THOMPSON, A. M., BROWN, J. C., KAY, J. W. and TITTERINGTON, D. M. (1988). A study of methods of choosing the smoothing parameter in image restoration by regularization. Unpublished manuscript.

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

DEPARTMENT OF STATISTICS
UNIVERSITY OF GLASGOW
GLASGOW G12 8QQ
SCOTLAND
UNITED KINGDOM

REJOINDER

ANDREAS BUJA, TREVOR HASTIE AND ROBERT TIBSHIRANI

Bellcore, AT & T Bell Laboratories and University of Toronto

We thank the discussants for their interesting comments and contributions, and the editors and referees for considerable efforts that led to many improvements in this work. We must also thank the intrepid reader, if he or she is still with us, for weathering his or her way through this long article. The many questions given at the end of the paper and the ideas and issues raised by the discussants, indicate (happily) that this is an active area of research.

The discussants address a wide variety of issues in considerable detail. We try to address their comments and questions below. Before addressing each discussant in turn, we would like to present our views on several topics raised collectively by some.

1. The Bayesian paradigm. It seems that our silence about the Bayesian side of smoothing was so loud that it called for equally loud corrective measures from several discussants. *Cox*, *Kohn and Ansley*, *Chen*, *Gu and Wahba* and *Eubank and Speckman* remind us how useful the Bayesian paradigm can be for developing inferential procedures and algorithms. However, in the absence of a repeated sampling or subjective probability justification for the prior, the Bayesian framework is just a heuristic. In such cases, inferences derived from the Bayesian model must be justified through their sampling properties.

There are of course examples where the assumption of a random function has ample justification and where the prior represents a useful frequentist modeling assumption. This is usually called the stochastic process interpretation of the underlying function. For example, the Yates (1939) random effects model for incomplete block designs (we thank Dr. Peter Green for bringing our attention to this area) can be cast as a semiparametric regression model [Green (1985) and Green, Jennison and Seheult (1985)]. Here the "smoother" for fitting the random incomplete block effects is generated by a natural (noninformative) prior. More informative priors allow for spatial trends of various complexity. Wilkinson, Eckhert, Hancock and Mayo (1983) and the many discussants give a useful overview of this important area. If the assumption of an underlying random

process is sensible, Bayesian inferential methods have a frequentist meaning and they encompass the additional variability introduced by the randomness of the estimated function. In all other cases, we subscribe to the pragmatic Berger–Wahba philosophy [Wahba (1983), who refers to Berger; *Eubank and Speckman* agree]: "Derive confidence intervals based on some prior distribution, then forget the prior and see how well the intervals can be expected to perform on cases of interest." Given a useful and "natural" prior, the Bayesian formulation certainly has appeal: Posterior based estimation and inference are conceptually simple and automatic. How should we then interpret the role of the prior if it does not have a frequentist meaning? We comment on estimation and inference separately.

1.1. *Priors and estimation.* We will not fault the prior for its influence on the posterior mean, since it gives us exactly the same estimate as penalized least squares. At most, we will say that we do not need the prior in order to understand and to justify how a smoother behaves. The penalization terms $\lambda \int f''(x)^2\,dx$ and $\mathbf{f}^t(S^- - I)\mathbf{f}$ do not require a Bayesian interpretation to tell us in what directions shrinking will occur. An eigenanalysis of S will do just fine. Another matter is the creative use of the prior for algorithmic and analytic purposes. An example is of course the translation of the Bayesian framework into a state space representation as was done by Kimeldorf and Wahba (1971) and Wecker and Ansley (1983). The latter exploited this translation for the design of $O(n)$ algorithms for splines. A second and simpler example is the application of priors to additive models: If f_j has prior covariance $\sigma^2 K_j$, then $f_+ = \Sigma f_j$ has the prior covariance $\sigma^2 K_+ = \sigma^2 \Sigma K_j$, assuming the priors are uncorrelated with each other. Since $K_j = S_j(I - S_j)^{-1}$ and $S_+ = K_+(I + K_+)^{-1}$, we immediately obtain Proposition 3. This is mentioned by Cox, and it also follows from Gu, Bates, Chen and Wahba's (1988) rules for reproducing kernels of additive models. We learn from this that the prior covariance or reproducing kernel transform $K = S(I - S)^{-1}$ of a strictly shrinking smoother S is more natural than the smoother itself as far as combining components in additive models is concerned. This fact is also at the heart of the algorithms mentioned by *Chen, Gu and Wahba*, who use additivity of K-transforms to make computation independent of the number p of additive components. Another example for the role of the K-transform is mentioned below in the context of interaction smooths.

