Berger, J. O., and Pericchi, L. R. (1996), "The Intrinsic Bayes Factor for Model Selection and Prediction," *Journal of the American Statistical Association*, 91, 109–122.
Casella, G., and George, E. I. (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.
Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society*, Ser. B, 80, 331–350.
Erkanli, A., and Gopalan, R. (1996), "Bayesian Nonparametric Regression: Smoothing Using Gibbs Sampling," in *Bayesian Statistics and Econometrics: Essays in Honor of Arnold Zellner*, eds. D. Berry, K. Chaloner, and J. Geweke, New York: Wiley, pp. 267–277.
Gelfand, A. E., and Smith, A. F. M. (1990), "Sampling-Based Approaches to Calculating Marginal Densities," *Journal of the American Statistical Association*, 85, 398–409.
George, E. I., and McCulloch, R. E. (1993), "Variable Selection via Gibbs Sampling," *Journal of the American Statistical Association*, 88, 881–889.
Hardle, W., and Korostelev, A. (1996), "Search for Significant Variables in Nonparametric Additive Regression," *Biometrika*, 83, 541–549.
Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge, U.K.: Cambridge University Press.
Hastie, T. J., and Tibshirani, R. J. (1987), "Generalized Additive Models: Some Applications," *Journal of the American Statistical Association*, 82, 371–386.
Kohn, R. (1983), "Consistent Estimation of Minimal Model Dimension," *Econometrica*, 51, 367–376.
Min, C., and Zellner, A. (1993), "Bayesian and non-Bayesian Methods for Combining Models and Forecasts With Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89–118.
O'Hagan, A. (1995), "Fractional Bayes Factors for Model Comparison" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 57, 99–138.
Raftery, A., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179–191.
Rao, C. R. (1973), *Linear Statistical Inference and Its Applications* (2nd ed.), New York: Wiley.
Schwartz, G. (1978), "Estimating the Dimension of a Model," *The Annals of Statistics*, 6, 461–464.
Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–344.
Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1762.
Wahba, G. (1978), "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society*, Ser. B, 40, 364–372.
——— (1990), *Spline Models for Observational Data*, CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59, Philadelphia: SIAM.
Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995), "Smoothing Spline ANOVA for Exponential Families, With Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy," *The Annals of Statistics*, 23, 1865–1895.
Wong, C., and Kohn, R. (1996), "A Bayesian Approach to Additive Semiparametric Regression," *Journal of Econometrics*, 74, 209–235.
Wood, S., and Kohn, R. (1998), "A Bayesian Approach to Robust Nonparametric Binary Regression," *Journal of the American Statistical Association*, 93, 203–213.
Wood, S., Shively, T. S., and Kohn, R. (1996), "Model Selection in Spline Nonparametric Regression," working paper, Australian Graduate School of Management.

Comment

Babette A. BRUMBACK, David RUPPERT, and M. P. WAND

1. INTRODUCTION

This article by Tom Shively, Robert Kohn, and Sally Wood provides a very effective solution to the difficult model selection problem in multiple predictor semiparametric regression, adding to a long list of impressive work of this type by Kohn and coauthors. As usual, the authors have done a thorough job, with lots of simulation testing to ensure good performance of their proposed strategy.

As the article's title suggests, the model and fitting procedure can be broken into two components: (a) variable selection through Bayesian modeling and Markov chain Monte Carlo (MCMC) schemes, and (b) function estimation. Concerning (a), we certainly have less expertise about computational Bayesian methods than the authors, and there is little we can add to their MCMC algorithm. Most of our research involves function estimation, and most of our comments are on this topic, though Section 4 briefly discusses non-Bayesian methods of variable selection. Also, for simplicity, our discussion deals only with the Gaussian case.

2. FUNCTION ESTIMATION

The function estimation component consists of a model expressed as a line plus an integrated Wiener process (with arbitrary variance) and with fitting achieved using state-space formulation and application of the Kalman filter.

2.1 Complexity

Additive models continue to gain popularity in applied research as a flexible and interpretable regression technique. For example, they are now used routinely in environmental epidemiology studies conducted at the Harvard School of Public Health. The input is a spreadsheet of numbers; one column representing measurements on the response variable, the remainder being measurements on each of the covariates. The output is a set of curves and coefficients, along with variability estimates, which describe the effects of each

Babette A. Brumback is Postdoctoral Research Fellow and M. P. Wand is Associate Professor, Department of Biostatistics, School of Public Health, Harvard University, Boston, Massachusetts 02115. David Ruppert is Professor, School of Operations Research and Industrial Engineering, Cornell University, Ithaca, NY 14853. These comments benefited from a conversation with Jim Hobert. This research was supported by National Science Foundation grant DMS-9804058 (Ruppert) and by U.S. Environmental Protection Agency grant R 824757 (Wand).

