Bayesian Dynamic Regression Trees

Inference and Learning for Streaming Data

Michael Ferreira (ferreima@tcd.ie) and Simon P Wilson (swilson@tcd.ie)
Trinity College Dublin

1. Contribution

We introduce a flexible system of modelling streaming data in a Bayesian regression setting. We combine two regression methods, Bayesian Cart by Chipman et al. 1998 and the Kalman Filter as derived by Meinhold and Singpurwalla 1983 because both minimise the mean square error based on the conditional expectation. Like Chipman et al. 2010 we form an ensemble of trees and perform inference over the weighted sum of the trees. Similar work has been done by Taddy et al. 2011 and Gramacy and Lee 2008.

2. Trees and Filters

A tree divides up a large covariate space, $\mathcal{X}$, using splitting threshold rules which assign observations to each of the partitions. This both provides a prior structure on the covariate space and concentrates the likelihood of the observations to each partition based on the conditional expectation: $E[y_t \mid X_{t,i} \dots X_{t,i+j}] = z_t$

Focusing estimation on these partitions induced by the data is a well trodden method and is successful in many machine learning fields.

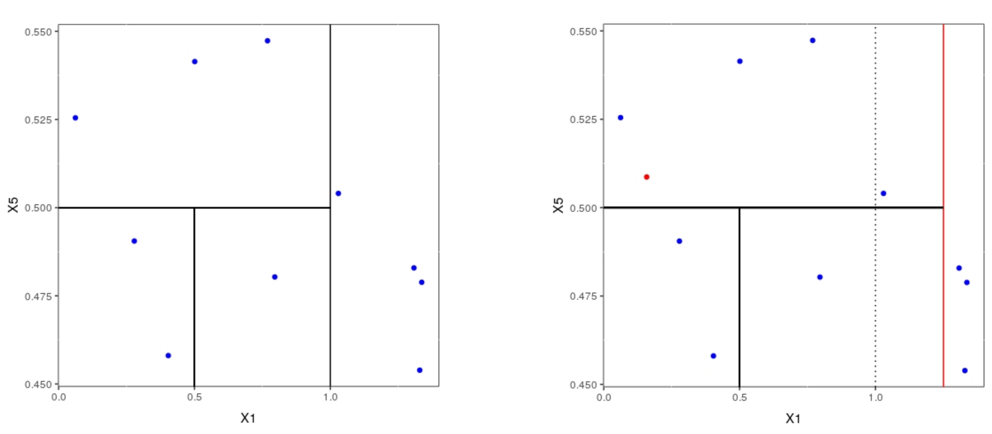

Figure 1: An evolving tree space.

The Kalman filter prediction of the next observation is based on the conditional expectation of the previous state: $E[y_t \mid z_t] = HFz_{t-1}$, where $y_t = Hz_t + v_t$ and $z_t = Fz_{t-1} + w_t$

Again, this is a well known method in stochastic control and there exist many improvements and extensions, for instance, parameter estimation and non-linear dynamics. Using an adaptation of the Kalman filter to a sensor network, Sinopoli et al. 2004 developed an intermittent Kalman filter which we use as means of updating the $1 \dots K_T$ filters of each tree.

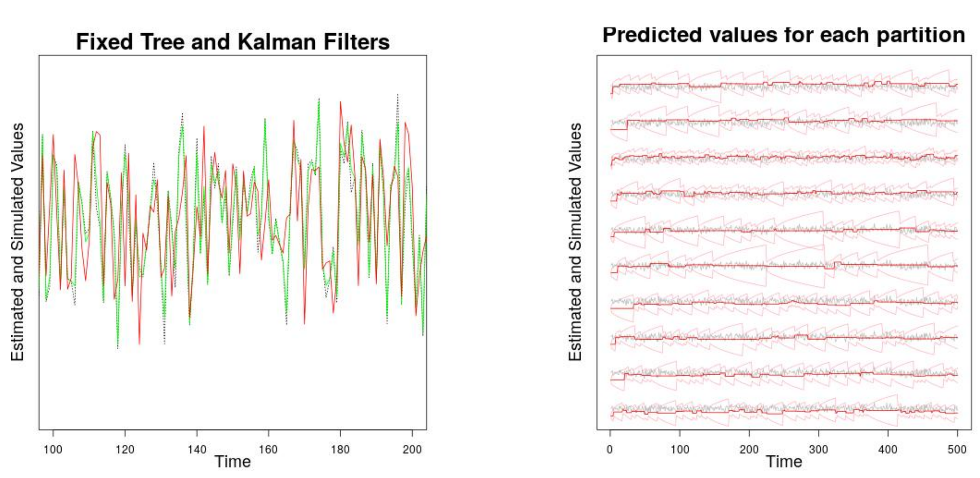

Figure 2: Estimating a process by creating smaller subprocesses.

