2. Statistical problems like appropriate goodness-of-fit tests, confidence bounds and selection of important or elimination of unimportant covariates should be dealt with.

REFERENCES

GASSER, TH. and MÜLLER, H.-G. (1984). Estimating regression functions and their derivatives by the kernel method. *Scand. J. Statist.* **11** 171–185.

GASSER, TH., MÜLLER, H.-G. and MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. *J. Roy. Statist. Soc. Ser. B* **47** 238–252.

GASSER, TH., SROKA, L. and JENNEN-STEINMETZ, CH. (1986). Residual variance and residual pattern in nonlinear regression. *Biometrika* **73** 625–633.

GEMAN, S. and HWANG, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.* **10** 401–414.

JENNEN-STEINMETZ, CH. and GASSER, TH. (1988). A unifying approach to nonparametric regression estimation. *J. Amer. Statist. Assoc.* **83** 1084–1089.

KNEIP, A. and GASSER, TH. (1988). Convergence and consistency results for self modeling nonlinear regression. *Ann. Statist.* **16** 82–112.

MÜLLER, H.-G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231–239.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898–916.

ZENTRALINSTITUT FÜR SEELISCHE GESUNDHEIT
ABTEILUNG BIOSTATISTIK
POSTFACH 5970
D-6800 MANNHEIM 1
WEST GERMANY

INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
HEIDELBERG
WEST GERMANY

ROBERT KOHN AND CRAIG F. ANSLEY

University of New South Wales and University of Auckland

We would like to congratulate the authors for presenting us with such a broad overview of this important topic, and in particular on their proof of the convergence of the backfitting method. The linear smoother to which they have paid the most attention is the cubic spline smoother. Now smoothing splines can be represented as signal extraction estimates in a model where the unknown regression function is generated by a stochastic process. This allows us to take a model-based approach to smoothing and estimating the components of an additive model using smoothing splines, and in this comment we wish to contrast this approach with that of the authors. A model-based approach for estimating the additive components has much to commend it because: (i) All assumptions are stated explicitly. (ii) It is a comprehensive approach which is able to deal with a variety of problems including polynomial smoothing splines. (iii) Unlike ad hoc approaches such as running means and medians, the model-based approach can deal with unequally spaced data. (iv) It suggests reasonable ways of estimating unknown parameters either by maximum likelihood or Bayesian methods. (v) It provides a framework for doing statistical inference, that is, for setting confidence intervals for the unobserved components and the unknown

parameters. (vi) It suggests efficient algorithms for doing the required computations. (vii) For each model-based approach there is usually an equivalent penalized least-squares problem which provides some measure of how reasonable our model is.

1. Stochastic model for one-dimensional smoothing. Wahba (1978) showed that the solution to the penalized least-squares problem (4) in the paper could alternatively be obtained by signal extraction as

$$E\{f(t)|y(1),\ldots, y(n)\} \tag{1}$$

with

$$y(i) = f(t_i) + e(i), \tag{2}$$

where $f(t)$ generated by the stochastic differential equation $d^2f(t)/dt^2 = \sigma\lambda^{-1/2}\, dW(t)/dt$. The $e(i)$ are independent $N(0, \sigma^2)$, and independent of the Wiener process $W(t)$, and λ is the smoothing parameter. For simplicity, we ignore the complications caused by the initial conditions in the solution of the differential equation as they are not central to our discussion. Details are given in Kohn and Ansley (1988). The smoothness properties of the solution (1) are obtained directly for stochastic models by Kohn and Ansley (1983).

One immediate consequence of the model (2) is that we can apply results from the time-series filtering and smoothing literature to obtain very fast and accurate algorithms to compute both the optimal solution and Bayesian confidence intervals. These algorithms are based on a *state space representation* for the stochastic process $f(t)$ and the observations $y(i)$ in (2). Details are given by Kohn and Ansley (1987a) and Wecker and Ansley (1983), who apply the Kalman filter and the fixed point smoothing algorithms to find the optimal solution (1) and the Bayesian confidence intervals. In particular, our approach which is based on a continuous-time prior yields model-based confidence intervals at arguments lying between the observed values t_i. A later Bayesian approach proposed by Silverman (1985), page 13, gives the same confidence intervals at the arguments t_i but narrower ones in between these arguments, as shown by Kohn and Ansley (1988, pages 418–419).This is most easily seen when $e(i)$ is identically 0 in (1) as then Silverman's (1985) confidence intervals have zero length. Kohn and Ansley (1988), pages 415 and 416, also resolve the paradox mentioned by Wahba (1983) and Silverman (1985), page 13, who observe that the solution to the optimal smoothing problem is smoother than the sample paths generated by the prior and posterior distributions for $f(t)$.

