KOH, E. (1989). A smoothing spline based test of model adequacy in nonparametric regression. Ph.D. dissertation, Univ. Wisconsin, Madison.

WAHBA, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *J. Roy. Statist. Soc. Ser. B* **40** 364–372.

DEPARTMENT OF STATISTICS
UNIVERSITY OF ILLINOIS
CHAMPAIGN, ILLINOIS 61820

R. L. EUBANK AND P. SPECKMAN

Texas A & M University and University of Missouri, Columbia

The authors are to be congratulated on this interesting and thought-provoking paper. They have raised a number of important questions and issues concerning additive model methodology. We will discuss some of these below. Throughout, our comments will be restricted to the case of symmetric smoothers having eigenvalues in $[0, 1]$.

1. Exact and approximate concurvity. This paper contains a thorough treatment of the fundamental issues of existence and uniqueness of solutions for the normal equations arising from additive model estimation. The authors show that these equations will have multiple solutions in certain cases. This raises questions as to how analyses should proceed in the presence of exact concurvity. Results from linear models would suggest that if $\mathbf{f}$ represents any solution to the normal equations, then one should only examine functionals $\mathbf{l}^t\mathbf{f}$ of the solution that are "*estimable*" in the sense that $\mathbf{l}^t\mathbf{g} = 0$ whenever $\hat{\mathbf{P}}\mathbf{g} = 0$. Such functionals are invariant under all choices of solutions to the normal equations and will have unique expectations. According to Theorem 5 of the paper, "estimable" functionals are provided by np-vectors in the orthogonal complement of the linear span of vectors $\mathbf{g}^t = (\mathbf{g}_1^t, \ldots, \mathbf{g}_p^t)$ with $\mathbf{g}_j \in M_1(S_j)$ and $\mathbf{g}_+ = \mathbf{0}$. In particular we see that $\mathbf{f}_+$ is derived using "estimable" functionals.

Another approach to solving the normal equations for linear models of less than full rank is to reparameterize to obtain a full rank model. This is essentially what the authors have accomplished in Section 4.4 by extracting the projection parts from the smoothers, if linear dependencies are also eliminated from $M_1(S_1) + \cdots + M_1(S_p)$. The $\tilde{\mathbf{f}}_j$ are therefore obtained using "estimable" functionals and perhaps they are what should be studied when there is exact concurvity.

However, it seems to us that instances where an analysis should actually proceed in the presence of exact concurvity without some type of remedial action are rare. For example, in the case of smoothing splines, $M_1(S_j)$ is the linear span of the constant vector and $\mathbf{x}_j$. By Theorem 5 the concurvity space consists only of the constant vector unless the $\mathbf{x}_j$ are linearly dependent. In this latter case at least one of the variables should be dropped from the analysis to obtain meaningful estimates.

The real issue here seems to be approximate concurvity. As before we will draw an analogy with the linear regression case. In that setting approximate

concurvity means multicollinearity. Its presence causes difficulties in separating effects in the model with the consequence that parameter estimates may be poor. An investigator wishing to understand specific effects in a model with a high degree of multicollinearity is almost forced to consider partial regression fits and to study the effects of adjusted variables, perhaps with some variables omitted. We believe that this paradigm may have some merit in nonparametric additive models as well.

An example where a partial regression type approach has been suggested in the additive model framework is the semiparametric model discussed by the authors in Section 5.4. In that application, one variable, x_1, is exactly linear while another is nonparametric. The backfitting algorithm converges to $\hat{\beta} = (\mathbf{x}_1^t(I - S_2)\mathbf{x}_1)^{-1}\mathbf{x}_1^t(I - S_2)\mathbf{y}$ and $\hat{\mathbf{f}}_2 = S_2(\mathbf{y} - \mathbf{x}_1\hat{\beta})$. An algorithm attributed to Denby and derived independently by Speckman (1988) leads instead to estimates of the form $\hat{\beta} = (\mathbf{x}_1^t(I - S_2)^t(I - S_2)\mathbf{x}_1)^{-1}\mathbf{x}_1^t(I - S_2)^t(I - S_2)\mathbf{y}$ and $\hat{\mathbf{f}}_2 = S_2(\mathbf{y} - \mathbf{x}_1\hat{\beta})$. The latter approach can be motivated by partial regression methodology. Let $\mathbf{x}_{1\cdot 2} = (I - S_2)\mathbf{x}_1$ denote the residuals of $\mathbf{x}_1$ after smoothing by S_2. Thus $\mathbf{x}_{1\cdot 2}$ represents the information from $\mathbf{x}_1$ that is unique to that variable. If $\mathbf{y}_{\cdot 2} = (I - S_2)\mathbf{y}$, then $\hat{\beta} = (\mathbf{x}_{1\cdot 2}^t\mathbf{x}_{1\cdot 2})^{-1}\mathbf{x}_{1\cdot 2}^t\mathbf{y}_{\cdot 2}$, the estimate that would be obtained from the partial regression residual plot of $\mathbf{y}_{\cdot 2}$ on $\mathbf{x}_{1\cdot 2}$.