3. Base Model

The $p(z_{tk} \mid T, \theta_T, x^t, y^t)$ is derived from the Kalman filter and because the Equations in (**??**) are both normally distributed we can derive an exact form of $p(T \mid \theta_T, x^t, y^t)$:

$$p(T)\prod_{k=1}^{K_T}\int p(z_{0,k})\prod_{i=1}^{t} p(y_i \mid z_{i,k}, T)^{\mathrm{I}_{\mathrm{i,k}}}$$
$$\cdot\, p(z_{i,k} \mid z_{i-1,k}, u_t, T)dz_k^t$$
$$= p(T)\prod_{k=1}^{K_T}(|2\pi W_0|)^{-\frac{1}{2}}$$
$$\cdot\left(\prod_{i=1}^{t}(|2\pi W_k||2\pi A_i|)^{-\frac{1}{2}}(|2\pi V_k|)^{\frac{-\mathrm{I}_{\mathrm{i,k}}}{2}}\right)\cdot$$
$$\exp\left[-\frac{1}{2}\left(\mu_{0,k}W_0^{-1}\mu_{0,k} - d_0^T A_i^{-1} d_0 + \right.\right.$$
$$\left.\left.\sum_{i=1}^{t}\mathrm{I}_{\mathrm{i,k}}\, y_{i,k}^T V_k^{-1} y_{i,k} + u_i^T G^T W_k^{-1} G u_i - d_i^T A_i^{-1} d_i\right)\right].$$

The main point to note is that this computation is $O(n)$ because the current term only relies on the previous term.

4. Tree Evolution

Initially we followed a standard MH algorthim which provides local reversible moves as described by Chipman et al. 1998. However, there is degeneracy of this Markov chain because the local tree mutations do not explore the full tree space. To alleviate this we have extended the stochastic search of the tree space by using simulated tempering by Geyer and Thompson 1995 and more complex moves inspired by Wu et al. 2007 and Pratola 2016.

5. Streaming

Exchangeability of responses based on conditional data allows us to to develop a window-like streaming algorithm. If the data is arriving faster than it can be processed, $\lambda_a < \lambda_s$ then we can randomly select inputs within the window size and discard/store those the preceded the selected input.

The size of the window is based on the decisions and resources available to the analyst including, tree size (K_T), rate of data arrival, (λ_s), rate of algorithm (λ_a), choice of model type (state only estimation, dual estimation, parameter learning, variable or model selection).

6. Extensions and More

One of the reasons for using the Kalman Filter is that many extensions exist. In particular, the Unscented Kalman filter allows for nonlinear dynamics to be modelled and is an improvement over the Extended Kalman Filter. The calcuation of the posterior uses the idea that the marginialisation over z_t in the posterior calculation can be approximated by the expectation w.r.t. z_t and $p(T \mid \theta_T, x^t, y^t)$.

There are also methods developed by Mehra 1970 and others that allow us to adapt the algorithm for inference on the variance of the state W_t.

A further adaptation under development is to declare each leaf as a Gaussian marginal so the each tree represents a Gaussian process, i.e. the leaves of the tree represent the components of a vector $\in \mathbb{R}^{|K_T|}$ where each leaf may be $\in \mathbb{R}^m$. In this case we have fixed tree sizes (fixed number of leaves).

