Bayesian Analysis (2016) **11**, Number 3, pp. 921–927

Comment on Article by Pratola[*,†]

Christopher M. Hans[‡]

1 Overview

Pratola addresses a specific and challenging problem: the construction of Metropolis–Hastings (MH) proposal mechanisms for regression tree models that are both *efficient* and *effective*. *Efficiency* in this context relates to per-iteration computation time, which is desired to be kept to a minimum. *Effectiveness* in this context relates to the mixing of the resulting chain and its ability to avoid becoming trapped in local modes. As is typical when designing Markov chain Monte Carlo (MCMC) algorithms, these desiderata must be balanced against each other. Moves that are computationally efficient often result in slow-mixing chains, while moves that result in fast-mixing chains—if mechanisms for proposing such moves can even be found—are often accompanied by a high computational burden. Balancing these desiderata is quite difficult, both in general and in the particular case of Bayesian regression tree models.

Pratola's approach to improving MCMC efficiency and effectiveness in the Bayesian regression tree model setting draws on two existing and commonly-used MH moves. Pratola generalizes these moves to be more aggressive in exploring the posterior without sacrificing much computational efficiency. The first is a move based on a rotation mechanism (see, e.g., Sleator et al., 1988) used by Gramacy and Lee (2008) in the sampling of Bayesian treed Gaussian process models. The generalized rotate proposal developed by Pratola allows for nontrivial changes to be made to the interior structure of the tree. Critically, the nodes involved in the rotation need not all split on the same variable, as was the case in earlier implementations. The second is a move that changes a cutpoint and/or splitting variable that has been implemented in various ways in the literature (e.g., Chipman et al., 1998; Dennison et al., 1998; Chipman et al., 2002; Wu et al., 2007; Gramacy and Lee, 2008; Chipman et al., 2010). The generalized perturb-within-change-of-variable move allows for flexibility in moving the cutpoint for a given split while using the covariates to inform the proposal distribution. The resulting generalized moves are *computationally local* yet *structurally global* in that they allow the chain to avoid becoming trapped in local modes while restricting computation at any given iteration to localized regions of the tree.

The goal of balancing computational speed with algorithmic effectiveness arises in many computational settings. When thinking about such a balance, I am often reminded of van Dyk and Meng (1997), who describe a search for "a 'free and better lunch,' not a 'better but expensive lunch' in terms of human and computational effort" in the context of designing ECM (Meng and Rubin, 1993) and ECME (Liu and Rubin, 1994)

[*]Main article DOI: 10.1214/16-BA999.

[†]This work was supported by the U.S. National Science Foundation under award number DMS-1310294.

[‡]Department of Statistics, The Ohio State University, Columbus, Ohio, U.S.A., hans@stat.osu.edu

 DOI: 10.1214/16-BA999B

optimization algorithms. Such a free lunch is indeed sometimes available. In their study of maximum likelihood estimation for the multivariate t distribution via the EM algorithm (Dempster et al., 1977), Meng and van Dyk (1997) describe how a minor change to the "standard" data augmentation results in an algorithm with an optimal rate of convergence. Moreover, the speedup can be obtained by changing only one portion of a single line of code implementing the "standard" algorithm.

While ideal, such an economical improvement is the exception rather than the rule when it comes to algorithm design. The new MH mechanisms Pratola introduces represent a step toward this ideal as measured by empirical comparisons to existing approaches. As demonstrated in Section 2.3 (and elsewhere) of Pratola's paper, the new MH moves allow the chain to mix across modes much more freely than under existing approaches. As demonstrated in Section 5.2 of the paper, the new MH moves can result in a larger effective sample size per second. The lunch, however, is only "free" to the user because Pratola pays for part of the coding bill himself by providing details for the algorithmic heavy-lifting in the paper. As evidenced in Section 3 of the paper, defining the rotation operator, $\mathcal{R}$, and deriving the MH acceptance probability are nontrivial exercises and represent important contributions of the work. Equipped with these results, Pratola provides pseudocode in the Supplementary Material that can incorporated into code implementing MCMC algorithms for a variety of Bayesian regression tree models. The result is a fairly straightforward "add-on" to existing algorithms that can improve mixing without sacrificing too much computational efficiency.

