Bayesian hidden Markov Model for DNA segmentation : A prior sensitivity analysis

Darfiana Nur, David Allingham, Judith Rousseau, Kerrie Mengersen, Ross Mcvinish

Bayesian Hidden Markov model for DNA Sequence Segmentation : A prior sensitivity analysis

Darfiana Nur [a,1], David Allingham [b]
Judith Rousseau [c], Kerrie L. Mengersen [d] and Ross McVinish [d]

[a] *School of Mathematical and Physical Sciences, University of Newcastle, Callaghan, NSW 2308, Australia,*
[b] *ARC Centre of Excellence for Complex Dynamic Systems and Control, University of Newcastle, Callaghan, NSW 2308, Australia,*
[c] *CEREMADE, Université Paris-Dauphine, Place du maréchal deLattre de Tassigny, Paris 75016, France,*
[d] *School of Mathematical Sciences, Queensland University of Technology, Brisbane, QLD 4001, Australia*

Abstract

The focus of this paper is on the sensitivity to the specification of the prior in a hidden Markov model describing homogeneous segments of DNA sequences. An intron from the chimpanzee α-fetoprotein gene, which plays an important role in embryonic development in mammals is analysed. Three main aims are considered : (i) to assess the sensitivity to prior specification in Bayesian hidden Markov models for DNA sequence segmentation; (ii) to examine the impact of replacing the standard Dirichlet prior with a mixture Dirichlet prior; and (iii) to propose and illustrate a more comprehensive approach to sensitivity analysis, using importance sampling. It is obtained that (i) the posterior estimates obtained under a Bayesian hidden Markov model are indeed sensitive to the specification of the prior distributions; (ii) compared with the standard Dirichlet prior, the mixture Dirichlet prior is more flexible, less sensitive to the choice of hyperparameters and less constraining in the analysis, thus improving posterior estimates; and (iii) importance sampling was computationally feasible, fast and effective in allowing a richer sensitivity analysis.

Keywords: DNA sequence; hidden Markov model; Bayesian model; sensitivity analysis; α-fetoprotein; Markov chain Monte Carlo; importance sampling.

1 Introduction

Many genome sequences display heterogeneity in base composition in the form of segments of similar structure. A number of statistical techniques have been developed to identify these homogeneous DNA segments, as reviewed in Braun and Müller (1998). One technique, proposed in Churchill (1989), describes DNA sequence structure using a hidden Markov model (HMM) which is, in essence, a mixture model with Markov-dependent component indicators (Macdonald and Zucchini, 1997). Sequence analysis using HMMs is now a standard approach (Durbin et al., 1998) in the comparatively young science of bioinformatics and is a

[1] Corresponding author. Tel.: +61 2 49215547; fax: +61 2 49216898
E-mail address : Darfiana.Nur@newcastle.edu.au (D. Nur), David.Allingham@newcastle.edu.au (D. Allingham), rousseau@ceremade.dauphine.fr (J. Rousseau), k.mengersen@qut.edu.au (K. L. Mengersen), r.mcvinish@qut.edu.au (R. McVinish).

fundamental component of many gene-finding algorithms which identify and delineate genes in the human and other genomes (Defonzo, 2007).

Bayesian inference procedures and algorithms have revolutionized the field of computational biology (Liu and Logvinenko, 2003) due to the development of computationally-intensive simulation-based methods such as Markov chain Monte Carlo (MCMC), which are available in software such as WinBUGS (Lunn et al., 2000), and has led to the adoption of increasingly complex models in many situations.

A sometimes controversial aspect of the Bayesian approach is the need to specify prior distributions for the unknown parameters. In certain situations these priors may be very well defined. However, for complex models with many parameters, the choice of priors and conclusions of the subsequent Bayesian analysis are usually validated through a prior sensitivity analysis, as presented here.

For DNA sequence segmentation, a DNA sequence can be thought of as the observed process which evolves independently or dependently given an unobserved Markov chain which locates the position of the segment types. The parameters in this model are the base (nucleotide) transition probabilities for the segment types and the transition matrix of segment types. Boys et al. (2000) presented a Bayesian solution to the segmentation problem using HMMs when the number of segments is known. These results were generalised in Boys and Henderson (2004) to the case in which the number of segments is unknown. In Boys et al. (2000) and Boys and Henderson (2004), the prior knowledge for base transition probabilities in each segment was weak but the prior beliefs about the transition matrix for the segment types were strong. The authors discussed briefly the sensitivity of their conclusion to the choice of prior, especially for the transition matrix for the segment types, but no details were given. Their articles raise fundamental questions about limitations in model specification and bring to the forefront the issue of how far one can refrain from making prior assumptions about a model while keeping it feasible in practice. This prompts the important question of the impact of these priors on resultant inferences.

This paper has three main aims. The primary aim is to undertake a sensitivity analysis of the priors of a Bayesian hidden Markov model for DNA sequence segmentation. We employ Markov chain Monte Carlo via a short and easy-to-use program in BRuGS ("Bayesian analysis using Gibbs Sampler in R"). The sensitivity analysis includes a traditional approach, varying the prior distributions for base transition probabilities for each segment type and for the transition matrix of segment types. A sequence of Dirichlet priors is considered for the former and Dirichlet and mixture Dirichlet priors for the latter. The second aim of this paper is to introduce an alternative approach to sensitivity analysis that employs importance sampling of an MCMC chain obtained from the traditional approach. Our focus is on the feasibility and computational efficiency of this approach for comparing a large number of priors simultaneously in a more comprehensive sensitivity analysis. The results are applied to the segmentation of a benchmarking DNA sequence, intron 7 of the chimpanzee α-fetoprotein gene.

2 Methods

2.1. The hidden Markov model

A DNA sequence $\mathbf{y} = y_1, y_2, ..., y_n$ can be considered as a realisation of a random process $Y_1, Y_2, ..., Y_n$ where $Y_t \in \{a, c, g, t\}, t = 1, 2, ...n$, represent the four nucleotides adenine, cytosine, guanine and thymine, respectively, and n represents the length of the sequence. For convenience, the data can be encoded as 1,

2, 3, 4 for a, c, g and t, respectively. Suppose that there are at most r types of homogeneous segment within the DNA sequence. The (hidden) segment type at location t will be denoted by $S_t \in \{1, 2, ..., r\}$ for $t = 1, 2, ..., n$.