1.2. *Priors and inference.* In contrast to the posterior mean, the prior has a strong effect on the posterior covariance, and hence the inference derived from it. The posterior covariance is $\sigma^2 S$, whereas the covariance for $\hat{\mathbf{f}}$ based on a fixed function is $\sigma^2 SS^t$ (as in Section 2.7). We notice $S \geq SS^t$, which implies that the fixed function variance $\text{var}_\varepsilon \hat{\mathbf{f}} = \sigma^2(SS^t)_{ii}$ is smaller than the posterior variance $\text{var}_{\varepsilon, f} \hat{\mathbf{f}} = \sigma^2(S)_{ii}$. The difference is accounted for as follows. The mean square error matrix for $\hat{\mathbf{f}}$ is

$$E_\varepsilon(\hat{\mathbf{f}} - \mathbf{f})(\hat{\mathbf{f}} - \mathbf{f})^t = \sigma^2 SS^t + \mathbf{b}\mathbf{b}^t,$$

where the bias vector is $\mathbf{b} = E_\varepsilon \hat{\mathbf{f}} - \mathbf{f}$. This bias term depends on the unknown $\mathbf{f}$, but if we average it with respect to its prior, we get $E_{\varepsilon, f}\mathbf{b}\mathbf{b}^t = (I - S)K(I - S)^t = \sigma^2(S - SS^t)$. From a frequentist point of view, the Bayesian covariance represents an average mean squared error with regard to the prior.

Now as long as

1. the true functions are random,
2. the prior represents their variability well and
3. we choose to average over, rather than condition on, this variability

the Bayesian pointwise confidence regions are appropriate.

More often, however, the true functions are considered fixed. There are at least two routes that one can take in this instance. On the one hand, following the Berger–Wahba philosophy one may justify posterior inference, for example, with desirable fixed function frequentist properties. Wahba (1983) gives asymptotic evidence that posterior confidence intervals are just wide enough to account on the average for variance and bias simultaneously (if a suitable bandwidth is chosen). On the other hand, one may separate assessment into

1. diagnostics to detect bias problems and
2. inference based on standard errors which account for variance due to observational errors.

In this spirit, we showed in our paper standard error intervals based on the diagonal of the fixed effect covariance $\sigma^2 SS^t$, which are narrower than the posterior intervals based on $\sigma^2 S$. Diagnostics for bias are probably not yet developed to the necessary extent, but in comparison to traditional linear models smoothing technology faces a lesser bias problem in the first place. If automated bandwidth choice is used, the variance and bias aspects get traded off against each other in a (hopefully) near optimal fashion.

2. Smoothing parameters. A smoother that uses a data-driven choice for the smoothing parameter is nonlinear, and hence strictly speaking is beyond the scope of our paper. From a practical point of view, the selection of smoothing parameters remains an important question and has generated a large literature. We have studied linear smoothers not necessarily out of preference but out of convenience: It seems very difficult to prove convergence results for nonlinear smoothers. The hope, of course, is that the results for the linear setting will shed light on the nonlinear case. Therefore, we largely agree with those discussants (especially *Titterington*) who argue in favor of data-driven bandwidth choice.

2.1. *Fixed smoothing parameters.* Although our goal was not so much to develop an alternative fixed-bandwidth smoothing methodology, we see nothing wrong with fixing a number of degrees of freedom for each variable in an additive fit (*Breiman* and *Titterington* disagree). This is a way of allocating degrees of freedom to each variable, and allowing the smoother to use those degrees of freedom in a flexible way. It is a simple extension of linear model fitting. You

have to learn to walk (fixed smoothing parameters) before you can run (smoothing parameter selection). We admit that the translation of smoothing parameters into degrees of freedom is nontrivial for most smoothers. However, for Hastie's (1988) simplified splines, this translation is straightforward and computationally cheap.

2.2. *Automatic smoothing parameter selection.* Often we really do want to find out automatically how much smoothing should be done. In additive models where there are potentially more than one smoothing parameter, the problem encompasses model selection as well. The direct approach would be to minimize a global (generalized) cross-validation criterion over all of the smoothing parameters. This seems computationally formidable (see *Breiman's* challenge). There are several compromise approaches:

1. *Chen, Gu and Wahba* fix the ratios θ of the smoothing parameters for smoothing splines, and estimate a common multiplier λ by generalized cross-validation. Gu and Wahba (1988) use a Newton–Raphson search to find the $(p-1)$-dimensional θ. Their procedure still requires $O(n^3)$ computations, although it is claimed to converge fairly rapidly.