Recently, a new smoothing algorithm for state space models was obtained by Kohn and Ansley (1987b, 1989) which is faster and numerically more stable than existing spline smoothing algorithms. It also enables the evaluation of the cross-validation and generalized cross-validation function without additional computation.

2. Additive models. We view the observations as a sum of zero mean independent Gaussian components. It will suffice to consider the three-

component model

$$y(i) = f_1(s_i) + f_2(t_i) + e(i), \qquad i = 1, \ldots, n, \tag{3}$$

with the third component being the residual $e(i)$ which we assume is independent $N(0, \sigma^2)$. The indices s_i and t_i belong to index sets $I^{(1)}$ and $I^{(2)}$, respectively, and each index set is either a subset of Euclidean space or an integer lattice in such a space. Typically, each of $I^{(1)}$ and $I^{(2)}$ will be a subset of the real line or the set of integers. Let Y be a vector of observations, and define $f_{1,\text{obs}} = \{f_1(s_1), \ldots, f_1(s_m)\}'$, $\hat{f}_1(s) = E\{f_1(s)|Y\}$, $\hat{f}_{1,\text{obs}} = E(f_{1,\text{obs}}|Y)$ and put $\text{var}(f_{1,\text{obs}}) = \sigma^2 V_1$ and $A_1 = V_1(I + V_1)^{-1}$. We define $f_{2,\text{obs}}$, $\hat{f}_2(t)$, $\hat{f}_{2,\text{obs}}$, V_2 and A_2 similarly for the second stochastic process $f_2(t)$. Then $\hat{f}_1(s)$ and $\hat{f}_2(t)$ are the best estimates of $f_1(s)$ and $f_2(t)$. By a double-conditioning argument we have that

$$\hat{f}_{1,\text{obs}} = E\{E[f_1(s)|Y, f_{2,\text{obs}}]|Y\} = A_1(Y - \hat{f}_{2,\text{obs}}). \tag{4}$$

Similarly we have that $\hat{f}_{2,\text{obs}} = A_2(Y - \hat{f}_{2,\text{obs}})$ giving the normal equations (19) in the paper and leading directly to the backfitting algorithm. We note that both A_1 and A_2 are symmetric positive-definite matrices with eigenvalues in the interval $(0, 1)$.

Suppose that conditionally on $f_{2,\text{obs}}$ we have an algorithm to compute $E[f(s)|Y, f_{2,\text{obs}}]$. Then we also have an efficient way to compute $\hat{f}_1(s)$ once we have $\hat{f}_{2,\text{obs}}$ because

$$\hat{f}_1(s) = E[f_1(s)|Y, f_{2,\text{obs}} = \hat{f}_{2,\text{obs}}]. \tag{5}$$

In particular, the backfitting algorithm allows us to compute $\hat{f}_1(s)$ for all $s \in I^{(1)}$ and not just $\hat{f}_{1,\text{obs}}$. More generally, if g is some functional of the $\{f_1(s)\}$ stochastic process, for example a derivative, then we can similarly obtain $\hat{g} = E(g|Y)$. We can deduce another important result from (5). Suppose that the stochastic process for $\{f_1(s)\}$ is generated by the linear stochastic differential equation, possibly more general than the second-order differential equation underlying cubic spline smoothing models. Then $\hat{f}_1(s)$ is a smooth function of s with its smoothness properties described in Kohn and Ansley (1983).

Kohn and Ansley (1988) show the equivalence between the stochastic model (3) and the multicomponent penalized least-squares problem (21) in the paper. Because the normal equations for backfitting follow directly from the stochastic model, as shown above, the equivalence between smoothing by backfitting and by penalized least squares is immediate.

The additive components model was investigated by Wecker and Ansley (1982) using a stochastic modeling algorithm, with the components estimated by splines, and estimation carried out by the backfitting algorithm, which Wecker and Ansley (1982) called alternating projection. This paper also gives practical examples and indicates a method of proof of the convergence of the backfitting algorithm based on von Neumann's (1950) alternating method and its extension by Halperin (1962), although details are not given. The method of proof depends on casting the penalized least-squares problem as a projection problem in Hilbert space as in Kohn and Ansley (1988). Details will be given elsewhere.

