Comment on
“Bayesian backfitting” by T. J. Hastie and R. J. Tibshirani

R. Dennis Cook and Iain Pardoe *

28 January, 2000

* R. Dennis Cook is Professor and Iain Pardoe is Graduate Student at the School of Statistics, University of Minnesota, St. Paul, MN 55108.
Their work was partially supported by the National Science Foundation.

1 Introduction

Hastie and Tibshirani propose an intriguing idea, neatly linking Bayesian modeling of the functions in a generalized additive model with Gibbs sampling to obtain posterior realizations of these functions. Since their procedure utilizes only smoother matrices for individual predictors, S_j, partial residuals, $\mathbf{r}_j$, and normal random vectors, $\mathbf{z}_j$, the method would appear to be applicable to any models with additive components that can be expressed in the form $S_j\mathbf{y}$.

A natural question to ask of any proposed methodology is "to what use can it be put?" Hastie and Tibshirani's examples, while interesting in themselves, left us questioning what information could be gleaned from plots such as Figures 1 and 2 for the ozone data and Figure 3 for the growth curves data. For example, do the individual realizations in Figure 2 add anything to the information already provided by the point-wise posterior intervals? Figure 4 goes some way to addressing these thoughts with a graphical display of two functionals of the posterior realizations. We decided to pursue these thoughts in a different direction—that of model checking—and we outline our findings in Section 2. We discuss other potential applications in Section 3, and make some more general comments in Section 4.

2 Marginal Model Plots

The goal of a regression analysis can be expressed as inference about the dependence of an unknown cdf F of the conditional random variable $y \mid \mathbf{x}$ on the value of $\mathbf{x}$. Consider a generic regression model for $y \mid \mathbf{x}$ represented by the cdf M; estimating this model gives rise to an estimated cdf $\widehat{\mathrm{M}}$. We now consider graphics for comparing selected characteristics of F to the corresponding characteristics of $\widehat{\mathrm{M}}$. We use the fact that $\mathrm{F}(y \mid \mathbf{x}) = \mathrm{M}(y \mid \mathbf{x})$ for all values of $\mathbf{x}$ in its sample space if and only if $\mathrm{F}(y \mid h) = \mathrm{M}(y \mid h)$ for all functions $h = h(\mathbf{x})$. This is a more general version of the approach proposed by Cook and Weisberg (1997) which sets $h = \mathbf{a}^T\mathbf{x}$, where $\mathbf{a} \in \mathbb{R}^{\mathrm{p}}$. In particular, we focus on comparing a non-parametric estimate of the mean of $y \mid h$ to the corresponding mean computed from $\widehat{\mathrm{M}}$, for various functions h.

For some fixed h, plot y versus h. Add a non-parametric mean estimate, say a cubic smoothing spline with fixed degrees of freedom, to the plot—denote this $\widehat{\mathrm{E}}_{\mathrm{F}}(y \mid h)$, where E_{F} denotes expectation under F. We wish to compare this mean estimate with a mean estimate under $\widehat{\mathrm{M}}$, $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y \mid h)$, where $\mathrm{E}_{\widehat{\mathrm{M}}}$ denotes expectation under $\widehat{\mathrm{M}}$. Since $\mathrm{E}_{\widehat{\mathrm{M}}}(y \mid h) = \mathrm{E}[\mathrm{E}_{\widehat{\mathrm{M}}}(y \mid \mathbf{x}) \mid h]$, we can obtain $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y \mid h)$ from a non-parametric mean estimate for the regression of the fitted values under M, $\mathrm{E}_{\widehat{\mathrm{M}}}(y \mid \mathbf{x})$, on h. We can then add this to the plot to obtain a marginal model plot (MMP) for h; this can be thought of as a plot for checking the model in the (marginal) direction h. Using the same method (and smoothing parameter) to obtain this estimate as that used to obtain the mean estimate under F allows point-wise comparison of the two estimates, since any estimation bias should cancel. See Bowman and Young (1996) for further discussion of this point. If the model is a close representation of F, we can expect that for any quantity h the marginal mean estimates should agree, $\widehat{\mathrm{E}}_{\widehat{\mathrm{M}}}(y \mid h) \approx \widehat{\mathrm{E}}_{\mathrm{F}}(y \mid h)$.

Ideas for selecting which MMPs (ie which functions h) to consider in practice are given in Cook and Weisberg (1997), with additional discussion provided in Cook (1998) and Cook and Weisberg (1999). Some examples of useful MMPs include those for fitted values, individual predictors, and linear combinations of the predictors. Any indication that the estimated marginal means do not agree for one particular MMP suggests that the model could perhaps be improved; if they agree for a variety of plots, we have support for the model. The ideas above can be extended to variance estimates to provide further ways for checking models.