2 Connections to Literature on Related Problems

Large, structured model spaces pose many computational challenges, especially for posterior inference via MCMC. The regression tree models considered by Pratola are particularly challenging, however similar difficulties arise in the related areas of Gaussian graphical model determination and Bayesian regression modeling with many predictors. Basic algorithms that are only capable of making local moves tend, in all three model settings, to become trapped in or near local modes, especially as the size of the problem increases. Those who attended the 2016 ISBA World Meeting in Sardinia, Italy—where Pratola's paper was presented with discussion—may have attended Professor Peter Green's Foundations Lecture on "Graphical modeling and Bayesian structural learning" (Green, 2016). In his lecture, Prof. Green provided a detailed survey of approaches to modeling and computation for Gaussian graphical models (and, to a lesser extent, Bayesian linear regression models and models on trees). Those who were not able to attend the lecture may view it online.[1]

It is clear from both a survey of the literature and from Prof. Green's lecture that, in all three related model settings, the historical trend has been a shift from algorithms that rely on local moves to those that encourage global moves. For example, early

[1]A video of the lecture and the corresponding slides from the presentation will soon be available at http://videolectures.net/isba2016_green_graphical_modelling/; see also links at http://www.bayesian.org.

MCMC algorithms for decomposable Gaussian graphical models typically involved proposals where a single edge was added to, deleted from, or swapped out of the current model (e.g., Madigan and York, 1995; Guidici and Green, 1999; Armstrong et al., 2009). Computationally-aggressive approaches like the Shotgun Stochastic Search approach of Jones et al. (2005) relied on an extensive survey of the local neighborhood of the current model when constructing moves. More recent algorithms, such as the feature-inclusion stochastic search approach of Scott and Carvalho (2012), have moved toward mixing local moves with global moves in an attempt to escape local posterior modes.

A similar trend has occurred in the literature related to Bayesian regression model uncertainty and variable selection. Early algorithms focused on local moves that involved some combination of adding, deleting or swapping variables (e.g., George and McCulloch, 1993; Geweke, 1996; Smith and Kohn, 1996; George and McCulloch, 1997). Shotgun Stochastic Search approaches (Hans et al., 2007) again relied on an extensive cataloguing of local neighborhoods. As with Gaussian graphical models, recent algorithms for exploring regression model space have moved toward using a mixture of local and carefully-constructed global moves (e.g., Bottolo and Richardson, 2010). The global moves often represent a non-local change of dimension. For example, Xu (2011) describes an approach where, after a transformation on the parameter space, local moves in the transformed regression space correspond to "larger" moves in the original model space, allowing the chain to make global moves. Finally, notions of adaptation are now being used to guide searches for high posterior probability regression models (Nott and Kohn, 2005; Clyde et al., 2011).

Pratola's paper fits nicely into the stream of related literature, as it focuses on a move away from purely local (and low-dimensional) moves while at the same time incorporating notions of adaptiveness and pre-conditioning based on information in the predictors, all while maintaining computational efficiency.

3 Specifics

There are many ways in which the two new MH proposals could be incorporated into an MCMC algorithm. The paper explores a few potential options and provides guidance that is based on experimental evidence and intuition the author has gained by studying these problems in detail. After reading the manuscript and considering how the work fits into the broader literature on MCMC for related problems, I wondered about a few other specific approaches to implementation and how these approaches might fare relative to the ones presented in the paper.

First, as evidenced in the literature on MCMC for structural identification in the areas described above, the role of adaptation has been seen to be increasingly useful. This is particularly true when the goal is to tune proposal distributions that have the ability to make global (or, at least, non-local) moves. Designing such moves often requires knowledge about the posterior so as to avoid proposing global moves to low-probability regions. This knowledge can be accumulated on-the-fly via adaptation. Pratola discusses the potential for adaptation with respect to the α scaling parameter involved in the perturbation proposal, as well as with respect to the percentage of iterations during

which a rotate proposal is used in place of the traditional birth/death proposal. Rather than implementing a formal adaptive MCMC for α, Pratola opts to adapt α during a pre-burn-in run of the chain. Under this approach, α is updated on a regular schedule to tune the locality of the proposed moves by monitoring the acceptance rate of the chain. At the end of the adaptation period, α is fixed and the formal MCMC is started. This is reasonable, and it appears to work well in practice.

Thinking about other potential ways in which adaptation could be incorporated in to the algorithm, I wonder whether adapting the change-of-variable move might help increase the efficiency of the MCMC. Pratola's change-of-variable move relies on a pre-conditioning strategy whereby correlations between the predictor variables are used to inform the proposal. Pratola argues that this strategy helps the algorithm converge to the true posterior when there are highly correlated predictors, providing a more honest quantification of posterior uncertainty than is obtained by the standard algorithms. I wonder whether, rather than fixing this proposal distribution at the start of the chain via pre-conditioning, the proposal distribution might be adapted as the chain evolves, resulting in a post-conditioning of the change-of-variables move. By adaptively learning which changes-of-variable result in good moves throughout the space, the chain may be able to make more effective moves than can be obtained by pre-conditioning alone. Such an approach is reminiscent of the adaptive sampler proposed by Nott and Kohn (2005) in the context of Bayesian variable selection for the normal linear model. Exploring adaptations along these lines might shed more light on the ways in which correlations between predictors drive poor mixing of standard algorithms.