7. Simulation Study

The Bayesian Dynamic Regression Tree (BDRT) is compared to the Kalman Filter. A time-series data set was simulated using the Mackey-Glass non-linear time series with a τ of 20. To estimate this series we used:

$$H = \begin{bmatrix}1\\1\end{bmatrix},\quad F = \begin{bmatrix}0.9 & 0\\0 & 0.2\end{bmatrix},\quad W = \begin{bmatrix}0.1 & 0\\0 & 0.1\end{bmatrix},\quad y \in \mathbb{R},\quad z \in \mathbb{R}^2 \quad \text{and} \quad V = 0.03.$$

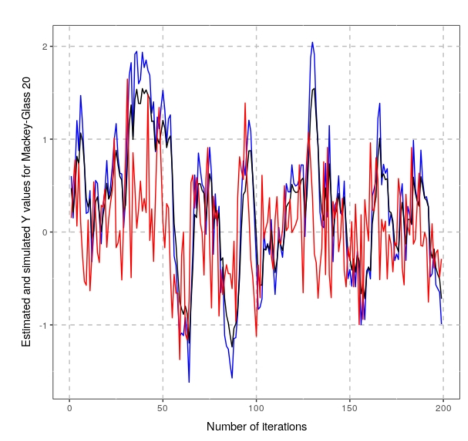

Figure 3: Showing observation predictions with BDRT and the Kalman Filter

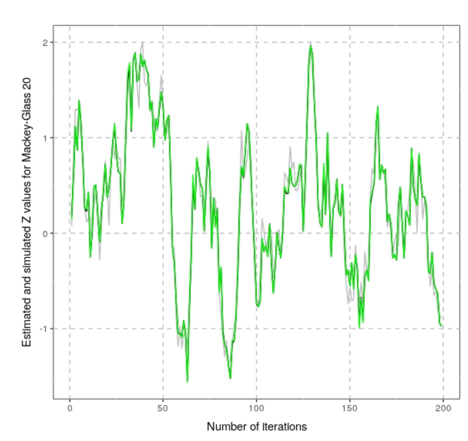

Figure 4: Showing latent state predictions with BDRT and the Kalman Filter

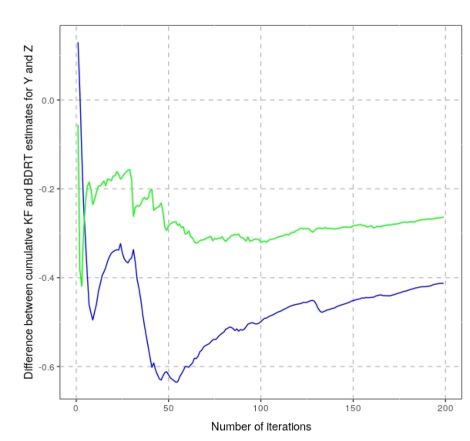

Figure 5: Difference in RMSE between BDRT and the Kalman Filter

Comparing different MCMC methods on BDRT. "MH" is the Chipman et al. method using Grow, Prune, Swap and Change. "BST" uses the same moves but has 10 levels of heating. The pseudoprior is estimated stochastically using the method suggested by Geyer and Thompson 1995. "MST" uses different moves: multigrow, multiprune, multichange, shift, and swap. The abound of growing, changing and pruning is temperature dependent.

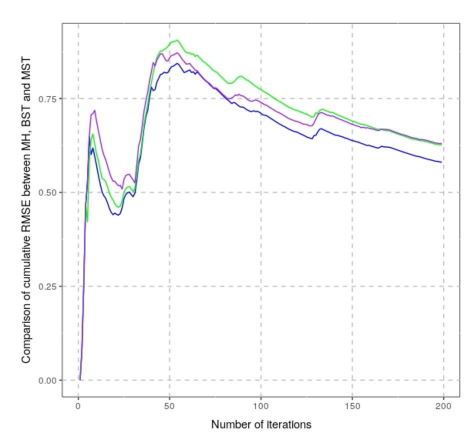

Figure 6: Comparing RMSE between the 3 different MCMC approaches.