3. Smoothing, signal plus noise models and cross-validation. The stochastic models (2) and (3) are examples of a class of models known in the time-series literature as signal plus noise models. In general, we can write such models as $y(i) = f(i) + e(i)$ with $f(i)$ the signal, $e(i)$ an independent $N(0, \sigma^2)$ noise sequence which is also independent of $f(i)$. Put $f_{\text{obs}} = \{f(1), \ldots, f(n)\}'$, Y the observation vector, $\sigma^2 V = \text{var}(f_{\text{obs}})$ and $A = V(I + V)^{-1}$. Kohn and Ansley (1989) show that

$$\hat{f}_{\text{obs}} = E(f_{\text{obs}}|Y) = AY \quad \text{and} \quad \text{var}(f_{\text{obs}}|Y) = \sigma^2 A,$$

which implies that the influence matrix A is equal to the conditional variance of the signal (up to the scalar σ^2). Therefore, if we have an efficient method for computing $\text{var}[f(i)|Y]$, $i = 1, \ldots, n$, then we have an efficient method for computing the diagonal elements A_{ii} of A. In their paper the authors assume that the smoothing parameters are given. If they are not, as is usually the case, then they can be estimated by the cross-validation function CV and the generalized cross-validation function G defined, respectively, as

$$\text{CV} = \sum_{i=1}^{n} \{y(i) - \hat{f}(i)\}^2 \Big/ \sum_{i=1}^{n} \{1 - A_{ii}\}^2,$$

$$\text{G} = \sum_{i=1}^{n} \{y(i) - \hat{f}(i)\}^2 \Big/ \{1 - \text{tr}\, A/n\}^2.$$

Thus, if we use an algorithm based on the stochastic model to compute the conditional variances efficiently, then we can compute CV and G efficiently. In other words, cross-validation and Bayesian confidence intervals can be computed by the same algorithms.

4. Evaluation of the likelihood. Maximum likelihood, with the likelihood based on the stochastic model, is another way to estimate unknown parameters. This was done successfully by Wecker and Ansley (1983) for smoothing a single function, with a marginal likelihood approach proposed by Kohn and Ansley (1987a). The likelihood for the model (3) will be Gaussian with exponent $-\zeta(Y)/2\sigma^2$ and with denominator $\sigma^n(\det D)^{1/2}$, where $D = I + V_1 + V_2$ and $\zeta(Y) = Y'D^{-1}Y$. It is computationally expensive in general to evaluate $\det(D)$ but if we can evaluate $\hat{f}_{\text{obs}} = \hat{f}_{1,\text{obs}} + \hat{f}_{2,\text{obs}}$ in $O(n)$ operations, then we can evaluate $\zeta(Y)$ in $O(n)$ operations because $\zeta(y) = Y'(Y - \hat{f}_{\text{obs}})$. Our suggestion, used successfully in time series, is to estimate unknown parameters, with the exception of σ^2, by minimizing $\zeta(Y)$. If $f_2(t) = z(t)'\beta$, then the computation simplifies considerably. Then $p(Y)$, the density of Y, is given by $\int p(Y|\beta)p(\beta)\, d\beta$ so that $\zeta(Y) = Y'K'\{I - KZW^{-1}Z'K'\}KY$ and $D = Q(K'K)^{-1}W^{-1}$, where the matrices K and W are defined as $K'K = (I + V_1)^{-1}$ and $W = Q^{-1} + Z'K'KZ$, and Z is the $n \times r$ matrix with the ith row $z(t_i)'$. If $f_1(s)$ is generated by a state space model, then both $\zeta(Y)$ and $\det(D)$ can be evaluated in $O(n)$ operations using the Kalman filter.

5. Semiparametric models. In Sections 3.2 and 5.4 of the paper the authors discuss semiparametric models, which are additive models where one or more of the components is assumed to be linear. The case where all but one of the terms is assumed to be linear, often known as partial spline smoothing, was first discussed by Ansley and Wecker (1983) using a state space approach. This paper gave a number of examples and preceded all the papers on this topic referenced by the authors. The solutions for $\hat{\beta}$ and f_2^{∞} given in (36) of the paper are those appearing in Ansley and Wecker (1983), and discussed in more detail in Kohn and Ansley (1988). An efficient method for evaluating the estimate $\hat{\beta}$ is given by Kohn and Ansley (1985), extending the method used by Ansley and Wecker (1983) and Wecker and Ansley (1983).

More generally, suppose that in model (3), $f_2(t) = z(t)'\beta$, where $z(t)$ is $r \times 1$ vector of regressors and β has the prior distribution $N(0, \sigma^2 Q)$ with Q an $r \times r$ nonsingular matrix. Then, $p(\beta|Y) = p(Y|\beta)p(\beta)/p(Y)$, where $p(\beta|Y)$ is the conditional density of β given Y with the rest of densities defined similarly. Let $\hat{\beta} = E(\beta|Y)$ and define the matrices W, K and Z as in part 4 above. It follows with a little algebra that $\hat{\beta} = W^{-1}Z'K'KY$ and $\text{var}(\beta|Y) = \sigma^2 W^{-1}$. Therefore $\hat{f}_{1,\text{obs}} = A_1(Y - Z\hat{\beta})$, $\hat{f}_{2,\text{obs}} = Z\hat{\beta}$, $\text{var}(f_{2,\text{obs}}|Y) = Z\,\text{var}(\beta|Y)Z'$ and

$$\begin{aligned}\text{var}(f_{1,\text{obs}}|Y) &= \text{var}(f_{1,\text{obs}}|Y, \beta) + \text{var}\{E[f_{1,\text{obs}}|Y, \beta]|Y\} \\ &= \sigma^2 A_1 + A_1\,\text{var}(f_{2,\text{obs}}|Y)A_1'.\end{aligned}$$