Consider, for example, a MMP for the fitted values for Hastie and Tibshirani's ozone data example with four predictor variables. The plot in Figure 1 shows a systematic discrepancy between the (black) mean estimate under F and the (gray) mean estimate under $\widehat{\mathrm{M}}$; the mean estimate under $\widehat{\mathrm{M}}$ is too low on the left, too high in the middle, and too low again on the right. Both mean estimates were calculated using the S-Plus function `smooth.spline` with (the default) four degrees of freedom.

On the other hand, relative to the variation in the data, the mean estimate under $\widehat{\mathrm{M}}$ does not appear to be too far from the mean estimate under F. So, are the discrepancies enough to indicate any potential for model improvement? Porzio and Weisberg (1999) provide some frequentist methodology to address this issue: point-wise reference bands to aid visualization and statistics to calibrate discrepancies. Hastie and Tibshirani's procedure also provides methodology to address this issue. They make the well-taken points that we can make use of the individual realizations of the posterior distributions of the functions in an additive model, and display the posterior distributions of *interesting functionals* of them. They also note that we can carry out Bayesian inference for any quantity of interest. This would appear to offer a Bayesian way to aid visualization in a MMP, with potential possibilities for calibrating discrepancies.

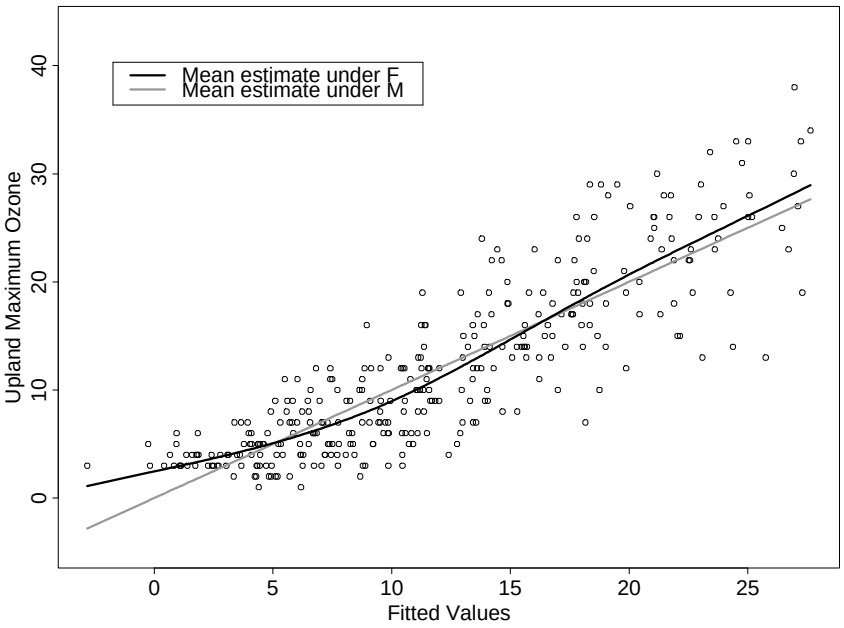

Figure 1: Marginal model plot for the fitted values for the additive model fit to four air pollution variables.

For any particular MMP, it would be useful to display mean estimates for the individual realizations from the posterior distribution of the fitted values, where the fitted-value realizations are just $\sum_{j=0}^{p} \mathbf{f}_j^t$, $t = 1, 2, 3, \ldots$ and $\mathbf{f}_j^t = S_j \mathbf{r}_j^t + \sigma S_j^{1/2} \mathbf{z}_j^t$. So, instead of adding the mean estimate under $\widehat{\mathrm{M}}$ to the plot of the mean estimate under F, we can instead add a mean estimate for each Gibbs sample, G^t, and obtain what we call a *Gibbs marginal model plot* (GMMP). If enough samples are taken, say 50 or 100, the Gibbs mean estimates will form an approximate mean estimate *band* under $\widehat{\mathrm{M}}$. This plot may provide a visual way of determining whether there is any evidence to contradict the possibility that $\mathrm{F}(y \mid h) = \mathrm{M}(y \mid h)$. Intuitively, if, for a particular h, the mean estimate under F lies substantially outside the mean estimate band under $\widehat{\mathrm{M}}$ (formed from the mean estimates under G^t), then perhaps the model can be improved. If, no matter what the function h is, the mean estimate under F lies broadly inside the mean estimate band under $\widehat{\mathrm{M}}$, then perhaps the model provides a reasonable description of the conditional distribution of $y \mid \mathbf{x}$. It would appear to be possible to supplement this purely graphical methodology with more formal Bayesian inference.