It seems intuitive that the GCV surface would vary most rapidly in the λ direction (overall amount of smoothing), in which case this parametrization is natural. We wonder if in fact it speeds up convergence. It also seems plausible that the relative factors θ might be poorly identifiable, especially in the presence of concurvity. There is a strong analogy to the estimation of variance components in random effects models (as noted by P. Green, personal communication), where similar identifiability problems exist. We offer a suggestion to alleviate this problem. Constrain the θ's in some sensible way, for example to be close to 1, and thus the relative amount of smoothing to be the same. This could be achieved by augmenting the GCV criterion with a ridge penalty of the form $\gamma\|\log(\theta)\|^2$, but would result in an additional *similarity* parameter γ. Other priors might also be considered, that give special weight to excluding variables entirely, or making the fit linear in particular variables (we have no suggestions on how to parametrize this).

2. Hastie (1988) describes a method for approximating the important eigenvectors of a smoothing spline (typically 8 or 10), and hence the smoother itself, in $O(n)$ operations. Given such an approximation for each term in the model, the entire additive model fit is simply a generalized ridge regression (as in Section 2 of *Eubank and Speckman*). Now we are in the same arena as *Chen, Gu and Wahba* above, except the cost per Newton–Raphson step is now only $O(n)$.

3. *Breiman* describes a backward selection procedure using fixed knot regression splines, an approach also taken by Friedman and Silverman (1989). The parameters are the number and positions of the "knots." Although we find this approach innovative, we have some reservations [see Hastie (1989)]. The procedure is akin to model selection in regression and shares one of its drawbacks when there are a large number of predictors: It is difficult to assess what you end

up with. Breiman emphasizes this point and proposes interesting ways to tackle it.

4. We have recently been using a backward stepwise procedure that is somewhat of a compromise between fixed smoothing parameters and global selection. Each variable can be either in the model with some fixed smoothing parameter (say span = 0.5 for a running-line smoother), fit linearly or absent from the model. Standard F-tests drive the selection procedure. We have found this to be quite useful, but have not yet studied its operating characteristics.

5. The previous stepwise approach for arbitrary smoothers can be quite slow. At each step an additive model is iterated till convergence. This suggests a further compromise. Start with the full model, with a nominal degrees of freedom for each term, and compute the GCV statistic using the approximation $\text{tr}(\mathbf{R}) = \sum_{j=1}^{p} \text{tr}(S_i)$. Attempt to "update" each term in the model by applying the appropriate smoother to the corresponding partial residuals, but selecting the smoothing parameter to minimize the (global) GCV criterion. In this step, one uses the usual univariate GCV criterion applied to the partial residuals, slightly modified to include the degrees of freedom for the other terms in the model. Having tried each of the p terms, incorporate the update corresponding to the minimum GCV. Continue until the criterion converges. A further modification to the smoothers allows the null fit (mean) or the linear fit to be candidates. This means that a valid step could be to remove a variable from a model, or make it linear. See Hastie (1989) for further details.

3. Interactions. In spite of the importance of the notion of interaction (pointed out by *Gasser and Kneip*), it is not a priori clear how it applies to smoother-based fitting methodology. We focus on two possible approaches to estimation: one by Breiman, and one by Barry (1983, 1986) and the Madison spline school [e.g., Wahba (1986)]. *Cox* also mentions the problem of inference for the presence of interactions.

As Breiman states, one could always estimate first-order interactions by a bivariate smooth of the residuals from an additive fit. We ourselves have tried this approach and found a perspective or contour plot of a bivariate kernel fit to be useful. Breiman proposes a more parsimonious estimate of interaction of the form $f(X_1) \cdot g(X_2)$, or $\sum_{j=1}^{J} f_j(X_1) \cdot g_j(X_2)$ if needed. We note a similarity between this idea for $J = 1$ and the thesis of Henry (1983). We find this approach attractive, especially in the case of data which require only one product term—the most interpretable situation. If two or more product terms are necessary, we wonder whether a contour plot of a bivariate fit is not more informative.

For greater parsimony, we may consider a Tukey 1-degree-of-freedom term of the form $\gamma \cdot h_1(X_1) \cdot h_2(X_2)$ fitted to the residuals $Y - \mu - h_1(X_1) - h_2(X_2)$ from the additive fit of $h_1(X_1)$ and $h_2(X_2)$. For higher-order interactions, general surfaces are difficult to interpret and display, and we do not know whether the generalization of Breiman's ideas to higher orders would be useful. A radical alternative would be to apply tree-based regression to the residuals [Breiman,

Friedman, Olshen and Stone (1984)]. This can reveal subgroups of observations which depart from additivity. Friedman's (1988) recent work on multivariate regression splines has a similar flavor.