Surprisingly, the two estimators of the linear regression coefficient turn out to be different with respect to their bias. One of us [Speckman (1988)] has analyzed this case for kernel smoothers and has found that in situations where there is concurvity between $\mathbf{x}_1$ and $\mathbf{x}_2$, the partial regression estimator is generally root-n unbiased. In contrast, the estimator obtained by backfitting has a bias rate that is more typical of nonparametric estimators and is usually larger than root-n. [See also Rice (1986).] This can be of practical significance when inference about β is the goal.

In view of these results we wonder if the following approach might be useful in cases of high concurvity. Order the variables with respect to their importance as $x_1, x_2, \ldots, x_p$. Any linear terms should be last for reasons of bias as discussed above. First adjust $\mathbf{x}_j$ for $\mathbf{x}_1$ to obtain $\mathbf{x}_{j\cdot 1} = (I - S_1)\mathbf{x}_j$, $2 \le j \le p$, and adjust $\mathbf{y}$ for $\mathbf{x}_1$ using $\mathbf{y}_{\cdot 1} = (I - S_1)\mathbf{y}$. Now let $S_{2\cdot 1}$ be the matrix defined for smoothing on $\mathbf{x}_{2\cdot 1}$. The contribution to the overall fit from $\mathbf{x}_{2\cdot 1}$ will then be $\hat{\mathbf{f}}_{2\cdot 1} = S_{2\cdot 1}\mathbf{y}_{\cdot 1}$. For $3 \le j \le p$, let $\mathbf{x}_{j\cdot 12} = (I - S_{2\cdot 1})\mathbf{x}_{j\cdot 1}$, $\mathbf{y}_{\cdot 12} = (I - S_{2\cdot 1})\mathbf{y}_{\cdot 1}$ and define $\hat{\mathbf{f}}_{3\cdot 12} = S_{3\cdot 12}\mathbf{y}_{\cdot 12}$. Proceed in this fashion until all variables have been fit. If at some stage the assumed model is linear in $\mathbf{x}_{j+1}$ replace $S_{j+1\cdot 1\cdots j}$ by projection onto $\mathbf{x}_{j+1\cdot 1\cdots j}$. One could check the norm of the adjusted fit and if it is "sufficiently" small, the variable can be considered unimportant and dropped. The final smooth is $\hat{\mathbf{f}} = S_1\mathbf{y} + S_{2\cdot 1}\mathbf{y}_{\cdot 1} + S_{3\cdot 21}\mathbf{y}_{\cdot 12} + \cdots$.

This procedure can be motivated from a least squares on populations perspective where the variables $X_{(j+1)\cdot 12\cdots j} = X_{j+1} - E[X_{j+1}|X_1, \ldots, X_j]$, $1 \le j \le p - 1$, are used rather than the original ones. It can also be viewed as a type of projection pursuit where, instead of taking directions which are linear combinations of the independent variables, nearly orthogonal directions are used. These directions are constructed by a Gram–Schmidt process where smoothers are used

rather than projections. In the case where all variables enter the model linearly the method reduces to ordinary least squares.

One possible objection to the above approach not present with ordinary regression is that the fit in general depends on the order of the independent variables. This is not the case when all smoothers are projections. To compensate for this, one could use a stepwise method of adding variables, choosing the adjusted variable which decreases residual sum of squares the most at each stage. Such a method would probably be computationally intensive, but that should not be a deterrent in the future if the method has other merit. We would be interested in the authors' views on some of these proposals.