My second thought about alternative approaches also relates to the strategy for the change-of-variables update. Under the pre-conditioning approach, Pratola notes that "if v_k is highly correlated with a single other variable v_j, then this formula [the pre-conditioning approach] will lead to proposals that stay at v_k about 50% of the time and propose transitions to v_j about 50% of the time." I wonder whether restructuring the proposal so that, in situations like this, the mechanism is heavily biased toward proposals *away* from v_k might lead to a more effective sampler, as it would avoid proposing a "move" to the current variable. This is motivated by the "Metropolized Gibbs sampler" of Liu (1996), which can be shown in certain setups to yield more efficient inference than samplers that allow proposals to the current state. Of course, such an approach would only be practically useful in the regression tree setting if improvements in mixing outweighed potential increases in computational cost due to the restructuring of the tree/model that would occur after a proposed change-of-variable.

Finally, the paper proposes two different types of improved moves, the rotate proposal and the perturbation(-within-change-of-variables) proposal. In Section 5, Pratola investigates the impact of these two proposals separately and simultaneously. It appears from the various examples that both proposal mechanisms help increase the effectiveness of the sampler, and that there is an interaction effect between the proposals in terms of both the diversity of tree structures explored and the resulting computation time. Further investigations along these lines to help understand the relative importance of the two proposals and the way in which they interact would be of interest, and may reveal other interesting properties of the posterior distribution over tree structures.

4 Summary

To summarize, Pratola has tackled a difficult problem that also arises in other discrete-structure setups. He has proposed improvements to existing MH methods that are customized to account for the topological structure of regression trees. The new methods, the development of which required a careful construction of a nontrivial operation on a tree structure, provide mechanisms for proposing global jumps in regression tree space that require only localized computation.

I would like to thank Pratola for writing a thoughtful paper that I very much enjoyed reading. As someone who does not work with tree structures on a regular basis, I appreciated that the paper was written with enough detail and clarity to enable a casual reader to extract the essence of the problem while at the same time tracking the details of the solution. I found the work to be both useful and interesting, and expect the methodology will have a positive impact on modeling practice in this area.

References

Armstrong, H., Carter, C. K., Wong, K. F. K., and Kohn, R. (2009). "Bayesian Covariance Matrix Estimation Using a Mixture of Decomposable Graphical Models." *Statistics and Computing*, 19: 303–316. MR2516221. doi: http://dx.doi.org/10.1007/s11222-008-9093-8. 923

Bottolo, L. and Richardson, S. (2010). "Evolutionary Stochastic Search for Bayesian Model Exploration." *Bayesian Analysis*, 5: 583–618. MR2719668. doi: http://dx.doi.org/10.1214/10-BA523. 923

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998). "Bayesian CART Model Search." *Journal of the American Statistical Association*, 93: 935–960. MR1631325. doi: http://dx.doi.org/10.2307/2670105. 921

Chipman, H. A., George, E. I., and McCulloch, R. E. (2002). "Bayesian Treed Models." *Machine Learning*, 48: 299–320. 921

Chipman, H. A., George, E. I., and McCulloch, R. E. (2010). "BART: Bayesian Additive Regression Trees." *The Annals of Applied Statistics*, 4: 266–298. MR2758172. doi: http://dx.doi.org/10.1214/09-AOAS285. 921

Clyde, M. A., Ghosh, J., and Littman, M. L. (2011). "Bayesian Adaptive Sampling for Variable Selection and Model Averaging." *Journal of Computational and Graphical Statistics*, 20: 80–101. MR2816539. doi: http://dx.doi.org/10.1198/jcgs.2010.09049. 923

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data via the EM Algorithm (with discussion)." *Journal of the Royal Statistical Society – Series B*, 39: 1–38. MR0501537. 922

Dennison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). "A Bayesian CART Algorithm." *Biometrika*, 85: 363–377. MR1649118. doi: http://dx.doi.org/10.1093/biomet/85.2.363. 921

George, E. I. and McCulloch, R. E. (1993). "Variable Selection via Gibbs Sampling." *Journal of the American Statistical Association*, 88: 881–889. 923

