Discussion of "Spline Adaptation in Extended Linear Models", by Hansen and Kooperberg

Hugh A. Chipman, Edward I. George, Robert E. McCulloch

December 5, 2001

This paper uses ideas for stochastic search implementations of adaptive Bayesian models, such as those outlined in Denison, Malick and Smith (1998 a,b) and Chipman, George and McCulloch (1998a) and effectively applies these ideas to logspline density estimation and triogram regression. Interesting comparisons are made to assess the effect of greedy search, stochastic search and model averaging. Such comparisons are valuable, since readily available computing power enables the construction of many methods, and an understanding of what works is important in developing new methodology.

It is very important to note the role of the prior when adaptive models are used in conjunction with stochastic searches. Inevitably, priors guide and temper our wandering in a large space of models. This benefit comes with a price: the need to select a prior that is appropriate for the problem at hand. It is important to acknowledge the simple fact that a prior choice represents a bet on what kind of models we want to consider.

If we skip to the end of the paper and read the discussion, what lessons have been learned? We have (i) ".. we have demonstrated a gain .. when appealing to the more elaborate sampling schemes" (relative to simple greedy search), and (ii) "priors play an important role". These things we know to be true in general from much experience. The question is: what should be done in practice?

In general, a practical approach usually involves first getting the prior specification down to a few hyper-parameters (about which we hopefully have some understanding) and then developing a scheme for making reasonable choices. At one end of the spectrum we can use automatic methods such as cross-validation to choose hyper-parameters that are appropriate for the problem at hand. At the other end of the spectrum we choose "reasonable values" based on our understanding and prior beliefs. Often, compromise strategies that combine a peek at the data with some judgment are effective and somewhat in the spirit of empirical Bayes. We believe Chipman, George, and McCulloch (2002) is a good example of this middle ground approach.

We have some general Bayesian insights that help us understand the effects of these hyper-parameters. Often we can think of prior in two stages: $p(M_k)$ a prior on "models", and $p(\theta_k|M_k)$ a prior on the parameters of a given model. A set of hyper-parameters would specify a choice for each of these components.

In section 2 of the paper, θ corresponds to the coefficients β, and M_k would be (K, t). Both choices can be important. Often we choose $p(M_k)$ to express the belief that the model is not too large. More subtle is the effect of a choice $p(\theta|M_k)$. If we make the prior too tight we will miss parameter values that give good fit to the data, diminishing the posterior probability on model M_k. If we make the prior too spread out, the likelihood will be washed out and again we diminish the posterior probability. These are the basic facts of odds ratio calculations.

In section 2 of the paper, the choices of A and λ are the hyper-parameters the determine the spread of the prior given the model. We know from the general insight outlined above that these choices will be influential. The paper discusses these choices in terms of penalties and the AIC. We find the basic Bayesian intuition about odds ratio calculations is also helpful in understanding what is going on. It may be helpful to recall that the AIC is just a (very poor) approximation to the odds ratio calculation.

Table 2 compares the performance of algorithms for various values of λ. We see that the choice of λ matters. What choice is best? It depends. Based on table 2, the authors state that choice (vii) is bad, yet it is best in several scenarios! The question remains: how do you choose λ?

While the authors consider the impact of different prior choices (e.g., for λ), methods for selection of the prior are not considered. Without such choices, the use of MCMC technology as a stochastic search by non-Bayesians is more limited.

One of the most important advantages of Bayesian methods in adaptive modeling problems is the effectiveness of stochastic search methods such as MCMC. In applications where the model space is complicated, constructing an effective chain can be challenging. For example, in the triogram regression problem, models are arranged somewhat hierarchically, with regions recursively subdivided into smaller and smaller triangles. Hierarchical structure makes the construction of an effective chain challenging because it constrains the possible set of proposals that can be made. Proposals making small local changes are easiest to make and most likely to be accepted, but a long succession of simple proposals may need to be accepted for the stochastic search to move on to a different posterior mode. With this dilemma in mind, we appreciate the importance of using good proposal steps in effective exploration of the model space. These transitions need to work within the model constraints (e.g. hierarchy in triograms) while not being so constrained as to have difficulty moving. Hansen and Kooperberg have effectively accomplished this by developing a set of proposal steps which move around the space in a natural way while respecting the nested nature of the models. In some problems, such as logspline density estimation, it may be easier to move around the space. In that case, the knots don't depend on the order in which they are added.

The authors use a single long chain to explore the model space, which can be an issue if the posterior on models has many sharp local peaks. In such situations, MCMC methods can tend to gravitate towards a single mode and have difficulty in moving to other regions of the model space. We expect such

issues to arise in the triogram regression problem, for example. Denison et. al. (1998b) use single chains as well, and by carefully controlling the early stages of the chain, achieve an algorithm which seems to explore a region of the model space around a single local maximum. We have found that another effective technique is to use multiple chains as a means of more fully exploring the space. Single and multiple chains were explored on a simulated dataset in Chipman, George and McCulloch (1998a) and the use of multiple chains resulted in a more complete exploration of the model space.

The authors examine the performance of Bayesian model averaging, which is an appealing and natural means of improving predictive accuracy. We are not surprised that greedy methods can be improved upon by a better search and model averaging. What does surprise us is the omission of a trivial (and often effective) frequentist competitor: Bootstrapping. The bootstrap has been used as a method of generating multiple models for model averaging (Breiman 1996), and as an easy way to improve upon greedy search algorithms (Tibshirani and Knight 1999). In this approach, multiple pseudo datasets are generated by resampling with replacement the rows of the data matrix, and a (often greedy) modeling algorithm is applied to each bootstrap dataset. Bootstrapping the data and averaging over models is an effective and easy way to model average. It enhances the search by perturbing the data and letting the greedy algorithm converge to different local maxima. Predictions are improved by averaging across all the different models. We have carried out some experiments with bootstrapping in the context of Bayesian CART (Chipman, George and McCulloch 1998b). In the example we considered, we found that bootstrapping identified a wider variety of good models than a single greedy search, but the models identified by a bootstrap algorithm were still a subset of those identified by Bayesian stochastic search procedures.

References

Breiman, L (1996), Bagging Predictors, *Machine Learning*, 26, 123–140.

Chipman, H. A., George, E. I., and McCulloch, R. E. (1998a) Bayesian CART Model Search (with discussion), *Journal of the American Statistical Association*, 93, 935–960.

Chipman, H. A., George, E. I. and McCulloch, R. E. (1998b) Making Sense of a Forest of Trees, *Proceedings of the 30th Symposium on the Interface*, S. Weisberg, Ed., Interface Foundation of North America. p 84. – 92.

Chipman, H. A., George, E. I, and McCulloch, R. E. (2002) Bayesian Treed Models, to appear, *Machine Learning*.

Tibshirani, R. and Knight, K (1999) Model Search by Bootstrap "Bumping", *Journal of Computational and Graphical Statistics*, 8, 671–686.