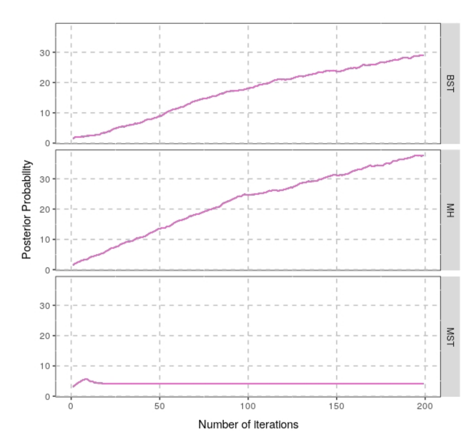

Figure 8: Average tree size as the algorithm progresses

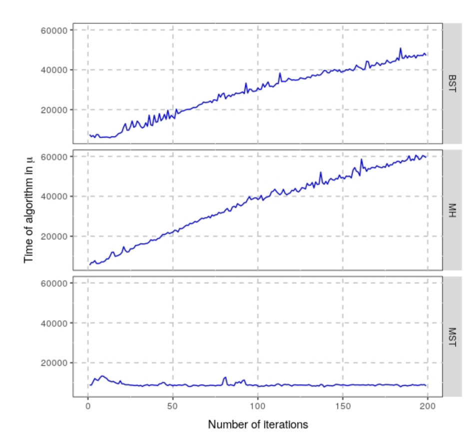

Figure 10: Time comparisons between different MCMC methods.

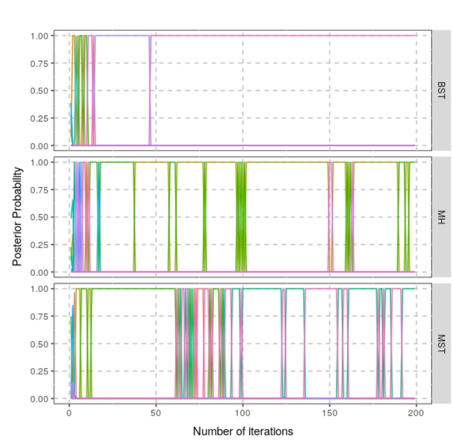

Figure 7: The mixing of the tree probabilities.

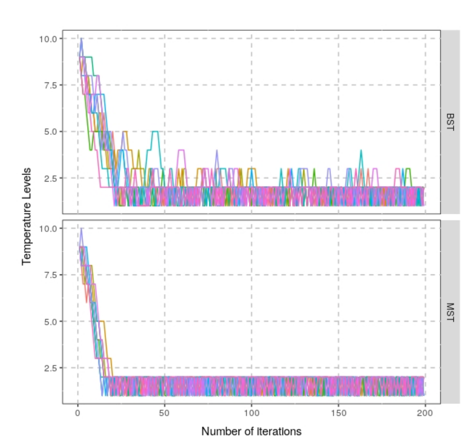

Figure 9: Temperature traversal of the trees.

References

Chipman, H. A., E. I. George, and R. E. McCulloch (1998). "Bayesian CART model search". In: *Journal of the American Statistical Association* 93.443, pp. 935–948.

Chipman, H. A., E. I. George, R. E. McCulloch, et al. (2010). "BART: Bayesian additive regression trees". In: *The Annals of Applied Statistics* 4.1, pp. 266–298.

Geyer, C. J. and E. A. Thompson (1995). "Annealing Markov chain Monte Carlo with applications to ancestral inference". In: *Journal of the American Statistical Association* 90.431, pp. 909–920.

Gramacy, R. B. and H. K. H. Lee (2008). "Bayesian Treed Gaussian Process Models With an Application to Computer Modeling". In: *Journal of the American Statistical Association* 103.483, pp. 1119–1130.

Mehra, R. (Apr. 1970). "On the identification of variances and adaptive Kalman filtering". In: *IEEE Transactions on Automatic Control* 15.2, pp. 175–184. ISSN: 0018-9286.

Meinhold, R. J. and N. D. Singpurwalla (1983). "Understanding the Kalman filter". In: *The American Statistician* 37.2, pp. 123–127.

Pratola, M. T. (Sept. 2016). "Efficient Metropolis-Hastings Proposal Mechanisms for Bayesian Regression Tree Models". In: *Bayesian Anal.* 11.3, pp. 885–911.

Sinopoli, B., L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, and S. S. Sastry (2004). "Kalman filtering with intermittent observations". In: *IEEE transactions on Automatic Control* 49.9, pp. 1453–1464.

Taddy, M. A., R. B. Gramacy, and N. G. Polson (2011). "Dynamic trees for learning and design". In: *Journal of the American Statistical Association* 106.493, pp. 109–123.

Wu, Y., H. Tjelmeland, and M. West (2007). "Bayesian CART: Prior Specification and Posterior Simulation". In: *Journal of Computational and Graphical Statistics* 16.1, pp. 44–66.