A very coherent framework for interactions was initiated by Barry (1986) and continued by the Madison spline school [e.g., Wahba (1986)]. This framework offers the flexibility of ANOVA modeling in a semiparametric setup. In Barry's work it becomes obvious how powerful a heuristic the Bayesian approach can be: He assigns independent "natural" priors to the four components of an ANOVA decomposition $f(X_1, X_2) = \mu + \alpha(X_1) + \beta(X_2) + \gamma(X_1, X_2)$, and derives corresponding spline-type estimates for each component. Wahba (1986) develops this theory free of Bayesian connotations based on penalization and reproducing kernels alone. We can give a flavor of the techniques in terms of linear algebra as follows. One starts with a decomposition of a smoother into a projection and a strictly shrinking part. For example, a cubic spline decomposes into the projection onto $\mathbf{1}$ and $\mathbf{x}$, and a shrinker described, for example, by $\|\mathbf{y} - \mathbf{f}\|^2 + \lambda \mathbf{f}^t K^{-1} \mathbf{f} = \min_{\mathbf{f}}$. The problem then consists of defining interaction terms of three types:

1. projection $\times$ projection,
2. projection $\times$ shrinker and
3. shrinker $\times$ shrinker.

Type 1 is standard: The usual interaction of two projection terms $\mathbf{x}_1$ and $\mathbf{x}_2$ is another projection term $\mathbf{x}_1 \circ \mathbf{x}_2$, the coordinatewise (Schur) product of $\mathbf{x}_1$ and $\mathbf{x}_2$. A type 2 interaction between, say, a projection onto $\mathbf{x}_1$ and a shrinker given by K_2 is another shrinker defined by $\|\mathbf{y} - \mathbf{x}_1 \circ \mathbf{f}_2\|^2 + \lambda \mathbf{f}_2^t K_2^{-1} \mathbf{f}_2 = \min_{\mathbf{f}_2}$. It is not hard to see that this amounts to $\|\mathbf{y} - \mathbf{g}\|^2 + \lambda \mathbf{g}^t K^{-1} \mathbf{g} = \min_{\mathbf{g}}$, where K is given by the Schur product $K = (\mathbf{x}_1 \mathbf{x}_1^t) \circ K_2$ (if all $x_{1i} \neq 0$). Last, it remains to define an interaction between two shrinkers given by K_1 and K_2 (type 3). To this end we reduce the shrinkers to a ridge regression by letting $K_i = W_i W_i^t$, such that $\|\mathbf{y} - \mathbf{f}_i\|^2 + \mathbf{f}_i^t K_i^{-1} \mathbf{f}_i = \min_{\mathbf{f}_i}$, becomes $\|\mathbf{y} - W_i \mathbf{b}_i\|^2 + \|\mathbf{b}_i\|^2 = \min_{\mathbf{b}_i}$. It should be intuitive to pick the matrix $W = \{\mathbf{w}_{1k} \circ \mathbf{w}_{2l}\}$ whose columns are all possible Schur products of columns $\mathbf{w}_{1k}$ of W_1 with columns $\mathbf{w}_{2l}$ of W_2. The K matrix for interaction becomes $K = WW^t$ which is easily seen to be $K = K_1 \circ K_2$. This shows also that the result is independent of the factoring $K_i = W_i W_i^t$. The rules presented here for forming interactions are in agreement with the Gu, Bates, Chen and Wahba (1988) rules for reproducing kernel tensor products. Together with the rule for additive fits (sum up the K's), this new rule for interactions (form their Schur products) shows again that the K-matrices are more natural objects than the smoothers themselves for strictly shrinking smoothers.

When putting several interaction terms into a model, one can of course supply each shrinking term with its separate smoothing parameter. A search over all these parameters simultaneously may not be feasible, but for up to 4 or so, Gu and Wahba (1988) tell us there is promise. Once again, we might consider a finite search over several discrete values of the smoothing parameters, corresponding to 0 (absent), 1, 2, 3 and 4 degrees of freedom for each term, which could be achieved by the usual all subsets search techniques.

4. Replies to individual discussants. Some of the discussants (and many writers) argue for their favorite kind of smoother. We venture to say that in our experience with relatively noisy data, in most cases the choice is not too important in that differences between smoothers are small relative to the difference between a smooth and a parametric (say linear) fit. On the other hand, there may be important differences in some problems, and we need more studies like those of Breiman and Peters (1988) to help assess the relative merits of the various smoothing techniques.

We now focus on some of the individual points raised. *Breiman's* discussion was covered in Sections 2 and 3 of this rejoinder, as was *Cox's* in Section 1—we are sorry that space does not permit us to discuss their contributions further. One small point regarding Cox's test for lack of fit: We note that Cleveland and Devlin (1988) have also derived approximate (frequentist) F-tests for this problem.