Journal of the American Statistical Association
September 1999, Vol. 94, No. 447, Theory and Methods

covariate on the mean response. Hence both input and output are rather simple mathematical concepts: numbers and univariate functions.

One feature that distinguishes each of the various additive model-fitting procedures is its level of complexity. The method proposed here requires knowledge of continuous time stochastic processes, state-space models, MCMC, and the Kalman filter. A researcher interested in learning about the mechanism by which his or her numbers are transformed into curves is destined to wade through some rather advanced and detailed mathematics. Is this level of complexity really necessary? We return to this issue in Section 3.

Another complexity-related issue, not mentioned in the article, is speed. For analysis of the heart attack data, 70,000 MCMC iterations are used; for the model with six predictors, each one of these iterations requires at least 12 smoothing operations. This means that fitting the full model requires nearly 1 million smoothing operations. It would be useful to know the time taken to fit such a model, and whether there is room for improvement by using a computationally less intensive smoother.

2.2 Implementability

We have some concerns about the implementability of the strategy proposed by Shively, Kohn, and Wood. The authors point out that the details of the smoothing component have been given by Wong and Kohn (1996). That article refers to algorithms of Ansley and Kohn (1990) and Carter and Kohn (1994). Carter and Kohn (1994) referred to the book by Anderson and Moore (1979) for details on the Kalman filter. Even after collecting each of these references and piecing them together, are there enough details for a talented C or Fortran (but otherwise uninitiated) programmer to code up the method? We are not very confident, as one of us (MPW) once tried to implement the smoothing spline algorithm of Kohn and Ansley (1987) but found some details missing (e.g., how does one compute $\chi(t)$?), and eventually had to give up. We thus suggest that the authors publish a generic, nontechnical, description of the algorithm used in this important article in a single place; perhaps in electronic form made available on the Internet. Otherwise, the use and extension of this methodology is likely to be confined to a small, select group.

3. A SIMPLE BAYESIAN FUNCTION ESTIMATOR

Is it really necessary to have function estimation depending on a continuous-time stochastic process model? In this section we describe an alternative Bayesian function estimator that is somewhat less complicated and more in keeping with the input and output of additive models. We first treat the case of ordinary nonparametric regression.

3.1 Ordinary Nonparametric Regression

Consider a set of scatterplot data $(x_i, y_i), 1 \le i \le n$, and the ordinary nonparametric regression model

$$y_i = f(x_i) + \sigma\varepsilon_i \tag{1}$$

where the ε_i are independent N(0, 1) variates. Let $\kappa_1, \ldots, \kappa_K$ be a set of distinct numbers inside the range of the x_i's and let $x_+ = \max(0, x)$. A *random coefficient linear regression spline* model for f is

$$f(x) = \beta_0 + \beta_1 x + \tau \sum_{k=1}^{K} u_k (x - \kappa_k)_+, \quad \text{where}$$

$$\mathbf{u} = \begin{bmatrix} u_1 \\ \vdots \\ u_K \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}) \tag{2}$$

is independent of $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^T$. In the situation where the positioning of the κ_k's follows roughly the distribution of the x_i's, this representation of f has similarities with (2). When $\tau = 0, f$ is linear, but for $\tau > 0$, the *truncated lines* $(x - \kappa_k)_+$ flexibly allow for nonlinearities in f in the same way as the integrated Wiener process. Note that more smoothness could be attained by instead using quadratic or cubic polynomials and truncated polynomials.