Assume that transitions between bases, $Y_{t-1} \rightarrow Y_t$, follow a first-order Markov chain, where the choice of transition matrix is determined by the hidden state S_t. Following Boys et al. (2000), we denote the 4×4 transition matrices for each segment type by $\mathcal{P} = \{P^{(1)}, P^{(2)}, ..., P^{(r)}\}$, where $P^{(k)} = (P^{(k)}_{ij})$. The update equations for the base transitions are

$$\begin{aligned} P(Y_t = y_t | S_t = s_t, y_1, ..., y_{t-1}, \mathcal{P}) &= P(Y_t = y_t | S_t = s_t, y_{t-1}, \mathcal{P}) \\ &= P^{(s_t)}_{y_{t-1} y_t}. \end{aligned} \tag{1}$$

The hidden state process of segment types is assumed to be a homogeneous first-order Markov chain with $r \times r$ transition matrix $\Lambda = (\lambda_{ij})$ such that

$$P(S_t = s_t | s_1, ..., s_{t-1}, \Lambda) = P(S_t = s_t | s_{t-1}, \Lambda) = \lambda_{s_{t-1} s_t}. \tag{2}$$

Assuming that Y_1 and S_1 follow independent discrete uniform distributions and by using (1) and (2), the likelihood for the model parameters $\mathcal{P}$ and Λ, given the observed DNA sequence $\mathbf{y}$ and the hidden segment types $\mathbf{s}$, is

$$L(\mathcal{P}, \Lambda | \mathbf{y}, \mathbf{s}) = \frac{1}{4r} \prod_{t=2}^{n} P^{(s_t)}_{y_{t-1} y_t} \lambda_{s_{t-1} s_t}.$$

The posterior distribution for the model parameters $\mathcal{P}$ and Λ and unobserved segment types $\mathbf{s}$ can be obtained by using Gibbs sampling with data augmentation. Let $\pi(\mathcal{P}, \Lambda | \mathbf{y}, \mathbf{s})$ be the posterior distribution of the parameters. Given the multinomial form of the likelihood function, we adopt a conjugate Dirichlet prior distribution, described in more detail below. Combining the likelihood with these priors using Bayes' theorem produces independent Dirichlet distributions for the rows of the transition matrices.

2.2. The priors

A choice of priors is available for the base transition probabilities for the segment types and the transition matrix of segment types.

A typical row of a base transition matrix and a row of the segment type transition matrix are denoted by $\mathbf{p}_i = (p_{ij})$ and $\lambda_k = (\lambda_{kj})$, respectively. Given that the likelihood is multinomial, the conjugate prior distributions for $\mathbf{p_i}$ and λ_k are a Dirichlet $\pi(\mathbf{p}_i) \propto \prod_{j=1}^{4} p_{ij}^{a_{ij}-1}$, $0 < p_{ij} < 1$, j=1,2,3,4, $\sum_{j=1}^{4} p_{ij} = 1$ and a Dirichlet $\pi(\lambda_k) \propto \prod_{k=1}^{r} \lambda_{kj}^{b_{kj}-1}$, $0 < \lambda_{kj} < 1$, k, j=1,2,...,r, where $\mathbf{a}_i = (a_{ij})$ and $\mathbf{b_k} = (\mathbf{b_{kj}})$ are the positive parameters of the distribution.

Similarly, following Boys et al. (2000), it is further assumed that the Dirichlet prior distributions, $\mathcal{D}$, for the rows of each transition matrix are independent, and therefore

$$\mathbf{p}_i^{(k)} = \left(p_{ij}^{(k)}\right) \sim \mathcal{D}(\mathbf{a}_i^{(k)}),\ i, j = 1, ..., 4,\ k = 1, ..., r, \tag{3}$$

$$\lambda_k = (\lambda_{kj}) \sim \mathcal{D}(\mathbf{b}_k),\ k, j = 1, ..., r. \tag{4}$$

Another possible prior distribution for this type of problem is a mixture of Dirichlet distributions (Brown et al., 1993), defined as

$$D = q_1 D_1 + ... + q_v D_v, \tag{5}$$

where $0 \leq q_i \leq 1$ such that $\sum_{i=1}^{v} q_i = 1$ and where each $D_i, i = 1, ..., v$, is a Dirichlet density function. The advantage of a mixture Dirichlet distribution lies in its flexibility for incorporating the segment types into the model.

The choice of prior for base transition matrices will depend on the purpose of analysis (Boys and Henderson, 2004). If the purpose is to screen general DNA sequences for possible homogeneous segments without any prespecified base transitions, as we pursue here, then a weak, or non-informative, prior should be used for $\mathcal{P}$. One such prior assigns $\mathbf{a}_i = (1, 1, 1, 1)$ along the rows in (3). Thus, components p_{ij} are distributed as Beta(1,3) with mean 1/4, standard deviation $\sqrt{3/80}$ and correlation between row elements $corr(p_{ij}, p_{ij'}) = -0.47$, for $j \neq j'$. Using this prior specification for each of the r segment types leads to an equivalent sequence length of $20r$ (Boys et al., 2000).

A generalisation of this prior, which we consider below, is

$$\mathbf{a}_i = (1 + \delta_1, 1 + \delta_2, 1 + \delta_3, 1 + \delta_4), \delta_m > 0, \ m = 1, .., 4, \tag{6}$$

along the rows in (3). A sensitivity analysis is thus facilitated by considering a range of values for each $\delta_m, m = 1, 2, 3, 4$.

Following Boys et al. (2000), a strong prior is assigned to the transition matrix for the hidden states. For detecting homogeneous segments, which are usually long segments, the assumptions in Boys et al. (2000) are adopted, namely that transitions between segment types are rare, that is, $E(\lambda_{kk}) \to 1$, and that the off-diagonal elements, λ_{kj} for $j \neq k$, in (4) are exchangeable so that the parameters of the Dirichlet distribution have the form $\mathbf{b}_k = (d, d, ..., d, c, d, ..., d, d)$ where c is the kth element. Again following Boys et al. (2000), we consider two cases: (i) fixing the mean $E(\lambda_{kk})$ and varying the standard distribution $sd(\lambda_{kk})$, and (ii) varying $E(\lambda_{kk})$ for a fixed $sd(\lambda_{kk})$. For a mixture Dirichlet, the number of components v of the mixture is set equal to the number of segments r, as in (5), where D_i follows (4) and the parameters have the form $\mathbf{b}_k = (d, d, ..., d, c, d, ..., d, d)$. A summary of the types of priors used for the sensitivity analysis is given in Table 1.