Chen, Gu and Wahba develop the theory of penalized least squares for function estimation with associated reproducing kernel (rk) theory. They apply it to additive models, and show how the additivity rule for rk's results in an algorithm which is independent of the number of components fitted. Gu and Wahba (1988) then indicate how simultaneous minimization of a GCV criterion over several smoothing parameters may be feasible via Newton and Raphson if one takes advantage of several speedups in matrix decompositions. As Gu, Bates, Chen and Wahba (1988) state, this approach does not use the special structure, for example, of univariate splines, but it provides a general computational framework for interaction splines. If special structure is present, one might try a modification of backfitting where each step is interleaved with one step of smoothing parameter optimization for the currently active variable. We have not implemented such an algorithm and therefore do not know whether it has viable performance.

A noteworthy point is that the computation of splines presented by *Chen, Gu and Wahba* leads to a natural separation of projections and shrinkage components. In other words, the linear regression on $\mathcal{M}(S_1) + \mathcal{M}(S_2) + \cdots + \mathcal{M}(S_p)$ is pulled out and treated as a separate block S_0. This is similar to our modified backfitting algorithm. There seems to be a deeper necessity in this separation of projection and shrinkage.

Chen, Gu and Wahba also make some intriguing philosophical points regarding what is so special about splines. With reference to Stein's (1987, 1988) work, they argue that asymptotically only the equivalence class of a covariance kernel matters, and that splines are wonderful because their kernels are the most parsimonious members of their respective classes. We are pleased with these results, but because

1. they are asymptotic and
2. they rely on the stochastic process interpretation of the underlying function,

we still feel a desire for other—preferably cruder—evidence. A picture of a successful fit does as much to convince us that smoothing splines are indeed wonderful. An asymptotically equivalent but less parsimonious smoother will

either perform as well on finite samples—in which case it is as wonderful as a spline—or else it will be discarded for reasons of finite sample performance rather than its lack of parsimony.

Eubank and Speckman lay some groundwork with their remarks on the relation between concurvity and estimability. As they notice, the fit $\mathbf{f}_+ = \Sigma \mathbf{f}_j$ is indeed fully estimable as in ordinary least squares regression. This makes it a considerably more manageable object in comparison to the full vector of components $\mathbf{f}^t = (\mathbf{f}_1^t, \mathbf{f}_2^t, \ldots, \mathbf{f}_p^t)$. We agree that the real issue is not exact but approximate concurvity, and we regret that there was not more space to make this point more vigorously. Theorem 5 on the structure of exact concurvity by itself could be misleading and lull the user of additive models in unwarranted safety. We have to point out over and over again that smoother-based models do not only increase the flexibility of fits but the problems with near dependence among predictors as well.

Some of us [Buja, Donnell and Stuetzle (1986)] are pursuing a route for concurvity analysis which is analogous to the use of smallest principal components in collinearity for linear regression. This would amount to a diagnostic tool, but it also lends itself to a more symmetric analysis of variables where none is singled out as a response.

Eubank and Speckman propose some ideas based on partial regression in order to deal with the concurvity problem. There are several obvious questions one might raise: Stepwise forward partial regression with smoothers does not seem to optimize an obvious single criterion, it depends on the ordering of the predictors, and it does not seem to lend itself for easy interpretation as the resulting fit is no longer formulated in terms of the original predictors. On the other hand, we would have to see a real-life application, mainly to judge the interpretability of smoother-adjusted variables. We would expect the importance of successive terms $S_1\mathbf{y}$, $S_{2\cdot 1}\mathbf{y}_{\cdot 1}$, $S_{3\cdot 21}\mathbf{y}_{\cdot 12}, \ldots$ to decrease rapidly as the smoothers are based on increasingly noisy predictors.

We wait to see the Speckman (1988) paper on the alternative estimator (attributed to us by Denby) of the semiparametric regression model (39) in the text. The different bias properties of this and the penalized least-squares estimators are interesting; do the mean squared errors also differ in order? We note that for symmetric S_2 this is equivalent to using the "twicing" smoother in the usual semiparametric fit.

Eubank and Speckman propose leverage diagnostics based on posterior variances (diagonal of S), while Hastie (1988) does the same with fixed-function variances (diagonal of SS^t). As they and we agree in our attitudes toward Bayesian methods, we can leave a comparison of relative merits of the proposals to future investigations. As to the question of affordable computation of such diagnostics, the best answer we have is that the smoothers themselves have to be made affordable. This is one of the major reasons for developing fast spline approximations in Hastie (1988) to permit $O(n)$ computations.

The polynomial-trigonometric regression approach of Eubank and Speckman surely recommends itself for its simple implementation. We are looking forward to seeing some plots of PTR fits in a few situations of interest. Especially, we

would be curious to know how well PTR can handle local behavior without creating artifacts outside the area of interest. For example, can PTR avoid the problems of polynomial regression with strong but local curvature?

