one transition occurs. The distribution of the transition time $T_1$ is that of a sampled Gaussian; in other words, let $\tau$ be a real-valued random variable with distribution $\mathcal{N}(E\tau, Var\tau)$; fix $P_1(T_1 = t) = P(\tau \in [t - 1/2, t + 1/2])$.

- Figures (4,5, 6) show example results for AR(2) process with correct prior. Figures (7,8, 9) give parallel results for incorrect prior.

$$\begin{align*}
\text{Figure 4: Correct Gaussian prior, } E\tau &= 1200, [Var\tau]^{1/2} = 10,
\end{align*}$$

- Data is generated for these trials, as follows. The true change time always occurs at $t = 1200$. Before change the white noise model ($A_0 = 0, \lambda_0 = 1$) holds; after change, an alternative model is sampled randomly and uniformly from the surface of the unit shell: $\{\theta_1 : K(\theta_1, \{A_0, \lambda_0\}) = 1\}$.

- The following parameters for priors are used: in the trials “correct prior”, $E\tau = 1200$, and the standard deviation $[Var\tau]^{1/2}$ varies among $\{10, 60, 360\}$, ranging from almost perfect information to a reasonably flat prior. In the trials “incorrect prior”, $E\tau = 1800$.
with the same variances. In every case, the hazard rate for every alternative hypothesis equals the overall hazard rate. Since changes are mutually exclusive we effectively assume all the prior transition probability is concentrated in the most likely transition. This follows most closely the philosophy of the generalized approach, as we will see.

- To generate alternative models in the multihypothesis sequential test, principal deviation sampling is used, with a single shell of Kullback’s information $= 2$.

**Relation to SPRT**

- Let’s consider how we can use this test for not just the change detection, but for the general sequential hypothesis test problem.
In this case the signal follows initially either $\theta_0$ or $\theta_1$, and never undergoes change. We set the prior probability for the $\theta_1$ hypothesis as: $F_{0,1} = P(\theta_0)$, and the transition probability as $\rho_t = 0 \forall t$.

- $\rho_t = 0$ implies the following simplification for the time update equation: $G_{t,1} = F_{t,1}$

- The measurement update simplifies accordingly:

$$F_{t+1,i} = \frac{f_{\theta_i}(y_{t+1}|y_{1:t})F_{t,i}}{\sum_{i=0}^{M} f_{\theta_i}(y_{t+1}|y_{1:t})F_{t,i}}, \forall i = 0, 1$$

- Now, define:

$$S_t = \log \frac{F_{t,1}}{F_{t,0}}$$

$$s_t = \log \frac{f_{\theta_1}(y_t|y_{1:t-1})}{f_{\theta_0}(y_t|y_{1:t-1})}$$
Substituting obtains the recursion:

\[ S_{t+1} = S_t + s_{t+1} \]

If the recursion begins with a *uniform* prior on the initial change hypothesis, i.e. \( F_{0,1} = 1/2 \), then we obtain \( S_0 = 0 \). Hence \( S_t \) is nothing but the log likelihood ratio:

\[ S_t = \log \frac{f_{\theta_1}(y_{1:t})}{f_{\theta_0}(y_{1:t})} \]

The optimal stopping rule specializes to:

\[ T = \inf \{ t \geq 1 : S_t \geq \gamma \} \]

which is nothing but the SPRT. Changing the prior \( F_{0,1} \) is equivalent to adding a constant offset to the SPRT statistic, or modifying the threshold.
Figure 8: Correct Gaussian prior, $E\tau = 1200$, $[Var\tau]^{1/2} = 60$

**Optimal stopping rules**

- We sketch the justification of

$$T = \inf\{t \geq 1 : F_{t,1} \geq \gamma\}$$

as an optimal stopping rule, omitting some of the more technical mathematical details which can be accepted on intuition.

- The situation at time $t$ is thus: We have observed $y_{1:t}$. We decide either to
  - Stop testing, declare a change has occurred at time $t$.
  - Take one more observation

- The generic optimal stopping rule (Shiryayev 1973) is to keep testing until the additional cost due to stopping is less than the
optimal additional cost due to continuation (take one more observation).

- The optimal additional cost at time $t$ is just the minimum of these two costs. We call this cost $J(F_{t,1})$ because it is a function only of the posterior probability of change at time $t$, as we will see.

- The additional cost due to stopping is the expected cost due to false alarm, which is $(1 - F_{t,1}) c_{FA}$.

- The additional cost due to continuation is 1, plus the expected optimal additional cost at time $t + 1$, which is $E_{F_{t,1}}(J(F_{t+1,1}))$.

Thus we have the recursion:

$$J(F_{t,1}) = \min\{(1 - F_{t,1}) c_{FA}, 1 + E_{F_{t,1}}(J(F_{t+1,1}))\}$$

- We stop at the first $t$ where $(1 - F_{t,1}) c_{FA} \leq 1 + E_{F_{t,1}}[J(F_{t+1,1})]$. 

Figure 9: Correct Gaussian prior, $E\tau = 1200, [Var\tau]^{1/2} = 360$
• Lemma $J(F_{t,1})$ is concave in $t$. Key insight in proof: Suppose $J(F_{t+1,1})$ is concave in $F_{t+1,1}$. The expectation preserves concavity, because it is a positive linear functional. Then $J(F_{t,1})$ is the pointwise minimum of a linear function and a concave function, which is also concave. The only technical problem here is the basis case (see Shirayev for details).

• Because the test depends only on $F_{t,1}$, given $F_{t,1}$ it is the same for every $t$. The concavity tells us the form of the test, which compares a linear function to a concave function. The region where the linear function, $(1 - F_{t,1})c_{FA}$, is less than the concave function, $\leq 1 + E_{F_{t,1}}[J(F_{t+1,1})]$, is generally an open interval $\{F_{\text{min},1}, F_{\text{max},1}\}$ It can be shown that $F_{\text{max},1} = \infty$, so the test must be a one-sided threshold.