2. Inference and diagnostics. A key tool used by Buja, Hastie and Tibshirani in establishing many of their results is the relationship they establish between estimation in additive models and penalized least squares. In this section we point out another way this connection might be exploited.

We will utilize the same notation as in Section 4.4 of the paper with $\tilde{S}_i$, $i = 1, \ldots, p$, the smoothers that have had the projections removed and set $\tilde{S}_{p+1} = H$, the projection operator for $M_1(S_1) + \cdots + M_1(S_p)$. Let $\tilde{S}_i = U_{1i} D_{\theta_i}^{-1} U_{1i}^t$ with D_{θ_i} a diagonal matrix containing the nonzero eigenvalues of $\tilde{S}_i$ which for $i \leq p$ are all necessarily in $(0, 1)$. A solution to the normal equations is provided by the unique minimizer β^* of $\|\mathbf{y} - \sum_{i=1}^{p+1} U_{1i}\beta_i\|^2 + \sum_{i=1}^{p} \beta_i^t (D_{\theta_i}^{-1} - I)\beta_i$. Thus, β^* can be viewed as a Bayes estimator corresponding to the case where $\mathbf{y}$, conditional on β, is $N(U_1\beta, \sigma^2 I)$, with $U_1 = [U_{11}, \ldots, U_{1,p+1}]$, and β has a normal prior that is partially improper over the β's corresponding to $U_{1,p+1}$. Under this Bayesian model, one can show that $\operatorname{var}(\mathbf{f}_+|\mathbf{y}) \equiv \operatorname{var}(U_1\beta|\mathbf{y}) = \sigma^2 U_1 (U_1^t U_1 + \Delta)^{-1} U_1^t = \sigma^2 V$ with Δ a block diagonal matrix containing the $D_{\theta_i}^{-1} - I$ and a zero matrix. One can also show that the unconditional variance of the residual vector $\mathbf{y} - U_1\beta^*$ is $\sigma^2(I - V)$. As in Eubank and Gunst (1986) the forms of the conditional variance–covariance matrices for the fitted values and residuals under the Bayesian model are analogous to those from ordinary linear regression and therefore suggest that we might mimic techniques from linear regression when formulating methods for diagnostic and inferential analysis. Even though the Bayesian model may not be tenable here, one can still proceed as if it were and see how tools developed from the Bayesian framework perform in practice. The results from doing this in other related settings have been surprisingly good.

To illustrate the idea, consider the problem of constructing confidence intervals for the value of the true best additive approximation to the regression function at the ith design point. The Bayesian model might lead us to use $f_{+i} \pm 2\hat{\sigma}\sqrt{v_{ii}}$ with f_{+i} the ith element of the estimator $\mathbf{f}_+$, v_{ii} the ith diagonal element of V and $\hat{\sigma}^2$ some estimator of σ^2 such as the residual sum of squares divided by $\operatorname{tr}(I - V)$ [cf. Wahba (1983)]. Interval estimates involving the $\tilde{\mathbf{f}}_i$, etc., can be derived similarly. Results in Wahba (1983) and Nychka (1988) have shown that for ordinary spline smoothing this Bayesian approach to interval estimation tends to account for estimation bias in a certain sense. Whether this

would be true for additive models is an open question; however, the possibility merits investigation. The Bayesian framework suggests that we could also use similar diagnostic measures to those used in ordinary regression analysis: namely, the leverage values v_{ii} and Studentized residuals $(y_i - f_{+i})/\hat{\sigma}\sqrt{1 - v_{ii}}$. These can be combined in various fashions to obtain diagnostic tools including measures of influence. Some specific proposals can be found in Eubank and Gunst (1986).

We have conveniently ignored the computational aspects of the proposals in this section. Perhaps the authors can offer some insight into questions of this nature.

3. Another possible smoother. Projection-type smoothers belong to the collection of symmetric smoothers with eigenvalues in [0, 1] towards which many of the results in this paper are directed. The authors point out that for some of these types of estimator the effective dimension of the normal equations may be much less than np so that direct solutions are possible using standard regression software. We briefly mention here a "new" projection-type smoother which has this latter property and is of potential value for additive model estimation.