Consider a GMMP for the fitted values for the ozone data. Hastie and Tibshirani kindly provided us with the S-Plus functions for implementing the ideas in their paper, as well as with help in using their code. This enabled us to construct the GMMP in Figure 2. The Gibbs sampling was carried out using the fully Bayesian procedure described in Hastie and Tibshirani's Section 4, with a warmup period of 300 iterations. The plot shows the (black) mean estimate under F lying mostly outside the mean estimate band under $\widehat{\mathrm{M}}$ (formed from 100 (gray) mean estimates under G^t). This appears to offer clear evidence that the fitted model can be improved.

As curious applied statisticians, we couldn't resist trying to see if we could come up with a better model for these data. One particular technique we applied was *Sliced Average Variance Estimation* (SAVE), introduced by Cook and Weisberg (1991) and developed by Cook and Lee (1999). SAVE is a model-free method for estimating the smallest subspace $\mathcal{S}$ of $\mathbb{R}^{\mathrm{p}}$ so that y and $\mathbf{x}$ are independent given the projection of $\mathbf{x}$ onto $\mathcal{S}$, $P_{\mathcal{S}}\mathbf{x}$. In words, all the information about y that is available from $\mathbf{x}$ is contained in $P_{\mathcal{S}}\mathbf{x}$. Following Li (1991), $\mathcal{S}$ is a *dimension reduction subspace* for the regression of y on $\mathbf{x}$. The smallest such $\mathcal{S}$ is called the *central* subspace, $\mathcal{S}_{y|\mathbf{x}}$ (Cook 1994; Cook 1998); SAVE yields a subspace estimate, $\mathcal{S}_{\mathrm{SAVE}} \subset \mathcal{S}_{y|\mathbf{x}}$. This estimate can then be used to postulate a model, as described by example below.

Since the additive model fit above appears unable to account for the curvature in the MMP for the fitted values, we felt that SAVE might be able to provide us with a better model. We used the SAVE methodology to infer the dimension of $\mathcal{S}_{y|\mathbf{x}}$ to be two, and obtained two linear combinations of predictors, w_1 and w_2, as an estimate of a basis for $\mathcal{S}_{y|\mathbf{x}}$.

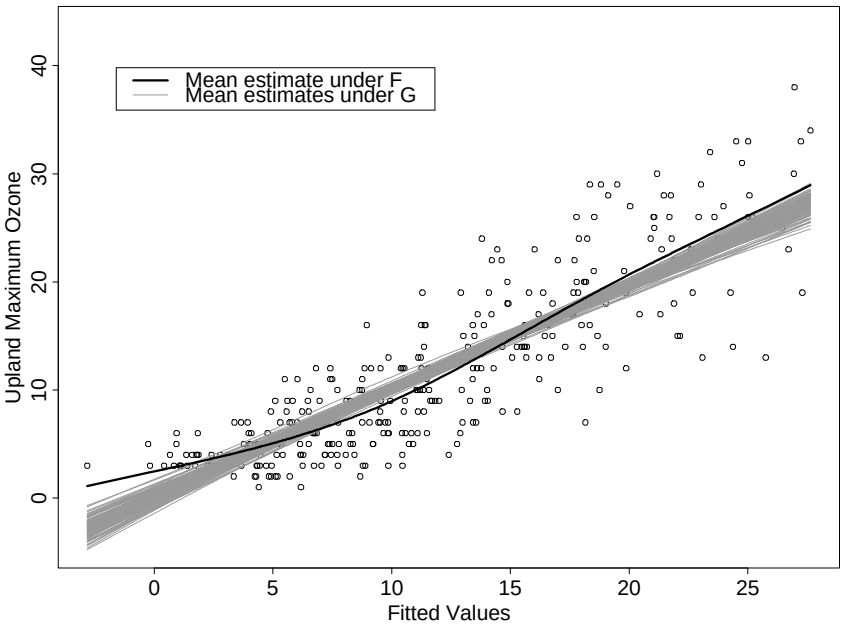

Figure 2: Gibbs marginal model plot for the fitted values for the additive model fit to four air pollution variables.