We can write the model conveyed by (1) and (2) in matrix notation as

$$\mathbf{y} = \mathbf{X}\beta + \tau\mathbf{Z}\mathbf{u} + \sigma\varepsilon, \qquad \begin{bmatrix} \mathbf{u} \\ \varepsilon \end{bmatrix} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

and

$$\mathbf{Z} = \begin{bmatrix} (x_1 - \kappa_1)_+ & \cdots & (x_1 - \kappa_K)_+ \\ \vdots & \ddots & \vdots \\ (x_n - \kappa_1)_+ & \cdots & (x_n - x_K)_+ \end{bmatrix}.$$

Equation (3) is recognizable as a normal linear mixed model and, for given τ and σ, the *best linear unbiased predictor* (BLUP) of $\mathbf{y}$,

$$\hat{\mathbf{f}} = \mathbf{X}\hat{\beta} + \tau\mathbf{Z}\hat{\mathbf{u}}, \tag{4}$$

where $\hat{\beta} = \{\mathbf{X}^T(\tau^2\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I})^{-1}\mathbf{X}\}^{-1}\mathbf{X}^T(\tau^2\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I})^{-1}\mathbf{y}$ and $\hat{\mathbf{u}} = \tau(\tau^2\mathbf{Z}\mathbf{Z}^T + \sigma^2\mathbf{I})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\hat{\beta})$ can be treated as an estimator of $\mathbf{f} = [f(x_1), \ldots, f(x_n)]^T$; extension to $f(x)$ for arbitrary x is straightforward. Note that $\hat{\mathbf{f}}$ can be rewritten as

$$\hat{\mathbf{f}} = \mathbf{C}(\mathbf{C}^T\mathbf{C} + \lambda\mathbf{D})^{-1}\mathbf{C}^T\mathbf{y},$$

where $\mathbf{C} = [\mathbf{X}\mathbf{Z}], \mathbf{D} = \mathrm{diag}(0, 0, 1, 1, \ldots, 1)$ and $\lambda = \sigma^2/\tau^2$. This shows that $\hat{\mathbf{f}}$ is equivalent to the *penalized spline smoother* of Eilers and Marx (1996) (also see Ruppert and Carroll 1999). However, the BLUP representation (4) lends itself, via Robinson (1991), to a host of other derivations, including one as a Bayesian estimator.

As shown in section 4.2 of Robinson (1991), if β is regarded as a parameter with a uniform, improper prior, then $\hat{\mathbf{f}}$ corresponds to the posterior mode. Section 6.4 of Robinson (1991) shows how the Kalman filter can be used to

compute (4); see also the discussion by Spall (1991) and the rejoinder by Robinson.

The representation of smoothing splines as BLUP's was laid out by Speed (1991) while discussing the work of Robinson (1991). But the simplicity of the BLUP representation for penalized splines is particularly striking. It also demonstrates that one is not confined to smoothing splines and Wiener processes if one insists on a Bayesian framework.

In addition to conceptual simplicity, the penalized spline approach also offers significant computational savings. This is because they are *low-rank* smoothers, as defined by Hastie (1996), and involve an order K matrix inversion rather than one of order n. The resulting computational advantages, such as not requiring backfitting, have been described by Hastie (1996) and Marx and Eilers (1998).

Also note that one can use variance component estimation techniques, such as restricted maximum likelihood (REML), to estimate τ^2 and σ^2 and thus choose the smoothing parameter λ in (4). This is analogous to the marginal likelihood smoothing parameter selection method mentioned by Shively, Kohn, and Wood for the state-space/smoothing spline approach (Ansley and Kohn 1985; Wahba 1985).

In our current implementations of this estimator, we select λ by minimizing generalized cross-validation (GCV), which is

$$\|\mathbf{y} - \hat{\mathbf{f}}\|^2/\{1 - A\ \mathrm{df}(\lambda)/n\}^2, \tag{5}$$

where A has a default value of 1 and $\mathrm{df}(\lambda)$ is the trace of the hat matrix $\mathbf{C}(\mathbf{C}^T\mathbf{C} + \lambda\mathbf{D})^{-1}\mathbf{C}^T$. (We introduce A because it has a value different than 1 in later discussion.)

How should the knots, $\kappa_1, \ldots, \kappa_K$, be selected? The number of knots K is *not* being used as a smoothing parameter; λ plays that role. One needs enough knots to ensure sufficient flexibility to fit the data, but after that additional knots do not change the fit much. We place the knots at sample quantiles of the unique x values so that there are equal, or nearly equal, numbers of x values between knots. To illustrate that the number of knots has little effect, we intended to include a plot showing 10- and 40-knot quadratic spline fits to the "sine" function used in the article. However, the two fits could not be distinguished by eye in any of the Monte Carlo samples we tried, so we omitted the plot! It takes about four knots to fit one cycle of a sine wave, so the fit deteriorates when there are less than eight knots. Ruppert and Carroll (1999) proposed a simple algorithm to ensure that there are enough knots.