Recently, Eubank and Speckman (1988) have studied the properties of a smoothing method called polynomial-trigonometric regression (PTR). Assuming that all predictors take on values in [0, 1], a generalization of PTR to estimation for additive models is obtained by regressing $\mathbf{y}$ on polynomials of order d_i in the x_i and the functions $\{\sin(2\pi k x_i), \cos(2\pi k x_i);\ k = 1, \ldots, \lambda_i\}$ for $i = 1, \ldots, p$. Usually the d_i are treated as fixed with $d_i = 2$ a standard choice. The number of sine and cosine terms, λ_i, $i = 1, \ldots, p$, are then manipulated to govern the amount of smoothing.

Given choices of the d_i and λ_i, PTR smoothers can be readily computed using standard linear models software. They provide an alternative to polynomial regression smoothers that are more numerically stable and, in particular, avoid the need for specialized tools such as orthogonal polynomials. Although concurvity may still pose a problem, it will be manifest as linear dependencies among transformed predictors and can be handled in the usual regression fashion.

While PTR smoothers lack the flexibility of more sophisticated techniques such as smoothing splines they seem to give quite satisfactory results in many cases and get high marks for their implementational simplicity. A smoother with this latter quality will probably be required if additive model methodology is to be adopted by the masses in the near future.

We have been able to show that in the case of only one predictor PTR is capable of attaining mean squared error convergence rates that are comparable to those of smoothing spline and (boundary corrected) kernel estimators. Thus we are led to conjecture that results such as those in Stone (1985), with PTR smoothers used instead of regression splines, may hold for the PTR method as well. We pose this as another potentially interesting research problem.

4. Concluding remarks. Professors Buja, Hastie and Tibshirani have given a masterful account of a number of topics in additive model estimation. A

measure of the success of any article is provided not only by the number of important problems that it solves but also by the number of new questions that it opens for investigation. On the basis of both these criteria we must judge the present article to be a solid success.

REFERENCES

EUBANK, R. and GUNST, R. (1986). Diagnostics for penalized least-squares estimators. *Statist. Probab. Lett.* **4** 265–272.

EUBANK, R. and SPECKMAN, P. (1988). Curve fitting by polynomial-trigonometric regression. Unpublished manuscript.

NYCHKA, D. (1988). Bayesian confidence intervals for smoothing splines. *J. Amer. Statist. Assoc.* **83** 1134–1143.

RICE, J. (1986). Convergence rates for partially splined models. *Statist. Probab. Lett.* **4** 203–208.

SPECKMAN, P. (1988). Regression analysis for partially linear models. *J. Roy. Statist. Soc. Ser. B* **50** 413–436.

STONE, C. (1985). Additive regression and other nonparametric models. *Ann. Statist.* **13** 689–705.

WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

DEPARTMENT OF STATISTICS
TEXAS A & M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

DEPARTMENT OF STATISTICS
UNIVERSITY OF MISSOURI
COLUMBIA, MISSOURI 65211

WALTER GANDER AND GENE H. GOLUB

Eidgenössische Technische Hochschule Zürich and Stanford University

The solution of linear algebraic equations arises in many situations in statistical computing. Most often the matrices are symmetric and positive definite and they may have some structure that can be taken advantage of; viz., Toeplitz matrices arise in time series and special algorithms are available for such problems (cf. [3]). It is unusual for matrices to be structured and nonsymmetric but this is the situation that arises in the paper by Buja, Hastie and Tibshirani. In addition, the system (19) the authors describe is singular though the nullspace can be determined without difficulty.

Very often for large structured systems, iterative methods are used. (We set aside the fact that $\hat{P}$ is singular at this time.) Thus one might split $\hat{P}$ and write

$$\hat{P} = M - N$$

and iterate as follows:

Given f_0
For $k = 0, 1, \ldots,$

$$Mf^{k+1} = Nf^k + \hat{Q}y \quad (\text{solve for } f^{k+1}).$$

It is important that solving the system

$$Mf^{k+1} = z^k \quad (\text{say})$$