A 3D plot of y versus w_1 and w_2 indicated that an interaction term, w_{12}, might also be important. So, we decided to fit an additive model: $\mathrm{E}(y \,|\, \mathbf{x}) = \alpha + f_1(w_1) + f_2(w_2) + f_{12}(w_{12})$. Smoothing splines with (the S-Plus default) four degrees of freedom were used to estimate the f functions. A GMMP for the fitted values for this model is shown in Figure 3. The plot shows the (black) mean estimate under F lying inside the mean estimate band under $\widehat{\mathrm{M}}$ (formed from the (gray) mean estimates under G^t). There is little evidence in *this* plot to suggest that the fitted model can be improved.

However, there is evidence from a MMP for one of the original predictors, Inversion Base Height, that this model too could be improved. Again, the discrepancy between the marginal mean estimates in this plot (not shown) is difficult to assess relative to the variability in the data. The corresponding GMMP in Figure 4 allows this discrepancy to be evaluated visually, and the plot reinforces the supposition that the model could possibly be improved (at least for low values of Inversion Base Height).

Having applied Hastie and Tibshirani's methodology to these data, GMMPs appear to offer a quick and easy way to graphically check models. The Gibbs sampling only needs to be done once for each model; with Hastie and Tibshirani's S-Plus code this is straightforward. The analyst can then cycle through a variety of GMMPs to get some guidance on whether (and how) an alternative model might provide an improvement. For example, in the above analysis, a next step might be to develop a model that deals with low values of Inversion Base Height more satisfactorily, say by increasing the degrees of freedom for the smoothers in the additive model, or by trying different smoothers such as *loess*.

Does a GMMP suffer the same shortcoming as Hastie and Tibshirani's Figure 2—namely, would we be able to obtain equivalent information by plotting point-wise posterior intervals instead of individual posterior realizations? The answer to this question would surely be yes, were it not for the fact that it is not clear how such intervals might be defined in practice. For example, posterior intervals could be calculated for the fitted values in an additive model by summing the posterior intervals for the individual functions in the model. It would then be straightforward to plot the point-wise intervals on a MMP for the fitted values. But, for MMPs for any other function h, it is unclear what point-wise posterior intervals should be defined to be. One possibility would be to smooth the point-wise upper and lower limits for the fitted values using the same method as used to obtain the mean estimates under F and $\widehat{\mathrm{M}}$, but it is not clear that this will give us point-wise posterior intervals for $\mathrm{E}_{\widehat{\mathrm{M}}}(y \,|\, h)$.

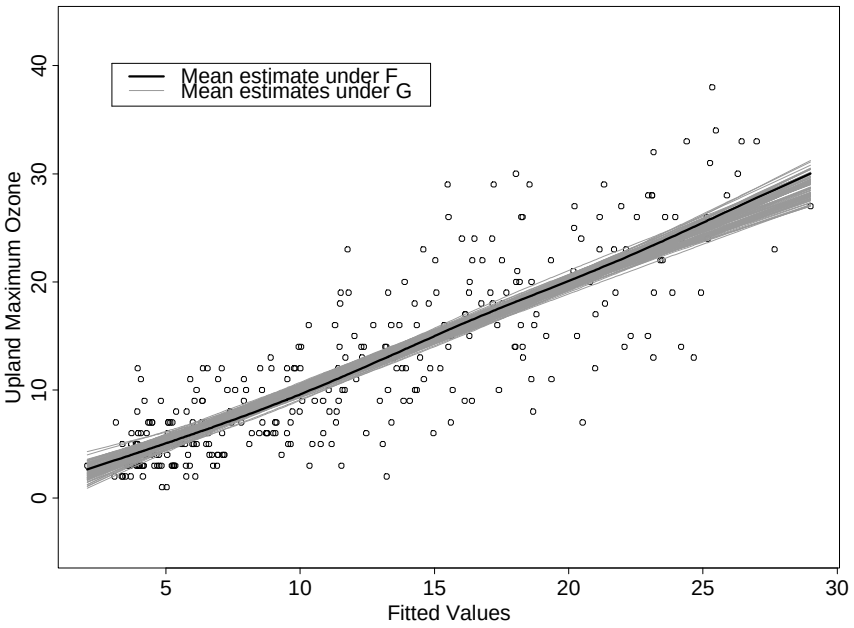

Figure 3: Gibbs marginal model plot for the fitted values for the additive model fit to (w_1, w_2, w_{12}).

3 Other Potential Applications

Returning to Hastie and Tibshirani's Figure 1, can *these* plots (of partial residuals versus individual predictors) be used for model checking? The answer to this question would appear to be no. The black curves are smooths of the partial residuals, $\mathbf{f}_j = S_j \mathbf{r}_j$, while the gray curves are the Gibbs posterior realizations, $\mathbf{f}_j^t = S_j \mathbf{r}_j^t + \sigma S_j^{1/2} \mathbf{z}_j^t$. These plots would appear to offer visualization only of the variability in the fitted functions. Appropriate plots for model checking in this context are GMMPs for the individual predictors, as shown for example in Figure 4.