We can similarly show that

$$\text{var}(f_{1,\text{obs}} + f_{2,\text{obs}}|Y) = \sigma^2 A_1 + \sigma^2(I - A_1)\text{var}(f_{2,\text{obs}}|Y)(I - A_1).$$

We deduce the important result that if $f_1(s)$ is generated by a state space model, then $\hat{f}_{1,\text{obs}}$, $\hat{f}_{2,\text{obs}}$ and for $i = 1, \ldots, n$, $\text{var}\{f_1(s_i)|Y\}$, $\text{var}\{f_2(t_i)|Y\}$, $\text{var}\{f_1(s_i) + f_2(t_i)|Y\}$, and the cross-validation functions CV and G can all be computed in $O(n)$ operations.

We note that if $Q = kI_r$ and we let $k \to \infty$, then β is diffuse and this is equivalent to taking β as a vector of parameters. The computation of W simplifies in the obvious way. The additive model $y(i) = f_1(s_i) + z(t_i)'\beta + e(i)$ is particularly important because: (i) We can use it to add independent regressors to a spline model. (ii) We can model one additive component as a spline and the other components as linear regressors or polynomial splines.

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.

HALPERIN, I. (1962). The product of projection operators. *Acta Sci. Math.* **23** 1962.

KOHN, R. and ANSLEY, C. F. (1983). On the smoothness properties of the best linear unbiased estimate of a stochastic process observed with noise. *Ann. Statist.* **11** 1011–1017.

KOHN, R. and ANSLEY, C. F. (1985). Efficient estimation and prediction in time series regression models. *Biometrika* **72** 694–697.

KOHN, R. and ANSLEY, C. F. (1987a). A new algorithm for spline smoothing based on smoothing a stochastic process. *SIAM J. Sci. Statist. Comput.* **8** 33–48.

KOHN, R. and ANSLEY, C. F. (1987b). A fast algorithm for smoothing, cross-validation and influence

in state space models. *Proc. Bus. Econ. Statist. Sec.* 106–113. Amer. Statist. Assoc., Washington.

KOHN, R. and ANSLEY, C. F. (1988). Equivalence between Bayesian smoothness priors and optimal smoothing for function estimation. In *Bayesian Analysis of Time Series and Dynamic Models* (J. Spall, ed.) 393–420. Dekker, New York.

KOHN, R. and ANSLEY, C. F. (1989). A fast algorithm for signal extraction, influence and cross-validation in state space models. *Biometrika* **76** 65–79.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1–52.

VON NEUMANN, J. (1950). *Functional Operators. Ann. Math. Studies* **2**. Princeton Univ. Press, Princeton, N.J.

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.

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

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.

AUSTRALIAN GRADUATE SCHOOL
OF MANAGEMENT
UNIVERSITY OF NEW SOUTH WALES
KENSINGTON 2033
NEW SOUTH WALES
AUSTRALIA

DEPARTMENT OF ACCOUNTING
AND FINANCE
UNIVERSITY OF AUCKLAND
PRIVATE BAG
AUCKLAND
NEW ZEALAND

D. M. TITTERINGTON

University of Glasgow

I am grateful to be granted the opportunity to comment on this interesting paper. It represents a synthesis of several smoothing techniques under one characterisation, it proposes a useful way of carrying out multiple regression that lies somewhere between multiple linear regression and the general additive models that underline ACE, and it investigates the properties of a practicable algorithm for obtaining the fit of the models to a set of data. There is much to discuss in the paper but, apart from a few brief comments and questions near the end, I should like to concentrate my remarks on a particular aspect, namely, the concept of degrees of freedom associated with the fitted models and the relationship with the choice of smoothing parameter.

I shall lead into my specific points by observing that, at first sight, the structure under consideration offers a variety of immediately applicable smoothing techniques, as indicated early on in Figure 2. However, a closer reading reveals that, if one is confronted with a particular set of data, the situation is not quite so straightforward. The authors remark that all their generic, linear techniques are characterised, in some guise, by a smoothing parameter. If, however, the choice of smoothing parameter is to be data-driven, then the linearity is lost. They are quite correct, of course, but unfortunately one finds repeatedly, in the literature, that the choice of a good smoothing parameter is considered to be a rather sensitive issue and that automatic, data-driven