George, E. I. and McCulloch, R. E. (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7: 339–373. 923

Geweke, J. (1996). "Variable Selection and Model Comparison in Regression." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 5*, 609–620. Oxford University Press. MR1425430. 923

Gramacy, R. B. and Lee, H. K. H. (2008). "Bayesian Treed Gaussian Process Models with an Application to Computer Modeling." *Journal of the American Statistical Association*, 103: 1119–1130. MR2528830. doi: http://dx.doi.org/10.1198/016214508000000689. 921

Green, P. J. (2016). "Graphical Modeling and Bayesian Structural Learning." ISBA World Meeting, Sardinia, Italy. Foundations Lecture. http://videolectures.net/isba2016_green_graphical_modelling/. 922

Guidici, P. and Green, P. J. (1999). "Decomposable Graphical Gaussian Model Determination." *Biometrika*, 86: 785–801. MR1741977. doi: http://dx.doi.org/10.1093/biomet/86.4.785. 923

Hans, C., Dobra, A., and West, M. (2007). "Shotgun Stochastic Search for "Large p" Regression." *Journal of the American Statistical Association*, 102: 507–516. MR2370849. doi: http://dx.doi.org/10.1198/016214507000000121. 923

Jones, B., Carvalho, C., Dobra, A., Hans, C., Carter, C., and West, M. (2005). "Experiments in Stochastic Computation for High-Dimensional Graphical Models." *Statistical Science*, 20: 388–400. MR2210226. doi: http://dx.doi.org/10.1214/088342305000000304. 923

Liu, C. and Rubin, D. B. (1994). "The ECME Algorithm: A Simple Extension of ECM With Fast Monotone Convergence." *Biometrika*, 81: 633–648. MR1326414. doi: http://dx.doi.org/10.1093/biomet/81.4.633. 921

Liu, J. S. (1996). "Peskun's Theorem and a Modified Discrete-State Gibbs Sampler." *Biometrika*, 83: 681–682. MR1423883. doi: http://dx.doi.org/10.1093/biomet/83.3.681. 924

Madigan, D. and York, J. (1995). "Bayesian Graphical Models for Discrete Data." *International Statistical Review*, 63: 215–232. 923

Meng, X.-L. and Rubin, D. B. (1993). "Maximum Likelihood Estimation via the ECM Algorithm: A General Framework." *Biometrika*, 80: 267–278. MR1243503. doi: http://dx.doi.org/10.1093/biomet/80.2.267. 921

Meng, X.-L. and van Dyk, D. (1997). "The EM Algorithm—an Old Folk-song Sung to a Fast New Tune." *Journal of the Royal Statistical Society – Series B*, 59: 511–567. MR1452025. doi: http://dx.doi.org/10.1111/1467-9868.00082. 922

Nott, D. J. and Kohn, R. (2005). "Adaptive Sampling for Bayesian Variable Selection." *Biometrika*, 92: 747–763. MR2234183. doi: http://dx.doi.org/10.1093/biomet/92.4.747. 923, 924

Scott, J. G. and Carvalho, C. M. (2012). "Feature-Inclusion Stochastic Search for Gaussian Graphical Models." *Journal of Computational and Graphical Statistics*, 17: 790–808. MR2649067. doi: http://dx.doi.org/10.1198/106186008X382683. 923

Sleator, D. D., Tarjan, R. E., and Thurston, W. P. (1988). "Rotation Distance, Triangulations, and Hyperbolic Geometry." *Journal of the American Mathematical Society*, 3: 647–681. MR0928904. doi: http://dx.doi.org/10.2307/1990951. 921

Smith, M. and Kohn, R. (1996). "Nonparametric Regression Using Bayesian Variable Selection." *Journal of Econometrics*, 75: 317–343. 923

van Dyk, D. A. and Meng, X.-L. (1997). "On the Orderings and Groupings of Conditional Maximizations within ECM-Type Algorithms." *Journal of Computational and Graphical Statistics*, 6: 202–223. MR1466588. doi: http://dx.doi.org/10.2307/1390931. 921

Wu, Y., Tjelmeland, H., and West, M. (2007). "Bayesian CART: Prior Specification and Posterior Simulation." *Journal of Computational and Graphical Statistics*, 16: 44–66. MR2345747. doi: http://dx.doi.org/10.1198/106186007X180426. 921

Xu, R. (2011). "Regression Model Stochastic Search via Local Orthogonalization." Ph.D. thesis, The Ohio State University. MR2995972. 923