Comment

D. M. Titterington

1. INTRODUCTION

It is a pleasure to comment on the paper by Dr. O'Sullivan. The paper represents admirable blends of review and new ideas, together with theory and application.

It seems that the world is saturated with inverse problems. At least, I am continually being surprised to discover further manifestations of the general structure and, sometimes, substantively innovative developments. I was particularly grateful to discover the work of Backus and Gilbert and to learn about the notion and use of representers.

The paper, of course, discusses a particular class of inverse problems, those which are ill-posed. Perhaps the most surprising feature of the literature on this topic is the comparatively late stage at which statisticians have made an impact. After all, a major reason for the inherent difficulties is the existence of the random noise terms, ε_t, in the model, and we note that the ubiquitous prescription for estimation is of the ridge-regression type, so it is certainly appropriate territory for statisticians. I should like to base the bulk of my remarks on the theme of what particular contributions statisticians can make to the development of the area.

Before I launch into this, I should admit that, as the paper points out, other mathematical specialties are also essential to a full treatment of the problem. Particular areas are those of functional analysis, matrix theory (singular-value decomposition), and numerical methods for optimization. So far as the last topic is concerned, the paper has concentrated on *linear* or *linearized* problems, so that the optimality criterion is a quadratic function for which the minimizer can be written down explicitly. In other cases we are left with a nonquadratic criterion, which leads to the requirement of numerical methods; see, for instance, the use of simulated annealing in finding a regularized image restoration by Geman and Geman (1984).

2. THE IMPACT OF STATISTICIANS

In this section I shall follow the pattern of the paper in concentrating on linear problems. As a result, and with some apologies to the author for apparently trivializing his achievements, the problems are "solved" by ridge-regression estimators of which (3.1) is an example. Crucial features of the prescription are a matrix Ω_2 and a scalar, λ.

Depending on one's statistical leanings, (λ, Ω_2) have different interpretations. For a Bayesian, they are hyperparameters in a notional prior density and the ridge-regression estimator is itself interpretable as a posterior mode. This Bayesian basis has the advantage, in principle, of permitting the construction of confidence regions for the true quantities of interest. Of course, the validity of such regions is dependent on whether the notional prior is a meaningful one. So far as repeated sampling confidence statements are concerned, more work requires to be done on the lines of Wahba (1983b) to see to what extent Bayesian statements carry similar confidence values from a frequentist point of view.

Non-Bayesians interpret λ and Ω_2 somewhat differently. They regard Ω_2 as the kernel of a roughness penalty function, usually chosen to reflect some (admittedly "prior") ideas about the local smoothness of the underlying functions and/or to lead to tractable prescriptions for the regularized estimators, in the form of splines, for instance. If one can extrapolate from the literature about kernel-based density estimation, the choice of Ω_2 (cf., the choice of kernel function) should not be crucial to the performance of the resulting estimator, computational difficulties apart. Certainly, from the non-Bayesian point of view, no one Ω_2 seems sacrosanct. This last statement appears to conflict with the views of the adherents of maximum entropy regularization, who contend that, in a wide range of problems, a roughness penalty based on Shannon entropy is fundamentally special, an opinion I do not share (Titterington, 1984).

The other parameter, λ, called variously the smoothing, ridge, or regularization parameter, is the one to which the estimators should be more sensitive. Furthermore it is here that the statistical impact is most obvious. In principle there is no problem to the Bayesian, in that λ is a parameter of the prior which is, of course, known! To other statisticians, it is natural to base the choice of λ on some criterion of how close the estimator is to the true, on average. As a result, we obtain the mean squared error criteria of Section 5 and the associated databased versions such as cross-validatory choice, now familiar in several types of smoothing problems. It has required

D. M. Titterington is Titular Professor and Head of the Statistics Department, University of Glasgow, Glasgow G12 8QW, Scotland.

statistical input, of Professor Wahba in particular, to establish these apparently natural approaches. Given the linearity, with the resulting rank-one update formulae of Section 5, the GCV score is easily computed (see also Silverman, 1984).

In the nonstatistical literature, the choice of λ has been approached differently and, in general, less satisfactorily, as described in the next section.