3.2 Extension to the Bivariate Additive Model

For the bivariate additive model (1) of Shively, Kohn, and Wood, one could replace g_1 and g_2 and (2) and (3) by

$$g_1(s) = \tau_1 \sum_{k=1}^{K_1} u_{1k}(s - \kappa_{1k})_+$$

and

$$g_2(t) = \tau_2 \sum_{k=1}^{K_2} u_{2k}(t - \kappa_{2k})_+,$$

where $\mathbf{u}_1 = [u_{11} \cdots u_{1K_1}]^T$ and $\mathbf{u}_2 = [u_{21} \cdots u_{2K_2}]^T$ each have the same distribution as u and the κ_{1k} and κ_{2k} are positioned to capture structure in directions 1 and 2.

Further research is required to check whether or not the other details of the Shively, Kohn, and Wood model and its fitting are affected by such a change. If they are not, then it would have the advantages that

- the complexity of the methodology is reduced due to not relying on Wiener processes, but rather on ordinary Gaussian random vectors and simple mathematical functions (i.e., truncated polynomials)
- computation is faster due to the low-rank nature of the smoothing operations.

One of us (DR) has implemented a penalized spline additive model in Matlab with the smoothing parameter selected by GCV and it is extremely fast, taking less than one second on a SPARC Ultra 1 for 400 observations and two predictors.

4. MODEL SELECTION

4.1 Simple Non-Bayesian Variable Selection

Both smoothing parameters and models can be selected by GCV. GCV certainly has an advantage in simplicity and speed compared to the MCMC implementation of Bayesian model selection. It would be interesting to learn how GCV compares in terms of precision.

GCV is used by Friedman's (1991) MARS and Stone, Hansen, Kooperberg, and Truong's (1997) POLYMARS to select variables from the set of all tensor product spline-basis functions. MARS and POLYMARS are not oriented toward additive models or variable selection. However, they have the capacity to select an additive model when one fits well and to eliminate a variable by selecting none of the basis functions containing that variable. A comparison of the authors' methods with these existing methodologies would be useful. Note that Stone et al. (1997) considered regression, generalized regression, and other problems that are all special cases of their "extended linear models."

Now consider additive modeling with several candidate predictor variables. A simple method for eliminating unnecessary variables in the penalized spline model of Section 3 would be to iterate the following backward elimination step. Given that the current model has $M > 0$ variables, compute GCV for each of the M submodels with one variable deleted. If the $(M-1)$-variable submodel minimizing GCV has smaller GCV value than the M-variable model, then replace the M-variable model with this "best" $M-1$-variable submodel and continue. Otherwise, retain the M-variable model and stop.

We tried this idea on the sampling design that the authors used in their Experiment 1 where the three functions were "flat," "linear," and "exponential." In 50 trials, variable 1 (corresponding to "flat") was correctly deleted 44 times. In the other six samples, no variables were deleted. In no case was either the second or third variable incorrectly deleted. For a single dataset, the entire model selection procedure takes only 3 seconds to run on an Ultra 1.

It is known that GCV, like C_p and the Akaike information criterion (AIC), tends to select too many variables, because it strives only for good prediction. From the prediction standpoint, occasionally retaining an unnecessary variable is preferable to occasionally omitting a needed variable. Schwarz's Bayes information criterion (BIC) is often used to decrease the probability of selecting unnecessary variables. BIC is similar to GCV with A in (5) equal to $\log(n)/2$. We repeated the experiment with this value of A, and the correct model was chosen all 50 times. This suggests that BIC or GCV with a BIC-like correction may be competitive with the more complex Bayesian methods. Clearly, more detailed comparisons would be interesting. Note that in this experiment $\log(n)/2$ is 3.0, which is similar to the A suggested by Stone et al. (1997), namely A around 2.5.

4.2 Interactions

There's an old rule-of-thumb in the response surface literature that whenever quadratic main effect terms are needed, then two-way interactions are likely also needed. This rule is based on empirical experience but is sensible mathematically since two-way interactions are also quadratic.

This rule-of-thumb suggest that additive models with nonlinear main effects probably are ignoring significant interaction terms. Of course, the user of additive modeling software may remain blissfully ignorant of these interactions. MARS and POLYMARS model selection can pick up these interaction terms. For this reason, unless the authors' methodology is expanded to handle interactions, MARS and POLYMARS may be more suitable for routine model selection.