Table 1
Summary of prior types used in the sensitivity analysis

Type	$p_i^{(k)}$ of (3)	λ_k of (4)	$E(\lambda_{kk})$	$sd(\lambda_{kk})$
I	$\mathcal{D}(1, 1, 1, 1)$	Dirichlet	0.99	0.003 0.005 0.010
II	$\mathcal{D}(1, 1, 1, 1)$	Dirichlet	0.98 0.99 0.995	0.010
III	$\mathcal{D}(1, 1, 1, 1)$	Mixture Dirichlet	0.99	0.003 0.005 0.010
IV	$\mathcal{D}(1, 1, 1, 1)$	Mixture Dirichlet	0.98 0.99 0.995	0.010
V	$\mathcal{D}(1 + \delta_1, 1 + \delta_2,$ $1 + \delta_3, 1 + \delta_4)$	Dirichlet or Mixture Dirichlet	0.99	0.010

2.3. Implementation

The results in this paper were obtained using MCMC in BRuGS, with a burn-in period of 750,000 iterations followed by 100,000 updates of thinning 10. For each parameter, CODA tests were run to check the convergence of the chain. The CODA tests include Geweke, Gelman and Rubin, Raftery-Lewis and Heidelberger and Welch. Most of these tests confirmed convergence after the nominated burn-in. In general, parameters Λ and $\mathcal{P}^{(k)}, k = 1, ..., r$, passed the convergence diagnostics using Gelman-Rubin and Heidelberger-Welch tests, indicating that stationarity was achieved (Mengersen et al., 1999).

Two approaches to sensitivity analysis were employed. First, a traditional approach was used for prior types I to IV by generating separate MCMC chains for each combination of parameters of interest. However, running one chain of the MCMC algorithm for complex models such as DNA segmentation modelling is computationally demanding, and so the set of combinations that can be inspected in this way is limited. For example, due to the limitations of BRuGS, the average computing time was about 5 hours per run for this relatively small dataset on a standard desktop PC. To address this typical problem, one MCMC chain is run using a 'baseline' set of parameters and then importance sampling (IS) (Besag et al., 1995) is applied to this chain for each of the other combinations of parameters. Our second approach to sensitivity analysis involved implementing this procedure for $\mathcal{P}$ in order to more explicitly assess the impact of moving away from a Dirichlet prior with equal (uniform) parameters.

The IS approach is undertaken as follows. Define the statistic of interest, $H_\pi(x) = E_\pi[h(\theta)|x]$, of some given function h on θ for which $\{\theta\}_{t=1}^T$ is a Markov chain whose stationary distribution is $\pi(\theta|x)$, and let $\pi'(\theta)$ be another prior. We approximate $H_\pi(x)$ by

$$\hat{H}_{\pi'}(x) = \frac{\sum_{t=1}^T h(\theta^t)w(\theta^t)}{\sum_{t=1}^T w(\theta^t)}, \tag{7}$$

where $w(\theta) = \pi'(\theta^t)/\pi(\theta^t)$. The approximating properties of $\hat{H}_{\pi'}(x)$ to $H_\pi(x)$ are given in McVinish et al. (2008). In particular, if $\pi' \leq M\pi$ for some constant M then $\hat{H}_{\pi'}(x)$ has the same ergodic properties as $\hat{H}_{\pi'}(x) = \frac{\sum_{t=1}^T h(\theta^t)}{T}$ which is the usual MCMC estimator.

While estimates from IS will work with sufficient large sample sizes, it is important to be able to assess the accuracy of a given estimate. Assuming certain conditions hold the central limit theorem can be applied to both numerator and denominator of $\hat{H}$. In particular if both w^2 and $h^2w^2(\theta)$ satisfy the drift condition of Theorem 17.0.1 of Meyn and Tweedie (1993), then there exist a central limit theorems for the numerator and the denominator and, applying the delta method it is seen that

$$(\gamma^2)^{-1/2}\left(\hat{H} - H_{\pi'}\right) \xrightarrow{d} \mathcal{N}(0,1),$$

where

$$\gamma^2 = [E^\pi(w(\theta))]^2\left[H_{\pi'}^2 \mathrm{var}(\bar{w}) + \mathrm{var}(\bar{hw}) - 2\mathrm{cov}(\bar{w}, \bar{hw})H_{\pi'}\right]$$

and

$$\bar{hw} = T^{-1}\sum_{t=1}^T h(\theta^t)w(\theta^t), \quad \bar{w} = T^{-1}\sum_{t=1}^T w(\theta^t).$$

In order to avoid estimating the variances and covariance separately we propose the following approach. Suppose both $E_\pi(w(\theta))$ and $E_\pi(h(\theta)w(\theta))$ were known. The sample mean of the series

$$Z(\theta_t) = \frac{E_\pi(h(\theta)w(\theta))}{E_\pi(w(\theta))}\left(\frac{w(\theta^t)}{E_\pi(w(\theta))} - \frac{h(\theta^t)w(\theta_t)}{E_\pi(h(\theta)w(\theta))}\right)$$

can be seen to have the same asymptotic variance as $\hat{H}$. Replacing $E_\pi(w(\theta))$ and $E_\pi(h(\theta)w(\theta))$ with their respective sample means an estimate of γ^2 can be obtained using one of the methods described in Geyer (1992). When conducting a sensitivity analysis where $\pi(\theta)$ is flatter than $\pi'(\theta)$, then $w(\theta)$ will typically be bounded and a central limit theorem for $\hat{H}$ is available with the above asymptotic variance as soon as the MCMC chain is V-ergodic with $V \geq 1$.

