Bayesian Analysis (2011) **6**, Number 4, pp. 529–532

Comment on Article by Wyse et al.

Paul Fearnhead*

I would like to start by congratulating the authors on a stimulating paper. The focus of the paper is on computationally efficient methods for analysing multiple changepoint models. They build on existing methods (Yao 1984; Barry and Hartigan 1992; Liu and Lawrence 1999; Fearnhead 2006) that allow for iid sampling from the posterior distribution for certain changepoint models. These existing methods require the ability to analytically calculate the marginal likelihood associated with any segment within the data. For analysing a data set with n data points, under the assumption of K changepoints, the CPU cost of these methods is $O(Kn^2)$.

There are two key ideas within this paper. The first is to use ideas from the integrated nested Laplace approximation (INLA) to approximate the segment marginal likelihoods. This enables you to apply these recursive methods to a much wider range of changepoint models, that is to models which include dependence of data within a segment when it is modelled through a Gaussian Markov random field. The second is to implement the recursions on a reduced set of possible changepoint times. This can be thought of as grouping each g consecutive data points into a single observation, and then running the recursions on this reduced set of observations. The motivation for this is purely computational – as it reduces the CPU cost of the recursions by a factor of g^2. In many applications this approximation has a natural interpretation. For example, for the coal-mining disaster data (Section 4 of the paper), different choices of g would relate to analysing data at different levels of aggregation: such as corresponding to data on the number of deaths each day, week, or year. Providing the distance between successive changepoints is larger relative the level of aggregation, we would expect any approximation error to be small.

The key feature of both ideas is to introduce some approximation, but with the gain of being able to analyse a much wider class of models and a much bigger size of data set. I would like to first discuss, via asymptotic arguments, in what sort of situations these approximations are likely to be small; and secondly to look at the idea and approach of summarising the inferences via a MAP estimate of changepoint positions.

*Department of Mathematics and Statistics, Lancaster University, Lancaster, UK, p.fearnhead@lancs.ac.uk

 DOI:10.1214/11-BA620A

1 Asymptotics

It is natural to consider what happens to the approximations as we analyse increasingly larger data sets. As we let the number of observations, $n \to \infty$, there are two extreme scenarios we could consider in terms of whether the number of changepoints, K, is fixed or whether on average it increases linearly with n. The former case is often called in-fill asymptotics, and corresponds to taking observations at higher frequency across a fixed region of interest. For example, if you are interested in a region of the genome and are making inferences based on single-nucleotide polymorphism (SNP) data, this regime would relate to having data from an increasing density of SNPs. The latter case corresponds to a fixed frequency of observations, but analysing an ever-increasing region. For the genetics example, this would correspond to SNP data at a fixed density, but analysing an increasing amount of a chromosome.

It seems that the methods in this paper are designed for the in-fill asymptotics (which in practice corresponds to $n >> K$). If we consider a fixed g, then in this limit the segment length will be much larger than g, and hence the error in using the reduced set of possible changepoint positions will be small. Furthermore, under in-fill asymptotics we may expect the error of INLA's approximation of the segment marginal likelihoods would be negligible. In fact, I wonder if there may be some theoretical results available to show that under these asymptotics, the proposed method would give consistent estimates of the changepoint positions. Such results may also allow g to increase with n, providing the rate is less than linear.

By comparison, the methods in the paper seem less suited to the other regime (which in practice corresponds to both n and K being large). However, there are alternative approaches that work well in this case. Firstly, for this sort of regime it feels natural to have a prior which models the segment lengths as being iid. This would naturally result in a model where the number of changepoints increases on average linearly as you analyse more data. One computational advantage of such a prior is that there are recursions for analysing changepoint models which are $O(n^2)$ rather than $O(Kn^2)$. This is already an important saving if $K = O(n)$.

Secondly, there are ways of approximating the recursions based on pruning possible values of the most recent changepoint prior to each time-point t (for such an approach, using ideas from particle filtering see Fearnhead and Liu 2007). Essentially these remove the computations that are related to certain large segments. In many applications the probability of these large segments is negligible, and these approximations can lead to

an $O(n)$ algorithm that has small approximation error.

Obviously there will be applications where both K and n/K are large. In these situations using a combination of this pruning idea, together with the use of a reduced set of possible changepoint location, would be sensible.

2 MAP estimation

Whilst it is useful to be able to summarise data through a point estimate of the number and position of changepoints, much information is lost if this is the sole output. For example consider inferences about the underlying latent field (e.g. as in Figure 8 for the well-log data). These inferences are now conditional on a specific estimate of the changepoint positions, and do not allow for any uncertainty in the inferences about the number and positions of the changepoints. In regions where the changepoints are clear, and there is little uncertainty about changepoint positions, this may not matter, but for applications where the changepoints are hard to detect this could have a sizeable impact on future inferences.

The other issue with using a MAP estimate to summarise inference about the changepoints is that the MAP estimate is not well-defined. For example whether you first choose the MAP estimate of the number of changepoints, and then the MAP estimate of their positions conditional on this number, or you use the joint MAP estimate of the number and position of changepoints, can lead to different estimates (Fearnhead 2005). Similarly, the suggested approach in the paper is to recursively choose the MAP estimate of the position of the first changepoint then the MAP estimate of the position of the second given the position of the first, until you calculate the MAP position of the final changepoint given the positions of the earlier ones. I believe this will give a different answer to calculating the joint MAP estimate of all changepoint positions. (It is possible to calculate the joint MAP estimate using a Viterbi algorithm, see Viterbi 1967; Fearnhead 2005).

One advantage of MAP estimation, rather than aiming to simulate from the posterior, is that there can be ways of substantially reducing the computational cost (as discussed in Section 2), but without introducing any error (see Killick et al. 2011, for an algorithm with complexity that can be $O(n)$).

References

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

Fearnhead, P. (2005). "Exact Bayesian curve fitting and signal segmentation." *IEEE Transactions on Signal Processing*, 53: 2160–2166.

— (2006). "Exact and efficient inference for multiple changepoint problems." *Statistics and Computing*, 16: 203–213.

Fearnhead, P. and Liu, Z. (2007). "Online inference for multiple changepoint problems." *Journal of the Royal Statistical Society Series B*, 69: 589–605.

Killick, R., Fearnhead, P., and Eckley, I. A. (2011). "Optimal detection of changepoints with a linear computational cost." *arXiv:1101.1438v2 [stat.ME]*.

Liu, J. S. and Lawrence, C. E. (1999). "Bayesian inference on biopolymer models." *Bioinformatics*, 15: 38–52.

Viterbi, A. J. (1967). "Error bounds for convolutional codes and an asymptotically optimal decoding algorithm." *IEEE Transactions on Information Theory*, 13: 260–269.

Yao, Y. (1984). "Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches." *The Annals of Statistics*, 12: 1434–1447.