Comments on 'Bayesian Backfitting' by Hastie and Tibshirani

Peter J. Green (University of Bristol)

I warmly congratulate the authors on this paper. I am sure they will succeed in broadening acceptance of the Bayesian paradigm in inference in regression, by providing this well-written and accessible treatment of the use of Markov chain Monte Carlo (MCMC) in fitting the important class of (generalised) additive models.

The paper promotes several ideas. It can be interesting and revealing to examine Bayesian analogues of familiar frequentist models and procedures; MCMC is important in Bayesian inference; Gibbs sampling is a convenient general recipe for MCMC; if that isn't available, try Metropolis-Hastings; Gibbs sampling is a close analogue of backfitting. None of these points are individually very original, of course! But it is very appealing to see their combination applied to additive models, with a number of practical details worked out to produce an efficient methodology, especially since **Splus** software to implement the resulting methodology is provided.

My comments focus on some of these practical details, and some other relations and connections to the proposed methods.

Bayesian function estimation

The reader who comes to this work from a background of backfitting in (generalised) additive models rather than experience of inference about functions and surfaces in the Bayesian paradigm might get an impression that this paper is close to the start-of-the-art in Bayesian function estimation. In fact, the authors do not claim this; researchers have been investigating and using Bayesian models and MCMC calculations for more complicated situations than this, almost from the earliest days of MCMC in statistics. One could even claim that a driving force in the broader acceptance of practical Bayesian methodology for inference in complex data structures has been that using MCMC methods there was a comparatively trivial computational penalty to be paid in moving on from simple models to more complicated ones (in the case of inference about functions, for example, replacing Gaussian priors on functions by non-Gaussian ones).

MCMC in statistics is generally accepted to have begun in statistical image analysis, and of course nonlinear smoothers, that do not destroy boundaries between objects in the scene, are routinely needed there. In discrete spatial settings, such as arise in pixellised images and region-based geographical and ecological problems, particular use has been made of pairwise-difference priors that are not Gaussian, but have heavier tails, for example

$$p(f) \propto \exp\{-\beta \sum_{i \sim j} \phi(f_i - f_j)\}$$

where f_i is the true image intensity in pixel i, and $i \sim j$ means the summation is over pairs of neighbouring pixels (see Geman and McClure, 1985; Green, 1990; Besag, et al., 1995 among many others). Instead of using $\phi(u) = u^2$, taking it to be $|u|$, or something more complicated such as $\log \cosh u$ or $-1/(1+u^2)$ has proved successful in many image restoration contexts. (The apparent connection here with M-estimation and robustness is not entirely superficial.)

These methods are all for *discrete* spatial problems, whether on lattices or irregular graphs. Nonparametric Bayesian surface fitting methods for *continuous* space include that of Heikkinen and Arjas (1998) for inference on a Poisson intensity, using ideas of model-averaging over appropriately-defined step functions to yield smooth posterior mean surfaces, and the variogram-based methods of Diggle, Tawn and Moyeed (1998). Turning to regression on more general, non-spatial, covariates, apart from the work of Denison, et al (1998) and Holmes and Mallick (1997) that is mentioned in this paper, and other methods investigated by these researchers and colleagues, there is the interesting approach of Müller, et al (1996), based on Dirichlet process priors.

Bayes from conviction or convenience?

The conceptual connection between smoothing methods, especially those based on penalised likelihood, and Bayesian formulations of inference about functions is often made, but there always seems to be an implicit or explicit warning: "don't take this too literally". Formal use of such connections is rarely made, a notable exception being Wahba's important 1983 paper, exploiting the Bayesian connection to construct confidence intervals about spline estimates. However, the frequentist properties of such intervals are well-known to be problematical.

This is all well-understood by the authors, but I do think that if they seek to use the Bayesian paradigm – and presenting credible intervals for functions is certainly doing that – they should try to take it a bit more seriously! There are several issues here.