An MCMC run was first undertaken to obtain the posterior distribution for $\mathcal{P}$ based on the uniform Dirichlet prior in (3) with $a_i = (1,1,1,1), i = 1,..,4$. We then used IS to conduct a local sensitivity analysis evaluating the posterior distribution for $\mathcal{P}$ based on a_i in (6) corresponding to all 80 combinations of $\delta_m = 0,1,2$ and $m = 1,2,3,4$ simultaneously, as shown in Section 3.4. This is also consistent with the condition that the baseline prior is flatter than the proposed priors.

Because of the smaller scale of the traditional sensitivity analysis, the results are evaluated by comparison of the posterior estimates and corresponding precision of the segment type locations and transition probabilities. The large-scale sensitivity analysis based on importance sampling is evaluated by computing the sum squared of distances between posterior distributions, calculated as $\Delta_l = \sum_{i=1}^{4}\sum_{j=1}^{4}|\hat{P}_{ijl}^{(k)} - \hat{P}_{ij1}^{(k)}|^2$, $l = 1,...,81$, where $\hat{P}_{ij1}^{(k)}$ are the posterior mean estimates of base transition probabilities given in segment k, as denoted by $P_{ij}^{(k)}$, obtained using MCMC and $\hat{P}_{ijl}^{(k)}$, $l = 2,...,81$ are obtained via IS. Larger values of Δ_l indicate greater sensitivity to the prior specification.

3 Results

3.1. The data

The structure of introns (non-coding regions within genes) is of interest because irreversible transpositions and other mutations are more likely to be preserved in these regions than in exons (coding regions), and hence the intron structure becomes a more reliable time guide for deciphering phylogenies (Nishio et al.,1995). A comprehensive review of segmentation methods for DNA sequences can be found in Braun and Müller (1998).

The intron data in which we are interested is intron 7 of the chimpanzee α-fetoprotein gene (Nishio et al.,1995) which does contain distinct homogeneous segments (Boys et al., 2000). This gene plays an important role in embryonic development in mammals; in particular unusual levels in pregnant women are associated with fœtal genetic disorders such as spina bifida and Down's syndrome. This gene is 18,867 base pairs long, composed of 15 exons separated by 14 introns. The full nucleotide sequence is in the GenBank sequence database under accession number U21916. Intron 7 starts at nucleotide 11,712 from 5' end of the α-fetoprotein gene and has a length of 1,968 base pairs.

3.2. Posterior summaries for Λ and $\mathcal{P}$

Table 2 presents the posterior mean and standard deviation estimates (to two decimal places) for the segment transition probabilities, Λ, and for the base transition probabilities, $\mathcal{P}$. Results are given for different prior standard deviations using both Dirichlet and mixture Dirichlet priors for Λ. The precision of the estimates are evaluated by the standard deviation estimates for the corresponding parameters.

Table 2

Posterior summaries of Λ and P using prior types I and III for a fixed $E(\lambda_{kk}) = 0.99$ and various standard deviations $sd(\lambda_{kk})$.