5. CONFIDENCE BOUNDS

The confidence bounds in Figures 7 and 8 widen from left to right, because the f_j functions are constrained to be 0 at their left boundaries. Are these confidence bounds scientifically meaningful? We prefer to estimate the mean response as a function of the jth predictor with the other variables fixed at some values (e.g., their means). Then the confidence bounds are roughly symmetric from left to right and are independent of the identifiability constraint. The confidence bounds are easily constructed in the penalized spline framework discussed earlier.

6. SUMMARY

The "signal extraction/state-space/Kalman filter" approach to semiparametric regression and time series modeling, started by Wecker and Ansley (1980) and continued by Ansley, Kohn, Shively, and collaborators is close to reaching the end of its second decade. It has been outstandingly successful, leading to flexible and elaborate models that incorporate, for example, function estimation, time series structure, missing-data interpolation, and outlier detection within an appealing model-based framework. (See the seven points listed at the start of Ansley and Kohn 1989.) Integration with MCMC methods in the 1990s has been a giant step forward and has made even more complex models feasible.

This discussion article is a landmark event for this research effort, allowing us discussants to pay testimony to its remarkable achievements. It is also an opportunity for constructive feedback. The concerns that we have raised about accessibility of the methodology are, we feel, quite genuine. Those regarding implementability do need to be addressed, if this vast battery of methodology is to be used for scientific enquiry. Our comments regarding alternative function estimation strategies are not meant to imply that Wiener process approaches should not have been used in the first place; for some time it has been thought to be the only reasonable way to formulate a Bayesian approach. But at a time when function estimation is finding its way into applications at an unprecedented rate, there is, in our view, a clear imperative for us "smoothers" to seek out simplifications that will make our methodology more palatable and user friendly.

ADDITIONAL REFERENCES

Anderson, B. D. O., and Moore, J. B. (1979), *Optimal Filtering*, Englewood Cliffs, NJ: Prentice-Hall.

Ansley, C. F., and Kohn, R. (1989), "Comment on 'Linear Smoothers and Additive Models' by Buja, A., Hastie, T., and Tibshirani, R.," *The Annals of Statistics*, 17, 540–543.

——— (1990), "Estimation, Filtering, and Smoothing in State-Space Models With Partially Diffuse Initial Conditions," *Journal of Time Series Analysis*, 11, 277–293.

——— (1985), "Estimation, Filtering, and Smoothing in State-Space Models With Incompletely Specified Initial Conditions," *The Annals of Statistics*, 13, 1286–1316.

Carter, C. K., and Kohn, R. (1994), "On Gibbs Sampling for State-Space Models," *Biometrika*, 81, 541–554.

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With B-Splines and Penalties" (with discussion), *Statistical Science*, 89, 89–121.

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics*, 19, 1–141.

Hastie, T. J. (1996), "Pseudosplines," *Journal of the Royal Statistical Society*, Ser. B, 58, 379–396.

Kohn, R., and Ansley, C. F. (1987), "A New Algorithm for Spline Smoothing Based on Smoothing a Stochastic Process," *SIAM Journal of Scientific and Statistical Computing*, 8, 33–48.

Marx, B. D., and Eilers, P. H. C. (1998), "Direct Generalized Additive Modeling With Penalized Likelihood," *Computational Statistics and Data Analysis*, 28, 193–209.

Robinson, G. K. (1991), "That BLUP is a Good Thing: The Estimation of Random Effects" (with discussion), *Statistical Science*, 6, 15–51.

Ruppert, D., and Carroll, R. J. (1999), "Spatially Adaptive Penalties for Spline Fitting," submitted.

Spall, J. C. (1991), Comment on "That BLUP is a Good Thing: The Estimation of Random Effects," by G. K. Robinson, *Statistical Science*, 6, 39–41.

Speed, T. (1991), Comment on "That BLUP is a Good Thing: The Estimation of Random Effects," by G. K. Robinson, *Statistical Science*, 6, 42–44.

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K. (1997), "Polynomial Splines and Their Tensor Products in Extended Linear Modeling" (with discussion), *The Annals of Statistics*, 25, 1371–1470.

Wahba, G. (1985), "A Comparison of GCV and GML for Choosing the Smoothing Parameter in the Generalized Spline Smoothing Problem," *The Annals of Statistics*, 13, 1378–1402.

Wecker, W. E., and Ansley, C. F. (1980), "Linear and Nonlinear Regression Viewed as Signal Extraction Problem," in *Proceedings of the Business and Economic Statistics Section, American Statistical Association*, pp. 44–51.