Gander and Golub explain the framework of iterative methods for linear systems as used in the numerical literature. We have not followed this formulation for didactic reasons, but the reader will easily recognize that $\hat{\mathbf{T}} = M^{-1}N$ for the Gauss–Seidel procedure. We are happy to learn that Golub and de Pillis (1988) have obtained optimal over-relaxation parameters in a situation which corresponds to the two-smoother case. Although they require what amounts to strictly shrinking smoothers, we are hopeful that their results throw light on the shrinking cases as well. Gander and Golub then proceed with an outline of ε-acceleration for iterations with slow convergence. This method is appealing since it is based on the same modularity as the Gauss–Seidel iterations: Only the basic modules for smoothing on each variable in turn are needed, and no matrix representation or other nonobvious form is required. ε-acceleration may give hope in many situations where componentwise solving or minimization performs unsatisfactorily. We wonder whether this method has been tried on mildly nonlinear problems as well, and if so with what results. We have just one question mark regarding the treatment of the null space of $\hat{\mathbf{P}}$: It is true that the theoretical nullspace is characterized by a linear dependency between the eigenspace for eigenvalue $+1$ of the smoothers and as such it is not hard to find. However, we have a problem with algorithms which force us to adjust for the presence of potential degeneracies as they may force us into giving up the modularity of Gauss and Seidel. It was the point of our convergence proof that for the Gauss–Seidel procedure no such adjustment is necessary. Of course, any such convergence result is somewhat dubious: Along directions of degeneracy, the Gauss–Seidel procedure does not move, and in directions of near degeneracy, it will move only slowly. Acceleration methods can do only so much: There is always the possibility of near degeneracy so close to exact, that for all practical purposes it is exact, yet unforeseen. Therefore, the statement that "the nullspace can be determined without difficulty" by Gander and Golub cannot be taken at face value. The example we give at the end of Section 3.7 should convince the reader that the possibility of near degeneracy is very real in actual data: The fact that flexible smoothers can approximate step functions quite well may turn them into cluster detectors in regression, but also produce "concurvity" due to multivariate clusters in predictor space. Problems of this sort are mentioned in the context of ACE and continuous correspondence analysis by Buja (1989). A forthcoming paper by Donnell, Buja and Stuetzle (1989) will deal with these issues in greater detail.

Gasser and Kneip mention an asymptotic connection between kernel, spline and k-nearest-neighbor smoothers in terms of a parametrization of design adaptation. This means that a smoother implicitly widens its bandwidth where the design is sparse, to a degree which depends on a parameter α. It is interesting to hear that k-NN smoothers often fare worse than either kernel or spline smoothers in Gasser and Kneip's empirical comparisons except in situations where design and response are favorably matched. We recall that the fixed-

bandwidth version of supersmoother [Friedman and Stuetzle (1982)] is a k-NN method.

Gasser and Kneip also hint at an ingenious $O(n)$ algorithm for kernel smoothers with polynomial kernels. As it stands, most important types of smoothers seem to have produced $O(n)$ implementations in the univariate case. The comparisons now are concerned with the proportionality constants. Some empirical results in this direction can be found in a recent report by Gasser, Koehler and Kneip (1988), where spline and kernel methods with automated bandwidth choice are compared.

Kohn and Ansley make a strong pitch for the stochastic process model in function estimation. We reiterate and clarify our previous objections: In adopting the assumption of random functions, one either takes a subjective Bayesian view of the function distribution as a prior—which we find unjustifiable in this situation—or else one commits a modeling error equivalent to a confusion of fixed and random effects in many practical situations. We do not deny that the time-series applications in economics considered for instance by Ansley and Wecker (1983), Section 3, call for a stochastic process assumption, but we fail to understand how this makes sense in physics and engineering data where the only variability of interest stems from measurement error. Even when the stochastic process assumption is adequate, one may still not want to account for the random effect variance but condition on the function and thus revert to the fixed effect model for the purposes of inference. In summary, we are unable to follow Kohn and Ansley if they maintain the notion that the random effects model will do in every situation. However, we see it as one of the positive outcomes of this discussion that it provided an opportunity to work out the conceptual problems in the choice of model assumptions.

It is apparent from this discussion that there have been some cross-connections in the literature; Kohn and Ansley point out some references to their work (and Wecker's) missed by us and other authors. We will note some additional references also missed by us and Kohn and Ansley. We address their points in the order encountered.