	Type I								Type III							
	Λ								Λ							
sd	mean				standard deviation				mean				standard deviation			
0.003	0.99	0.01	0.01		0.00	0.00	0.00		0.99	0.00	0.01		0.00	0.00	0.00	
	0.26	0.33	0.40		0.10	0.11	0.11		0.01	0.98	0.01		0.00	0.00	0.00	
	0.00	0.00	0.99		0.00	0.00	0.00		0.99	0.00	0.01		0.00	0.00	0.00	
0.005	0.99	0.01	0.01		0.01	0.00	0.00		0.99	0.01	0.00		0.00	0.00	0.00	
	0.23	0.34	0.43		0.15	0.18	0.19		0.00	0.99	0.01		0.00	0.00	0.00	
	0.00	0.00	0.99		0.00	0.00	0.00		0.00	0.99	0.01		0.00	0.00	0.00	
0.01	0.99	0.01	0.00		0.00	0.00	0.00		0.98	0.01	0.01		0.00	0.00	0.00	
	0.01	0.98	0.01		0.01	0.01	0.01		0.00	0.99	0.01		0.00	0.00	0.00	
	0.01	0.00	0.99		0.01	0.01	0.01		0.01	0.98	0.01		0.00	0.01	0.00	
sd	$P^{(1)}$								$P^{(1)}$							
0.003	0.26	0.25	0.19	0.30	0.17	0.17	0.13	0.16	0.35	0.15	0.26	0.24	0.02	0.02	0.02	0.02
	0.05	0.43	0.01	0.51	0.03	0.06	0.01	0.06	0.34	0.21	0.04	0.41	0.03	0.02	0.01	0.03
	0.21	0.22	0.18	0.39	0.14	0.12	0.12	0.15	0.33	0.20	0.24	0.23	0.02	0.02	0.02	0.02
	0.03	0.37	0.06	0.54	0.02	0.05	0.03	0.05	0.23	0.16	0.22	0.38	0.02	0.02	0.02	0.02
0.005	0.31	0.21	0.22	0.26	0.17	0.15	0.12	0.14	0.30	0.23	0.19	0.28	0.17	0.16	0.12	0.15
	0.05	0.42	0.01	0.52	0.03	0.06	0.01	0.06	0.05	0.42	0.01	0.52	0.03	0.06	0.01	0.06
	0.26	0.22	0.16	0.36	0.14	0.11	0.11	0.14	0.24	0.22	0.17	0.36	0.14	0.12	0.11	0.14
	0.03	0.36	0.07	0.53	0.02	0.05	0.03	0.05	0.04	0.37	0.07	0.52	0.02	0.05	0.03	0.05
0.01	0.35	0.16	0.23	0.26	0.02	0.02	0.02	0.02	0.31	0.22	0.20	0.27	0.17	0.15	0.11	0.14
	0.34	0.21	0.04	0.41	0.03	0.02	0.01	0.03	0.05	0.42	0.02	0.51	0.03	0.06	0.01	0.06
	0.32	0.20	0.21	0.26	0.03	0.02	0.03	0.03	0.25	0.21	0.17	0.37	0.14	0.11	0.11	0.14
	0.23	0.17	0.21	0.39	0.02	0.02	0.02	0.02	0.04	0.36	0.07	0.53	0.02	0.05	0.03	0.05
sd	$P^{(2)}$								$P^{(2)}$							
0.003	0.26	0.25	0.23	0.26	0.20	0.19	0.19	0.20	0.15	0.36	0.14	0.35	0.13	0.17	0.12	0.17
	0.26	0.24	0.22	0.28	0.20	0.19	0.18	0.20	0.04	0.44	0.01	0.50	0.03	0.06	0.01	0.06
	0.26	0.23	0.24	0.27	0.20	0.19	0.19	0.20	0.14	0.25	0.24	0.37	0.12	0.14	0.14	0.17
	0.25	0.24	0.27	0.24	0.19	0.19	0.20	0.19	0.03	0.38	0.05	0.54	0.02	0.05	0.02	0.05
0.005	0.27	0.24	0.23	0.26	0.20	0.19	0.19	0.20	0.35	0.15	0.25	0.24	0.02	0.02	0.02	0.02
	0.26	0.24	0.22	0.28	0.19	0.19	0.18	0.21	0.35	0.21	0.05	0.39	0.03	0.02	0.01	0.03
	0.25	0.24	0.24	0.27	0.19	0.19	0.19	0.20	0.33	0.20	0.24	0.23	0.03	0.02	0.02	0.02
	0.25	0.25	0.26	0.24	0.19	0.19	0.19	0.19	0.24	0.16	0.22	0.38	0.02	0.01	0.01	0.02
0.01	0.34	0.16	0.22	0.28	0.03	0.02	0.02	0.03	0.19	0.33	0.15	0.33	0.16	0.18	0.13	0.17
	0.34	0.21	0.04	0.41	0.03	0.03	0.01	0.03	0.04	0.45	0.02	0.49	0.03	0.06	0.01	0.06
	0.32	0.21	0.21	0.27	0.03	0.03	0.03	0.03	0.13	0.26	0.25	0.36	0.12	0.15	0.14	0.17
	0.23	0.17	0.20	0.40	0.02	0.02	0.02	0.03	0.03	0.39	0.05	0.53	0.02	0.06	0.02	0.06
sd	$P^{(3)}$								$P^{(3)}$							
0.003	0.35	0.15	0.26	0.24	0.02	0.01	0.02	0.01	0.25	0.25	0.25	0.25	0.19	0.19	0.19	0.19
	0.34	0.21	0.05	0.40	0.03	0.02	0.01	0.03	0.25	0.24	0.24	0.27	0.19	0.19	0.19	0.20
	0.33	0.20	0.24	0.23	0.02	0.02	0.02	0.02	0.25	0.25	0.25	0.25	0.19	0.19	0.20	0.19
	0.24	0.16	0.22	0.38	0.02	0.01	0.01	0.02	0.27	0.25	0.23	0.25	0.20	0.20	0.19	0.19
0.005	0.35	0.16	0.25	0.24	0.02	0.01	0.02	0.02	0.25	0.25	0.25	0.25	0.20	0.19	0.19	0.19
	0.34	0.21	0.05	0.40	0.03	0.02	0.01	0.03	0.26	0.24	0.24	0.26	0.20	0.19	0.19	0.20
	0.33	0.20	0.24	0.23	0.02	0.02	0.02	0.02	0.25	0.25	0.24	0.26	0.19	0.19	0.19	0.20
	0.24	0.16	0.22	0.38	0.02	0.01	0.02	0.02	0.25	0.25	0.26	0.24	0.19	0.19	0.19	0.19
0.01	0.32	0.11	0.52	0.05	0.13	0.06	0.14	0.04	0.25	0.24	0.25	0.25	0.19	0.19	0.19	0.20
	0.33	0.24	0.13	0.30	0.11	0.09	0.07	0.09	0.25	0.24	0.24	0.26	0.19	0.19	0.18	0.20
	0.33	0.19	0.36	0.11	0.07	0.06	0.06	0.04	0.25	0.25	0.24	0.26	0.19	0.19	0.19	0.20
	0.16	0.12	0.63	0.09	0.10	0.09	0.12	0.07	0.25	0.25	0.25	0.25	0.19	0.20	0.19	0.19

Under a Dirichlet prior for Λ, it appears that varying the standard deviation of the parameters has a substantial impact on the posterior mean estimates. In particular, increasing the standard deviation of these parameters, facilitating greater movement around the parameter space, enables a satisfactory solution to be found: compare, for example, the posterior standard deviation estimates for Λ, $P^{(1)}$ and $P^{(2)}$ for the prior with $sd = 0.01$ with those for $sd = 0.003$ and 0.005. On the other hand, there is a suggestion from the results for $P^{(3)}$ with $sd = 0.01$ that excessive vagueness in the priors leads to poorer estimates compared with those for $sd = 0.003$ and 0.005. Furthermore, under a Dirichlet prior for Λ, the posterior mean estimates of $P^{(2)}$ for $sd = 0.003, 0.005$ are similar to those of the assigned uniform prior.

The comparative flexibility of mixture Dirichlet priors for Λ is also reflected in this table: it appears that satisfactory posterior mean estimates (with smaller posterior standard deviations) can be identified under a wider range of prior sd values. The estimates of transition probabilities from the third segment to others are close to 1 suggesting the possibility of two segments model. Moreover, under a mixture Dirichlet prior for Λ, the posterior mean estimates of $P^{(3)}$ for all sd priors are similar to those of the uniform prior, showing that the existence of the third segment can be neglected. In the case of $P^{(2)}$, a dramatic decrease in standard deviations is observed when the mean converges to a satisfactory estimate. Similar results were obtained for type II and IV priors, with the mixture Dirichlet prior again outperforming the Dirichlet prior, and are omitted here.