The first concerns the use of the roughness penalty as a negative-log-prior on the regression function. Does the usual penalty $\lambda \int [f''(x)]^2 dx$ have particular merit in this context, over other (say) quadratic forms in f that are zero for constant or first-order f, or is the choice driven, as in straightforward smoothing, by the computational advantages? Specifically, do autocorrelations decline with lag in a reasonable way, and what about homogeneity of variance? (Answers to these questions are complicated by the partially improper nature of the prior in this case). Are Gaussian prior assumptions reasonable, or should we consider heavy-tailed modifications? Choice of functional form of the penalty has far greater consequences for the Bayesian procedure than for the simple smoother, as we are going to use it to generate much more subtle inferences, invested, presumably, with real probabilistic interpretations. To be fair, the authors are in good company in not raising these questions – they are almost never considered by anybody else either!

The second issue is the treatment of smoothing (tuning) parameters (or, equivalently, variance components). This is explicitly discussed, in Section 4 of the paper. I must say that I find the full Bayesian version much more compelling than either of the alternatives.

A third concern is about sensitivity to prior assumptions, always problematic in hierarchical models. It would be good to see at least an empirical study of the effect on posterior inference of variations in the authors' assumptions. I would anticipate that everything about the inference except the posterior means is actually rather sensitive.

The root-S approximation

I cannot be the only reader to worry about the quality of the approximation to $S^{1/2}\boldsymbol{z}$ proposed in Algorithm A.1. The approach looks simplistic – why should the size of the first neglected term of the series be a reliable guide to the size of the sum of *all* neglected terms? – and indeed a small numerical experiment bears out this concern.

Although an alternative, superior, approach using Cholesky decomposition is available for the cubic spline smoother, this case provides a convenient choice for a numerical check. Taking $x_i = i, i = 1, 2, \ldots, n = 100$ and drawing a single vector $\boldsymbol{z} \sim N(0, I)$, I compared $||S^{1/2}\boldsymbol{z} - S^{1/2}_{\mathrm{HT}}\boldsymbol{z}||$ with the tolerance on $||S\boldsymbol{z}''||$ used in the authors' algorithm, where $S^{1/2}_{\mathrm{HT}}$ represents their approximation. Table 1 shows that, especially if higher precision is sought, the method becomes both expensive and much less successful.

Table 1: Relationship between tolerance on first neglected term, number of terms included, and overall precision for algorithm A.1 ($||\cdot||$ is the sup norm).

$df = 6$, $\lambda = 2493.2$			$df = 12$, $\lambda = 105.55$		
$\|\|S\boldsymbol{z}''\|\|$	final b	$\|\|S^{1/2}\boldsymbol{z} - S^{1/2}_{\mathrm{HT}}\boldsymbol{z}\|\|$	$\|\|S\boldsymbol{z}''\|\|$	final b	$\|\|S^{1/2}\boldsymbol{z} - S^{1/2}_{\mathrm{HT}}\boldsymbol{z}\|\|$
10^{-1}	2	0.100	10^{-1}	2	0.139
10^{-2}	5	0.073	10^{-2}	6	0.105
10^{-3}	25	0.032	10^{-3}	25	0.069
10^{-4}	95	0.024	10^{-4}	175	0.042
10^{-5}	620	0.015	10^{-5}	847	0.042
10^{-6}	2875	0.010	10^{-6}	2702	0.043

Are better approximations available for an amount of work comparable, say, to Algorithm A.1 with a 10^{-2} tolerance? If not, I wonder if there is merit in turning the approach around, and taking $S^{1/2}$ to be some convenient (symmetric) linear smoother, and defining S to be its square, that is, the smoother obtained by applying $S^{1/2}$ twice? This obviously changes the prior covariance structure so we would have to re-visit that question. However, there is a clear potential for saving computational effort.