None of the smoothers we mentioned require evenly spaced data (and all can deal with tied x's). The idea of using state space representations for fast computations is fruitful, but there do exist other efficient $O(n)$ algorithms for computing smoothing splines, including the diagonal of S [see O'Sullivan (1985) and Woltring (1986)]. Which algorithms have numerical advantages would have to be investigated in a comparative study. The equivalence between smoothing by backfitting and penalized least squares can hardly be more immediate than the simple derivation in Section 3.3.2: Solving a linear algebra problem with the aid of stochastic processes is not necessarily the preferred method of proof. Kohn and Ansley claim as an advantage that the extrapolation to unobserved x-values is immediate, but this is the case under fixed function assumptions as well if the penalized LS problem is formulated in terms of function approximation in a suitable Sobolev space, an inherently finite-dimensional approach.

Friedman and Stuetzle (1981) introduced us to the backfitting algorithm for fitting the additive model, their so-called "projection selection" procedure.

Backfitting has appeared in the time-series literature where it is used in decomposition algorithms [X-11, Shiskin (1985) and Shiskin, Young and Musgrave (1967)], including the semiparametric model. Ansley and Wecker (1983) indeed refer to the semiparametric model, although we could not find equation (36) in that reference. We once again refer to the vast literature on the analysis of trends in field trials [Wilkinson, Eckhert, Hancock and Mayo (1983)], where semiparametric procedures were used as far back as 1937 (Papadakis)! We note, as was one by Green and Yandell (1985), that the expressions in equation (36) can be computed in $O(n)$ computations as long as the operation $S_2 z$ can be computed in $O(n)$; this is the case for most smoothers.

Wecker and Ansley (1982) devised the backfitting algorithm under the name of "alternating projection method." For convergence properties they refer to von Neumann's and Halperin's results, but no indication is given as to how they apply. Indeed, they do not, at least not in an immediate sense. The usual application of von Neumann–Halperin convergence results is to residuals. If the linear map S_j is a projection, one inner backfitting step of the form $\mathbf{f}_j \leftarrow S_j(\mathbf{y} - \Sigma_{k \neq j} \mathbf{f}_k)$ entails a residual update of the form $\mathbf{r} \leftarrow (I - S_j)\mathbf{r}$. One pass over all variables amounts to $\mathbf{r} \leftarrow (I - S_p) \cdots (I - S_1)\mathbf{r}$. Hence Halperin's theorems give us convergence on the residuals, but this is only the case if *all* smoothers are orthogonal projections. A less obvious but more powerful application is to the full vector of components $\mathbf{f}^t = (\mathbf{f}_1^t, \mathbf{f}_2^t, \ldots, \mathbf{f}_p^t)$, in which case one ends up with nonorthogonal idempotents denoted $\hat{\mathbf{T}}_j$ in our paper. These maps are "pseudo-orthogonal" projections with regard to the pseudoscalar product mentioned in the proof of Proposition 12. One could simply exclude the possibility of concurvity, in which case the pseudoscalar product becomes a true scalar product and Halperin's theorems give us convergence of backfitting. However, the point of our Theorem 8 is to avoid this assumption. Kohn and Ansley will have to come up with an equivalent technical fix, unless they are content with the weaker result. The virtue of allowing for concurvity is to show that intentionally or accidentally overparametrized models still lead to convergent backfitting.

Titterington discusses the choice of definition for degrees of freedom and the need for data-driven selection. We note that his method (of moments) for selecting λ rests on a Bayesian justification (and thus averages bias over the prior), whereas cross-validation does not. See Green (1985) for related method of moments techniques.

He is right about the nonconstancy of variance in the ozone concentration data. When the AVAS procedure [Tibshirani (1988)] is applied to these data, a cube root transformation of the response is suggested. We thank Titterington for the reference to Peters and Walker (1978a, b).

REFERENCES

ANSLEY, C. F. and WECKER, W. E. (1983). Extensions and examples of the signal extraction approach to regression. In *Applied Time Series Analysis of Economic Data* (A. Zellner, ed.) 181–192. Bureau of the Census, Washington.

BARRY, D. (1983). Nonparametric Bayesian regression. Ph.D. dissertation, Yale Univ.

BARRY, D. (1986). Nonparametric Bayesian regression. *Ann. Statist.* **14** 934–953.

BREIMAN, L. and PETERS, S. (1988). Comparing automatic bivariate smoothers. Technical Report, Dept. Statistics, Univ. California, Berkeley.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, Calif.

BUJA, A. (1989). Remarks on functional canonical variates, alternating least squares methods, and ACE. Technical Memorandum, Bellcore.

BUJA, A., DONNELL, D. and STUETZLE, W. (1986). Additive principal components. Technical Report, Dept. Statistics, Univ. Washington.

CLEVELAND, W. S. and DEVLIN, S. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *J. Amer. Statist. Assoc.* **83** 596–610.

DONNELL, D., BUJA, A. and STUETZLE, W. (1989). Additive principal components. Unpublished manuscript, Bellcore.

