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 methods of choice are strongly favoured. For the authors' results to have widespread practical application, therefore, extension to the case of data-driven choice of smoothing parameter is crucial.

At the end of Section 2.2 the authors promise "a linear method for fixing the degree of smoothing." However, the method, described in Section 2.7.3, involves adjusting the "degrees of freedom" associated with the fit to be equal to some prescribed number. The value "4" was used for degrees of freedom in their example. One then asks, "why 4?" Would one use the same value with other data sets? Surely not. In other words, if this method is adopted, I should expect that any sensible method of deciding on the value for "degrees of freedom" would have to be data-driven. For the case of the example in the paper, it would be interesting to know what happens if, say, cubic spline smoothing is used with λ chosen by generalized cross-validation, providing $\lambda = \hat{\lambda}$ and S-matrix $\hat{S}$, say. What values are then assumed by the three measures of "degrees of freedom"?

One of those measures, $\mathrm{tr}(S)$, has emerged from the literature on spline smoothing and on ridge regression as founded on the penalized least-squares formulation. For instance, Wahba (1983) suggests that $\mathrm{tr}(I - \hat{S})$—note that generalized cross-validatory choice is involved—be interpreted as equivalent degrees of freedom for *error*, and that

$$\mathrm{RSS}(\hat{\lambda})/\mathrm{tr}(I - \hat{S}) = \hat{\sigma}^2 \tag{1}$$

might be a reasonable estimator for the residual variance, σ^2. This is supported by an empirical study that forms part of Thompson, Brown, Kay and Titterington (1988). Under the guise of a one-dimensional image-restoration problem, a particular ridge regression problem was investigated in depth and $\hat{\sigma}^2$ was examined as an estimator for σ^2. Apart from the occasional "bad" $\hat{\lambda}$, leading to gross undersmoothing of the data, $\hat{\sigma}^2$ was very satisfactory in this respect. One can also turn (1) round. If σ^2 is known, or can be estimated consistently by $\tilde{\sigma}^2$, say, then solution of (1) provides a $\tilde{\lambda}$, say, that can be used as a data-driven choice of λ that should be comparable with cross-validatory choice but might not be so liable to producing "bad" values. (Such a $\tilde{\lambda}$ is called the equivalent degrees of freedom (EDF) choice for λ.) This is borne out by the empirical work of Thompson, Brown, Kay and Titterington (1988) and by Hall and Titterington (1987) who showed that, in a very simple regression problem and in the case of periodic spline smoothing, this latter method produces a λ that is the same order of magnitude as an optimal λ to which the cross-validatory $\hat{\lambda}$ is asymptotically equivalent.

Hall and Titterington (1987) also report a simulation study based on an example of periodic spline smoothing. Of relevance to the present discussion are the summary statistics (sample means and standard deviations) of the cross-validatory degrees of freedom $\mathrm{tr}(\hat{S})$. For each of six combinations of (n, σ), where n is the sample size, 100 replications were generated and the following results were obtained (standard deviations in brackets):

(n, σ)	(21, 0.1)	(21, 1.5)	(41, 0.1)	(41, 1.5)	(81, 0.1)	(81, 1.5)
$\mathrm{tr}(\hat{S})$	20.21(0.05)	14.71(2.45)	28.69(3.65)	15.39(4.2)	30.31(3.58)	16.58(5.17)

In general, the three measures for degrees of freedom will be similar when S is close to being orthogonal. If "optimal" smoothing is used, then this happens when the signal-to-noise ratio is large. Possibly the most trivial example, again taken from Hall and Titterington (1987), is that in which

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon},$$

$\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\operatorname{cov} \boldsymbol{\varepsilon} = \sigma^2 I$ and the penalised least-squares function is

$$\min_{\mathbf{f}} \{\|\mathbf{y} - \mathbf{f}\|^2 + \lambda\|\mathbf{f}\|^2\}.$$

Suppose an optimal λ is defined to be that for which

$$\mathbb{E}\|\mathbf{y} - \tilde{\mathbf{y}}\|^2$$

is minimised where $\mathbb{E}$ refers to the distribution of $\boldsymbol{\varepsilon}$. Since $\tilde{\mathbf{y}} = (1 + \lambda)^{-1}\mathbf{y}$, it turns out that $S = (1 + \lambda)^{-1}I$ and $\lambda = r^{-1}$, where $r = \mathbf{f}^T\mathbf{f}/(n\sigma^2)$, a signal-to-noise ratio. In this case

$$\operatorname{tr} S = n/(1 + \lambda) = n\{1 - \lambda + o(\lambda)\},$$

$$\operatorname{tr}(SS^T) = n/(1 + \lambda)^2 = n\{1 - 2\lambda + o(\lambda)\}$$

and

$$\operatorname{tr}(2S - S^TS) = n\{1 - \lambda^2 + o(\lambda^3)\}.$$

These are all equal, to order $O(1)$ if λ is small.

As promised, I conclude my comments with a few brief remarks.

(i) I am a little concerned that the analysis of the data in Figure 2 ignored what appears to be a marked nonconstancy of variance, and I wonder whether or not the ozone-concentration variable should have been transformed before fitting any curve.

(ii) A circulant approximation to the running-mean smoother matrix in Figure 1 would presumably lead to a theoretically more amenable method.

(iii) A version of the SOR method has been investigated by Peters and Walker (1978a, b), based on the EM algorithm for estimating parameters within models for finite mixtures. The range $1 < \omega < 2$ was crucial there in the context of convergence, and ω near 2 was best if the components of the mixture were not well separated.

REFERENCES

HALL, P. and TITTERINGTON, D. M. (1987). Common structure of techniques for choosing smoothing parameters in regression problems. *J. Roy. Statist. Soc. Ser. B* **49** 184–198.

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.

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