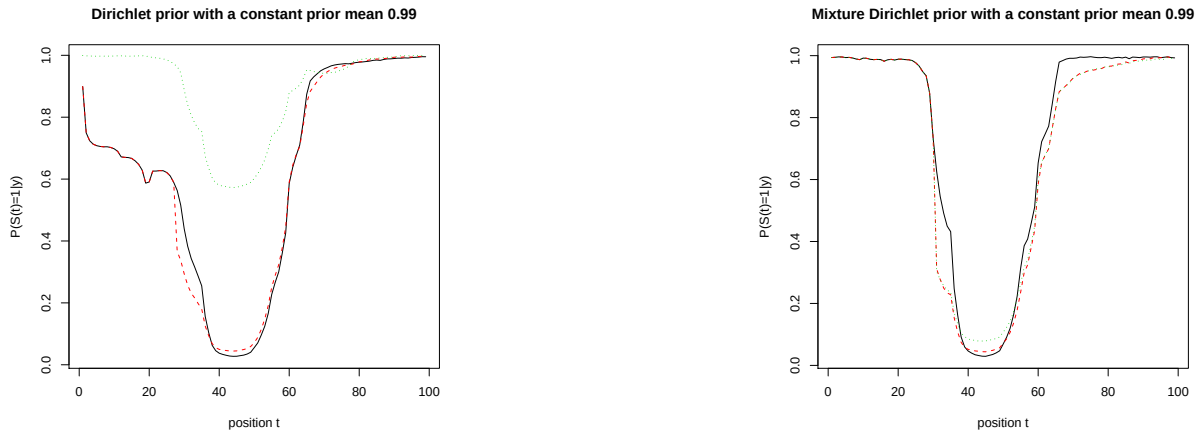

Fig. 1. Posterior probability of each location belonging to segment 1, with $r = 3$, using Dirichlet prior (left) and mixture Dirichlet prior (right), for the case of fixed mean $E(\lambda_{kk}) = 0.99$ and different standard deviations: 0.003 (solid line), 0.005 (dashed line) and 0.010 (dotted line).

3.3. Posterior summaries for segments

Figure 1 shows a close-up of the estimates of $P(S_t = 1|\mathbf{y})$ in the first 100 locations of the sequence, varying the standard deviation while fixing the mean for both the Dirichlet prior and the mixture Dirichlet prior. In the case of the Dirichlet prior, the plot shows that changes in segment type become less certain as the prior uncertainty about λ_{kk} increases. For example, in locations 40 to 50 for $sd = 0.01$, the probability of being in segment 1 is close to 0.5, but only 0.05 for $sd = 0.003$ and 0.005. On the other hand, for the mixture Dirichlet prior the variation among the three standard deviations is less for locations 30 to 100, showing that the analysis is less sensitive to the mixture Dirichlet prior than to the Dirichlet prior.

Figure 2 shows similar plots in which the mean is varied while fixing the standard deviation for the Dirichlet and mixture Dirichlet priors. These illustrations confirm that under the Dirichlet prior, lower mean values are associated with more frequent changes in segment type; for example, for mean 0.98, the sequence is in segment 1 for locations 1 to 30, fluctuates until location 60 and stays in segment 1 afterwards. A similar but stronger conclusion is evident for the mixture Dirichlet prior.

Figure 3 presents a further close-up of the posterior probability estimates for each location, $P(S_t = k|\mathbf{y}), k = 1, 2, 3$ in the first 100 locations of the sequence, for fixed mean $E(\lambda_{kk}) = 0.99$ and fixed standard deviation $sd(\lambda_{kk}) = 0.003$, for both Dirichlet and mixture Dirichlet priors. Whilst the Dirichlet prior estimates a small probability of the existence of segment type 3 in the first 100 locations, the mixture Dirichlet prior shows a stronger result, with probabilities are almost zero. The reduced uncertainty apparent in the latter case reflects the insensitivity to the mixture Dirichlet prior.

3.4. Posterior summaries for $\mathcal{P}$

The IS algorithm was used to compare 81 priors simultaneously, representing combinations 3^4 of $\delta_m = 0, 1, 2,\ m = 1, 2, 3, 4$ in (6). An MCMC run was first undertaken using the uniform Dirichlet prior $\delta_m = 0$, for $m = 1, 2, 3, 4$, and the results for the other priors were derived using IS as described in Section 2.3. Figure

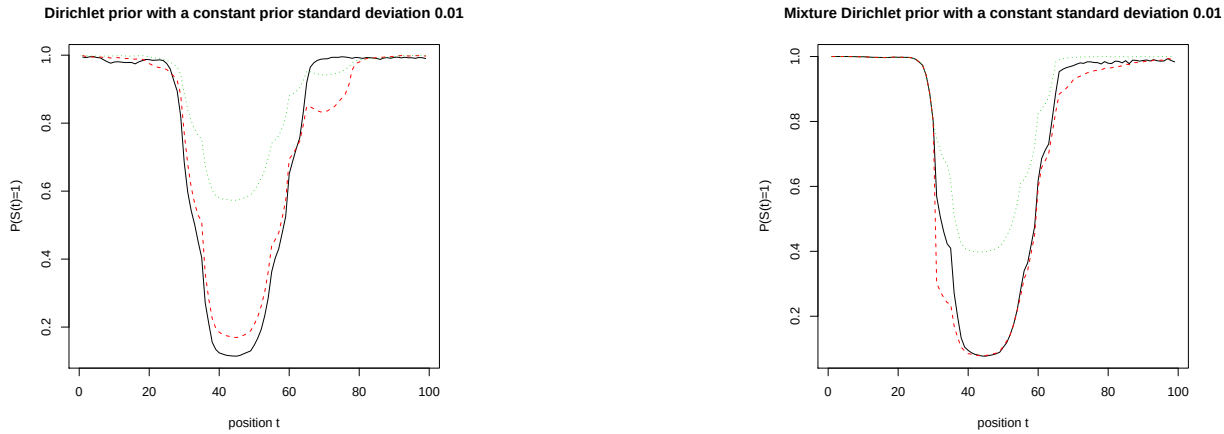

Fig. 2. Posterior probability of each location belonging to segment 1 with $r = 3$ using Dirichlet prior (left) and mixture Dirichlet prior (right), for the case of fixed standard deviation $sd(\lambda_{kk}) = 0.01$ and different means: 0.98 (solid line), 0.99 (dashed line) and for 0.995 (dotted line).