To a rough degree of approximation, there can be surprisingly little change – for example, with $x_i = i, i = 1, 2, \ldots, n = 100$ again, S_{10}^2 is close to $S_{8.07}$ where S_{df} is the cubic spline smoother with *df* degrees of freedom – the eigenvalues differ by a maximum of about 0.05, and their outputs are visually nearly indistinguishable.

Some MCMC details

There are several ideas on improving MCMC performance in the literature that could be beneficial in the present context. For example, the modifications to backfitting introduced at the end of Appendix A appear to be related to the ideas of 'hierarchical centering' in normal linear mixed models, of Gelfand, et al. (1995).

Organising variables in a Gibbs sampler into blocks to be simultaneously updated is a common strategy, and often pays off if the blocked variables are highly correlated and the multi-variable update is not expensive to implement. The authors' backfitting strategy is precisely an example of this idea. The variables $(f_j(x_{ij}))$ in the j^{th} block will indeed be strongly correlated. Many questions of MCMC strategy about blocking, updating schedules and re-parameterisation, precisely for the present case of multivariate Gaussian models, are discussed by Roberts and Sahu (1997), which is strongly recommended reading!

A posteriori, there are of course also strong correlations between some variables in different blocks, especially if the predictors are highly correlated themselves, and the anticipated impact of this of MCMC convergence provides another explanation for poor performance in this case.

In other contexts, the correlation between variance parameters and their associated sums-of-squares has proved damaging for MCMC performance; a commonly successful work-around has been to integrate out the variance parameter (assuming we are in the usual conjugate setting with normal random effects and inverse gamma hyperpriors), and then update the random effects by Metropolis-Hastings targetted at a t density. This approach may be useful if the sampler based on equation (24), for example, should mix slowly in a particular case.

Regarding the approach to generalised additive models in section 6, I have also been a great enthusiast (in other contexts) for Metropolis-Hastings based on a Gaussian approximation to the full conditional. People tell me this is not a free lunch, however. Whether the resulting chain is even geometrically ergodic is, I think, not fully understood. In similar problems, this depends on the relative size of the tails of the proposal and target densities (see, for example, Roberts and Tweedie, 1996), and so geometric ergodicity may be problematical in skew cases such as GAM's. Can the authors reassure me – is the sampler provably good?

Finally, brief mention should be made of the required length of MCMC runs. The authors are rather silent on this; a passing reference in the Discussion seems to suggest that only 100 sweeps were made, after burn-in, in one example. This seems very few for such a large-dimensional problem, especially if full posterior distributions are being estimated, not just their means. In all their emphasis on order-n computation, the authors do not allow for any increase in MCMC convergence times with n.

References

Besag, J., Green, P. J., Higdon, D. and Mengersen, K. (1995) Bayesian computation and stochastic systems (with discussion). *Statistical Science*, **10**, 3–66.

Diggle, P. J., Tawn, J. A. and Moyeed, R. A. (1998) Model-based geostatistics (with Discussion). Applied Statistics, **47**, 299–350.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, **82**, 479–488.

Geman, S. and McClure, D. E. (1985) Bayesian image analysis: an application to single photon emission tomography. In *Proceedings of the Statistical Computing Section* 12–18. Amer. Statist. Assoc., Alexandria, VA.

Green, P. J. (1990) Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging*, **9**, 84–93.

Heikkinen, J. and Arjas, E. (1998) Nonparametric Bayesian estimation of a spatial Poisson intensity. *Scandinavian Journal of Statistics*, **25**, 435–450.

Müller, P., Erkanli, A., and West, M. (1996) Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, **83**, 67–79.

Roberts, G. O and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterisation for the Gibbs sampler. *Journal of the Royal Statistical Society*, B, **59**, 291–317.

Roberts, G. O. and Tweedie, R. L. (1996) Geometric convergence and central limit theorems for multidimensional Hastings and Metropolis algorithms. *Biometrika*, **83**, 95–110.

Wahba, G. (1983) Bayesian confidence intervals for the cross-validated smoothing spline. *Journal of the Royal Statistical Society*, B, **45**, 133–150.