FRIEDMAN, J. H. (1988). Fitting functions to noisy data in high dimensions. Technical Report, Dept. Statistics, Stanford Univ.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modeling (with discussion). *Technometrics*. To appear.

FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

FRIEDMAN, J. H. and STUETZLE, W. (1982). Smoothing of scatterplots. Technical Report, Orion 3, Dept. Statistics, Stanford Univ.

GASSER, TH., KOEHLER, W. and KNEIP, A. (1988). A flexible and fast method for automatic smoothing and differentiation. Report 455, Univ. Heidelberg.

GOLUB, G. H. and DE PILLIS, J. (1988). Toward an effective two-parameter SOR method. Presented at the Conference on Iterative Methods for Large Linear Systems, October 19–21, 1988, Austin, Texas.

GREEN, P. J. (1985). Linear models for field trials, smoothing and cross-validation. *Biometrika* **72** 527–537.

GREEN, P. and YANDELL, B. (1985). Semi-parametric generalized linear models. *Generalized Linear Models. Lecture Notes in Statist.* **32** 44–55. Springer, Berlin.

GREEN, P., JENNISON, C. and SEHEULT, A. (1985). Analysis of field experiments by least squares smoothing. *J. Roy. Statist. Soc. Ser. B* **47** 299–315.

GU, C. and WAHBA, G. (1988). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. Technical Report 847, Dept. Statistics, Univ. Wisconsin, Madison.

GU, C., BATES, D. M., CHEN, Z. and WAHBA, G. (1988). The computation of GCV functions through Householder tridiagonalization with application to the fitting of interaction spline models. Technical Report 823, Dept. Statistics, Univ. Wisconsin, Madison.

HASTIE, T. (1988). Pseudo-smoothers and additive model approximations. Technical Memorandum, AT & T Bell Laboratories.

HASTIE, T. (1989). Discussion of "Flexible parsimonious smoothing and additive modeling" by J. H. Friedman and B. W. Silverman. *Technometrics*. To appear.

HENRY, D. (1983). Multiplicative projection pursuit. Ph.D. dissertation, Stanford Univ.

KIMELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95.

O'SULLIVAN, F. (1985). Discussion of "Some aspects of the spline smoothing approach to nonparametric regression curve fitting" by B. W. Silverman. *J. Roy. Statist. Soc. Sec. B* **47** 39–40.

PAPADAKIS, I. (1937). Methóde statistique pour des expériences sur champ. *Bull. Inst. Amél. Plantes á Salonique* **23**.

PETERS, B. C. and WALKER, H. F. (1978a). An iterative procedure for obtaining maximum likelihood estimates of the parameters for a mixture of normal distributions. *SIAM J. Appl. Math.* **35** 362–378.

PETERS, B. C. and WALKER, H. F. (1978b). The numerical evaluation of the maximum likelihood estimate of a subset of mixture proportions. *SIAM J. Appl. Math.* **35** 447–452.

SHISKIN, J. (1955). Seasonal computations of Univac. *Amer. Statist.* **9** 19–23.

SHISKIN, J., YOUNG, A. H. and MUSGRAVE, J. C. (1967). The X-11 variant of census method II seasonal adjustment program. Bureau of Census Technical Paper, U. S. Department of Commerce, Washington.

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

STEIN, M. L. (1987). Minimum norm quadratic estimation of spatial variograms. *J. Amer. Statist. Assoc.* **82** 765–772.

STEIN, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist.* **16** 55–63.

TIBSHIRANI, R. (1988). Estimating transformations for regression via additivity and variance stabilization. *J. Amer. Statist. Assoc.* **83** 394–405.

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

WAHBA, G. (1986). Partial and interaction splines for the semiparametric estimation of functions of several variables. Technical Report 784, Dept. Statistics, Univ. Wisconsin, Madison.

WECKER, W. E. and ANSLEY, C. F. (1982). Nonparametric multiple regression by the alternating projection method. *Proc. Bus. Econ. Statist. Sec.* 311–316. Amer. Statist. Assoc., Washington.

WECKER, W. E. and ANSLEY, C. F. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *J. Amer. Statist. Assoc.* **78** 81–89.

WILKINSON, G. N., ECKHERT, S. R., HANCOCK, T. W. and MAYO, O. (1983). Nearest neighbour (NN) anaylsis of field experiments (with discussion). *J. Roy. Statist. Soc. Ser. B* **45** 151–210.

WOLTRING, H. J. (1986). Fortran program "GVC on the netlib" public library (netlib research. att. com).

YATES, F. (1939). The recovery of inter-block information in variety trials arranged in three-dimensional lattices. *Ann. Eugenics* **9** 136–156.