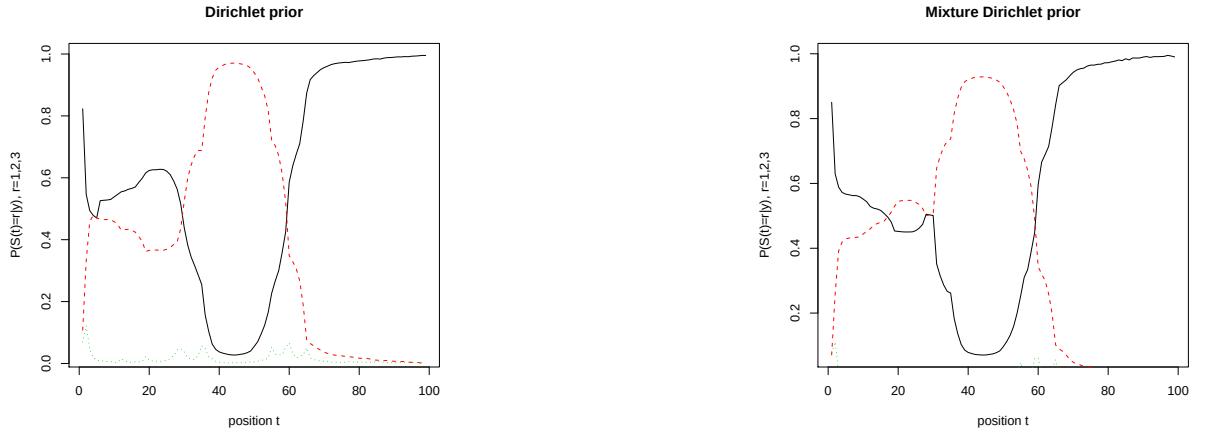

Fig. 3. Posterior probability of each location belonging to each segment type, segment 1 (solid line), segment 2 (dashed line) and segment 3 (dotted line), using Dirichlet prior (left) and mixture Dirichlet prior (right), for the case of fixed mean $E(\lambda_{kk}) = 0.99$ and fixed standard deviation $sd(\lambda_{kk}) = 0.003$.

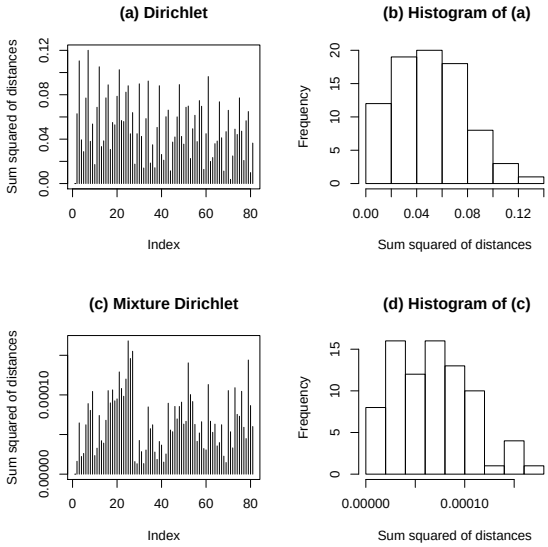

Fig. 4. Importance sampling distances using 81 Dirichlet priors for $\mathcal{P}$ and a Dirichlet or mixture Dirichlet prior for Λ with $E(\lambda_{kk})$=0.99, $sd(\lambda_{kk})$=0.01. (a) Sum of squared distances Δ_l, where "Index" refers to $l = 1, 2, ..., 81$ for a Dirichlet; (b) histogram of the values in (a); (c) sum of squared distances Δ_l, where "Index" refers to $l = 1, 2, ..., 81$ for a mixture Dirichlet; (d) histogram of the values in (c).

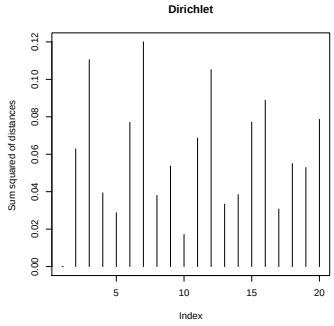

Fig. 5. The first 20 sum of squared distances Δ_l, where "Index" refers to $l = 1, 2, ..., 20$ for the Dirichlet of Figure 4(a).

4 shows the sum of squared differences, $\Delta_l, l = 1, ..., 81$, between posterior mean estimates for all P_{ij}, for $i = 1, 2, 3, 4$ and $j = 1, 2, 3, 4$, and those from the uniform Dirichlet prior. Figures 4(a) and (b) show results for a Dirichlet prior on Λ with a fixed mean 0.99 and standard deviation 0.01; Figures 4(c) and (d) show results for a mixture Dirichlet prior. In the former case, the sum of squared differences range between 0 and 0.12, whereas in the latter case, less sensitivity is exhibited, with much tighter ranges. Furthermore, we can see that estimates of $\mathcal{P}$ are potentially quite sensitive to the change in priors. This is shown more clearly in Figure 5; there is, for example, a substantial difference between the posterior estimates under a uniform prior and priors Dir(3,1,1,1) (index 3) and Dir(1,3,1,1) (index 7).

Finally, a model checking is briefly presented. One technique for checking the fit of a model to data is to draw simulated values from the posterior predictive distribution of replicated data and compare these samples to the observed data (Gelman, et al., 1993). From the previous results, it is apparent that two-segments HMM is preferred. Using the Type I prior with sd=0.01, the posterior predictive distribution was simulated using a two-segments HMM with the estimated parameters

$$\hat{\Lambda} = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix},$$

$$\hat{\mathcal{P}}^{(1)} = \begin{pmatrix} 0.35 & 0.15 & 0.26 & 0.24 \\ 0.34 & 0.21 & 0.05 & 0.40 \\ 0.33 & 0.20 & 0.24 & 0.23 \\ 0.23 & 0.16 & 0.22 & 0.38 \end{pmatrix},$$

$$\hat{\mathcal{P}}^{(2)} = \begin{pmatrix} 0.17 & 0.34 & 0.15 & 0.34 \\ 0.04 & 0.45 & 0.01 & 0.50 \\ 0.14 & 0.26 & 0.23 & 0.37 \\ 0.03 & 0.39 & 0.05 & 0.53 \end{pmatrix}.$$

The comparison of timeplots is depicted in Figure 6. It is clear that in general the model fits the data well, except that the model seems to underestimate or overestimate the observed data in some locations. For example, the nucleotide 'c' is underestimated more often compared to other nucleotides. But in the other hand, the nucleotides 'a' and 'g' were overestimated in the last 300 positions. Further methods to measure the goodness of fit can be elaborated using a test quantity such as a discrepancy measure or by calculating posterior predictive p-values.