3. ON THE CHOICE OF REGULARIZATION PARAMETER

A crucial instrument for the choice of λ is the residual sum of squares, RSS(λ), defined by

$$\mathrm{RSS}(\lambda) = m^{-1} \sum (z_i - \hat{z}_i)^2;$$

see Section 5.1 of the paper. RSS(λ) appears in the numerator of the GCV score and, in some contexts (see Wahba, 1983b, for instance), it is suggested that an estimator of the error variance, σ^2, can be obtained from

$$\mathrm{RSS}(\lambda) = \sigma^2\{1 - m^{-1}\mathrm{tr}\ H(\lambda)\}. \tag{1}$$

Conversely, if an estimator $\hat{\sigma}^2$ is available for σ^2, equation (1) provides another method of choosing λ. Hall and Titterington (1986b) show that, for a simple, ridge-regression problem, the method is, so far as λ is concerned, asymptotically equivalent to GCV.

The early workers in cross-validation, however, used $\hat{\sigma}^2$ to find $\lambda > 0$ such that

$$\mathrm{RSS}(\lambda) = \hat{\sigma}^2. \tag{2}$$

Comparison of (1) and (2) suggests that this method of choice leads to *over*smoothing, a fact established quantitatively, for some problems, by Hall and Titterington (1986b).

The heuristic motivation for using (2) is that, if the errors ε_i are independently, identically, and normally distributed, then

$$m^{-1} \sum_i (z_i - Ez_i)^2 \sim m^{-1}\sigma^2\chi^2(m)$$

so that

$$E\left\{m^{-1} \sum_i (z_i - Ez_i)^2\right\} = \sigma^2. \tag{3}$$

The two sides of (2) are then considered to be estimates of the two sides of (3).

The scale of differences in the degrees of smoothing resulting from (1) and (2) has been investigated by Hall and Titterington (1986a, 1986b) for particular inverse problems. As an example, consider a simple, ridge-regression structure in which the model is

$$\mathbf{z} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where $E(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{\varepsilon}) = \sigma^2 I$, and X is $m \times s$ of rank s. Suppose we assess the degree of smoothing by examining the values of the equivalent degrees of freedom for error,

$$\mathrm{EDF}(\lambda) = m - \mathrm{tr}\ H(\lambda).$$

To keep calculations simple, suppose $X'X = I$ $(s \times s)$, and define the signal to noise ratio to be $r = \boldsymbol{\beta}'\boldsymbol{\beta}/(s\sigma^2)$. Suppose we define λ_1 and λ_2 to be the solutions of

$$E\{\mathrm{RSS}(\lambda)\} = \sigma^2\{1 - m^{-1}\mathrm{tr}\ H(\lambda)\}$$

and

$$E\{\mathrm{RSS}(\lambda)\} = \sigma^2,$$

respectively. Then (Hall and Titterington, 1986b),

$$\mathrm{EDF}(\lambda_1) = m - sr/(1 + r)$$

and

$$\mathrm{EDF}(\lambda_2) = m - s/[1 + \{(1 + r)^{1/2} - 1\}^{-1}].$$

Some numerical values are given in the following table.

r	EDF(λ_1)	EDF(λ_2)
1	$m - 0.5s$	$m - 0.29s$
3	$m - 0.75s$	$m - 0.5s$
8	$m - 0.89s$	$m - 0.67s$

It is clear that the difference can be substantial, particularly in highly parameterized problems, that is, if s is of the same order of magnitude as m. It should be said, however, that more empirical work has to be done to confirm whether or not these differences are meaningful in practical terms.

4. ON THE VIRTUE OF SMOOTHING IN GENERAL

In this closing section, I should like to reemphasize the advantage that can be gained by regularization, thanks to the often spectacular reduction of the instability caused by noise. The advantage of methods such as cross-validation lies in their underlying methodological basis and theoretical properties. It would be of great interest to apply a cross-validatory-based smoothing rule to the inverse problems underlying positron emission tomography, as discussed by Vardi, Shepp and Kaufman (1985). (Their maximum likelihood solution corresponds to the unsmoothed version of any reasonable prescription.) In some problems there may also be scope for genuinely interpreting λ as a component of a prior density and estimating it accordingly (see Besag, 1986, for instance).