There are other plots used in model checking and regression diagnostics that can be difficult to assess relative to the variation in the data. Some examples include: residual plots; CERES plots, which are a generalization of partial residual plots and were introduced by Cook (1993); net-effect plots, which aid in assessing the contribution of a selected predictor to a regression and were introduced by Cook (1995). The ideas discussed above would appear to have a rôle to play in the analysis of such plots. Work is in progress on these issues, as well as on developing supplementary Bayesian inference methodology.

4 Miscellanea

Hastie and Tibshirani's procedure appears to live up to its claim of modularity and generality. Although the procedure derives from the back-fitting algorithm for fitting additive models, it could probably be applied fairly easily to other families of models such as generalized linear models. Whether the procedure could also be described as conceptually simple is perhaps more open to debate. For example, choosing priors for the variance components is far from trivial, and MCMC convergence should always be checked in practice. That said, there is clearly a wealth of potential applications for the posterior samples generated with this technique.

SAVE techniques can be applied using *Arc* (Cook and Weisberg 1999), a comprehensive regression program. Information about the program is available at the Internet site `www.stat.umn.edu/arc`.

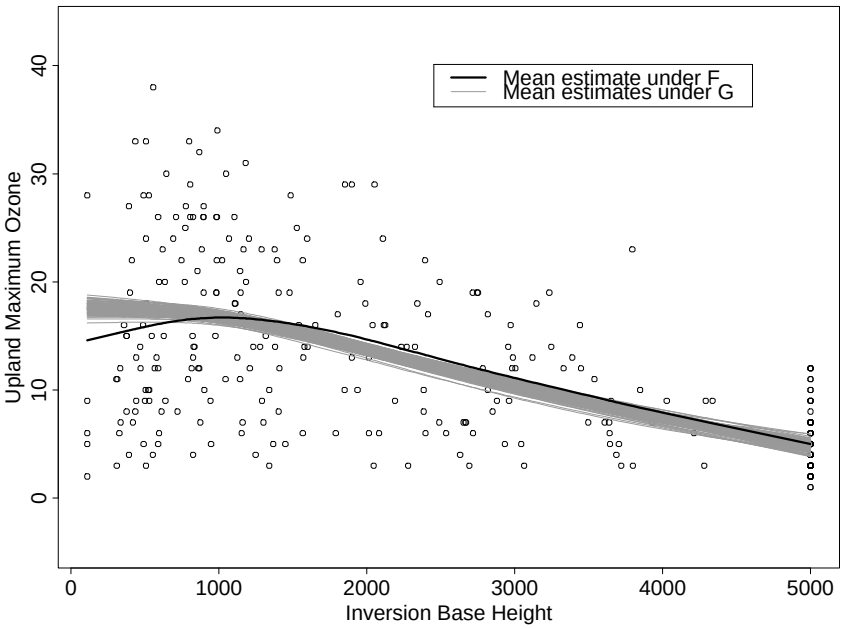

Figure 4: Gibbs marginal model plot for Inversion Base Height for the additive model fit to (w_1, w_2, w_{12}).

References

Bowman, A. W. and S. Young (1996). Graphical comparison of nonparametric curves. *Applied Statistics 45*, 83–98.

Cook, R. D. (1993). Exploring partial residual plots. *Technometrics 35*, 351–362.

Cook, R. D. (1994). Using dimension-reduction subspaces to identify important inputs in models of physical systems. In *Proceedings of the Section on Physical and Engineering Sciences*, Alexandria, VA, pp. 18–25. American Statistical Association.

Cook, R. D. (1995). Graphics for studying net effects of regression predictors. *Statistica Sinica 5*, 689–708.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions through Graphics*. New York: Wiley.

Cook, R. D. and H. Lee (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association 94*, 1187–1200.

Cook, R. D. and S. Weisberg (1991). Comment on "Sliced inverse regression for dimension reduction" by K.-C. Li. *Journal of the American Statistical Association 86*, 316–342.

Cook, R. D. and S. Weisberg (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association 92*, 490–499.

Cook, R. D. and S. Weisberg (1999). *Applied Regression Including Computing and Graphics*. New York: Wiley.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association 86*, 316–342.

Porzio, G. C. and S. Weisberg (1999). Tests for lack-of-fit of regression models. Technical Report 634, School of Statistics, University of Minnesota.