4 Discussion

We have performed a thorough local sensitivity analysis on priors in a Bayesian analysis of DNA sequence segmentation using hidden Markov models. Three main aims were addressed: (i) to assess the sensitivity to prior specification in hidden Markov models for DNA sequence segementation analysis; (ii) to examine the impact of replacing the standard Dirichlet prior with a mixture of Dirichlet distributions prior; and (iii) to propose and illustrate a more comprehensive approach to sensitivity analysis using importance sampling. We observed that (i) the posterior estimates obtained under were indeed sensitive to the specification and precision of prior distributions imposed on the segment types and transition matrix of segment types; (ii) compared with the standard Dirichlet prior, the mixture Dirichlet prior was more flexible to define, less sensitive to the choice of hyperparameters and less constraining in the analysis, thus improving posterior estimates; and (iii) the importance sampling was computationally feasible, fast and effective in allowing a richer sensitivity analysis.

Two implications for practice arise from this work. First, priors are an integral and influential part of complex Bayesian models such as the one considered here. A sensitivity analysis thus becomes crucial

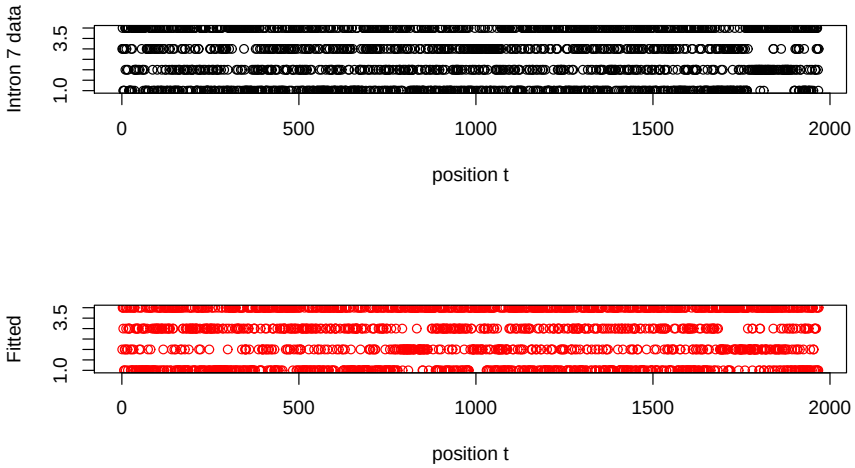

Fig. 6. The comparison of the intron 7 data (top) with the observations obtained using the model (bottom).

in assessing the impact of prior specification, particularly where these are *noninformative*. Secondly, the traditional approach of evaluating a limited number of alternative priors is not only inefficient but also potentially misleading. Innovative computational approaches, such as the importance sampling algorithm described here, enable a comprehensive evaluation of the space of possible priors, allowing us to adequately assess the impact of prior choice on posterior inference.

References

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems, *Stat. Sci.*, **10**, 3-66.

Boys, R.J, Henderson, D.A. and Wilkinson, D.J. (2000) Detecting homogeneous segments in DNA sequences by using Hidden Markov models, *Appl. Stat.*, **49**, 269-285.

Boys, R.J. and Henderson, D.A. (2004) A Bayesian approach to DNA sequence segmentation, *Biometrics*, **60**, 573-588.

Braun, J.V. and Müller, H.-G. (1998) Statistical methods for DNA sequence segmentation, *Stat. Sci.*, **13**, 142-162.

Brown, M.P., Hughey, R., Krogh, A., Mian, I.S., Sjölander, K. and Haussler, D. (1993) Using Dirichlet mixture priors to derive hidden Markov models for protein families. In Hunter, L., Searls, D. and Shavlik, J. (eds), *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology (ISMB-93)*. AAAI/MIT Press, Menlo Park, CA, pp. 47-55.

Churchill, G.A. (1989) Stochastic models for heterogeneous DNA sequences, *B. Math. Biol.*, **51**, 79-94.

De Fonzo, V., Aluffi-Pentini, F. and Parisi, V. (2007) Hidden Markov models in bioinformatics, *Current Bioinformatics*, **2**, 49-61.

Durbin, R. and Eddy, S.R. and Krogh, A. and Mitchison, G. (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian data analysis*. Chapman and

Hall, London.

Geyer, C. (1992) Practical Markov chain Monte Carlo (with discussion) *Statistical Science* **7**:473–483.

Liu, J.S. and Logvinenko, T. (2003) Bayesian methods in biological sequence analysis. In Balding, D.J., Bishop, M. and Cannings, C. (eds), *Handbook of Statistical Genetics*, 2nd edition. John Wiley & Sons, Chichester, pp. 66-93.

Lunn, D.J., Thomas, A., Best, N. and Spiegelhalter, D. (2000) WinBUGS – A Bayesian modelling framework: Concepts, structure, and extensibility, *Stat. Comput.*, **10**, 325-337.

MacDonald, I.L. and Zucchini, W. (1997) *Hidden Markov and Other Models for Discrete-Valued Time Series*. Chapman and Hall, London.

McVinish, R., Mengersen, K. L., Nur, D., Rousseau, J. and Guihenneuc-Jouyaux, C. *Use of Importance sampling for repeated MCMC*. (School of Mathematical Sciences, Queensland Univeristy of Technology, 2008).

Mengersen, K.L., Robert, C.P. and Guihenneuc-Jouyaux, C. (1999) MCMC convergence diagnostics: a "reviewww". In Berger, J., Bernardo, J., Dawid, A.P. and Smith, A.F.M. (editors), *Bayesian Statistics 6*. Oxford Sciences Publications, pp. 415-440.

Meyn, S.P. and Tweedie, R.L. (1993) *Markov Chains and Stochastic Stability*, Springer-Verlag, London.

Nishio, H., Gibbs, P.E.M, Minghetti, P.P, Zielinski, R. and Dugaiczyk, A. (1995) The chimpanzee α-fetoprotein-encoding gene shows structural similarity to that of gorilla but distinct differences from that of human, *Gene*, **162**, 213-220.