It seems to me that future research should place particular emphasis on the extension, to a wider range

of ill-posed problems, of loss-based methods for choosing smoothing parameters, supplemented by empirical checks that the resulting smoothed estimates are acceptable from a practical point of view. I look forward, in particular, to reading about the future exploits of the present author in this important area!

ADDITIONAL REFERENCES

BESAG, J. (1986). On the statistical analysis of dirty pictures (with discussion). To appear in *J. Roy. Statist. Soc. Ser. B.*

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Machine Intell.* **PAMI-6** 721–741.

HALL, P. and TITTERINGTON, D. M. (1986a). On some smoothing techniques used in image processing. To appear in *J. Roy. Statist. Soc. Ser. B.*

HALL, P. and TITTERINGTON, D. M. (1986b). Common structure of techniques for choosing smoothing parameters in regression problems. Revised manuscript in preparation.

SILVERMAN, B. W. (1984). A fast and efficient cross-validation method for smoothing parameter choice in spline regression. *J. Amer. Statist. Assoc.* **79** 584–589.

TITTERINGTON, D. M. (1984). The maximum entropy method for data analysis (with reply). *Nature* **312** 381–382.

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

Comment

Grace Wahba

Professor O'Sullivan has given us a nice overview of some of the issues in ill-posed inverse problems as well as some new ideas. The most important of these new ideas I believe are the following: a) the extension of the idea of averaging kernel to reproducing kernel spaces, with the resulting formula

$$\operatorname*{supremum}_{\|\theta\|^2 \le \mu^2} |\theta(t) - E\hat{\theta}(t)|^2 = \|e_t - A(t)\|^2 \mu^2$$

and b) a new approach to the history matching problem of reservoir engineering. The formula bears a not coincidental relationship to Scheffé's S method of multiple comparisons (Scheffé, 1959, page 65). In atmospheric sciences and possibly elsewhere, extensive historical data allows the construction of a prior covariance for the unknown θ, from which reasonable norms can often be constructed via the well known duality between prior covariances and optimization problems in reproducing kernel spaces. An example of the use of prior covariances based on historical meteorological data to establish penalty functions can be found in Wahba (1982a). The problems of reservoir engineering are extremely important and would benefit from the attention of statisticians. Letting

$$z_{ij} = u(x_i, t_j, a) + \varepsilon_i,$$

as in Section 4.2, the method of regularization estimate of a is the minimizer of

$$\frac{1}{n} \sum_{ij} (z_{ij} - u(x_i, t_j, a))^2 + \lambda J(a)$$

(see especially Kravaris and Seinfeld, 1985). This problem is particularly difficult since, not only is u a nonlinear function of a, but in general the relationship is only known implicitly as the solution to a partial differential equation. It is a good conjecture that the GCV for nonlinear problems as proposed in O'Sullivan and Wahba (1985) can be used to choose λ in this problem. The details are far from obvious but it looks like the present paper provides an important first step. Of course this history matching setup leads to some juicy experimental design problems—choice of the forcing function q, the location of the wells, and the times of observation.

Concerning robustness of the PMSE criteria (that is, minimizing PMSE also tends to minimize other, possibly more interesting loss functions), further remarks on that can be found in Wahba (1985, page 1381). The GCV extension proposed by the author is an interesting one. Let C be the matrix with ijth entry $c_i' c_j$. If C is the identity then the extension is the same as GCV. If C is a well conditioned matrix, then it appears that one can show that the minimizer of $EV(\lambda)$ is asymptotically near the minimizer of $EL(\lambda)$, the associated (estimable) loss function. You need $(1/T)\text{tr } HC$ to be small near the minimizer of EL. I think a problem may arise if you try to choose C to approximate $L(\lambda)$ of the form

$$L(\lambda) = \frac{1}{T} \sum_{i=1}^{T} |\theta(t_i) - \hat{\theta}_\lambda(t_i)|^2$$

where the problem is very ill-posed. Consider the operator which maps θ to euclidean m space via the formula $\theta - (\eta(x_1, \theta), \ldots, \eta(x_m, \theta))$. In practice the theoretical dimension of the range space of this

Grace Wahba is Professor, Department of Statistics, 1210 West Dayton Street, Madison, Wisconsin 53706-1693.