flat regions with occasional spikes. It might be useful to develop smoothing functionals J that mirror this.

Regarding the empirical selection of smoothing parameters, Rice (1986) sounds a cautionary note by constructing simple examples in which a choice of smoothing parameter giving a good value of predictive mean square error gives unacceptable errors for estimating θ and vice versa.

ADDITIONAL REFERENCES

RICE, J. (1986). Choice of smoothing parameter in deconvolution problems. *Contemp. Math.* **59** 137–151.

WAHBA, G. (1982). Constrained regularization of ill-posed linear operator equations, with applications in meteorology and medicine. In *Statistical Decision Theory and Related Topics III.* (S. Gupta and J. Berger, eds.) **2** 383–418. Academic, New York.

Comment

Freeman Gilbert

In a typical geophysical inverse problem one has

$$(1) \qquad d_j = D_j(f) + r_j\sigma_j, \qquad j \in \{1, \ldots, J\},$$

where

- d is a datum,
- D is the functional that maps f into d,
- f is the model,
- r is a unit variance random variable,
- σ is the assigned error, usually taken to be the standard deviation (Gaussian errors).

An error statistic is introduced, usually the χ^2 statistic

$$(2) \qquad \chi^2(f) = \sum_j [d_j - D_j(f)]^2/\sigma_j.$$

One defines the set

$$\{F_0(f)\text{: all } f \text{ such that } \chi^2(f) \leq X_0^2\},$$

where χ_0^2 is chosen to be the 99% or 95% confidence level, for example.

Except in very unusual circumstances, $F_0(f)$ is either empty or infinite-dimensional. In the former case, one increases χ_0^2 and seeks to fill $F_0(f)$. In the latter case, one desires to know about the members of $F_0(f)$.

One procedure is to use the method of regularization (MOR) to find a particular member of $F_0(f)$ (e.g., the smallest, the smoothest, the one closest to a particular f_0, the maximum entropy solution, $\max\{-f\log f\}$, etc.). Another procedure is to use a resolution method to find what features all f have in common or what are the resolvable averages of f. In any case one may wish to assert a priori conditions on f, such as prejudices about the shape or size of f that can be cast in the form of equation (1).

O'Sullivan has shown that the two procedures are connected and, taken together, can lead to improved methods of estimating bias. By generalizing the concept of averaging kernel, i.e., requiring the averaging kernel to assume certain shapes, one can estimate average bias as well as local bias. For linear problems, the matter appears to be resolved and depends only on the number and quality of the data and the span of their representers. For nonlinear problems one is confined to the neighborhood of the subject. O'Sullivan is to be congratulated for his original contribution to it.

Freeman Gilbert is Professor of Geophysics in the Institute of Geophysics and Planetary Physics, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California 92093.

Rejoinder

Finbarr O'Sullivan

It is a pleasure to thank the discussants Professors Gilbert, Rice, Titterington, and Wahba for their most interesting and stimulating comments. The ubiquity of inverse problems in areas like geophysics, medical imaging, and meteorology presents statisticians with wonderful opportunities to contribute to the development of science and technology. As Professor Wahba notes there are lots of open research questions many of which are at the interface of numerical analysis and statistics. It is an especially exciting time to be a statistician who computes.

I will begin the rejoinder by making some comments about the role of the singular value decomposition (SVD) in the analysis of retrieval characteristics of linear and nonlinear inversion methods. This is followed by some briefer remarks about priors, the selection of smoothing parameters, and the hat matrix.

1. RETRIEVAL CHARACTERISTICS AND THE SVD

The SVD is a familiar object in inverse problems (see Andrews and Hunt, 1977; Cullum, 1980, for example). The analysis mentioned by Rice reveals the nature of information visible in the observed data. Wahba (1980) has used this to define the degree of ill-posedness of an inverse problem. Cullum (1979) uses the SVD in coming up with guidelines for the choice of norm in the method of regularization (see also Nychka et al., 1984). SVD of the averaging kernel is useful in understanding retrieval characteristics.

1.1 Linear Inversion

A linear inversion method can be decomposed as

$$\hat{\theta} = S\mathbf{z} = E\hat{\theta} + \delta \tag{1}$$

where $E\hat{\theta}$ is the systematic component of $\hat{\theta}$ and δ is the random component. From the averaging kernel calculus, we have that $E\hat{\theta} = A\theta$ with $A = [X'X + m\lambda\Omega_2]^{-1}X'X$. The random component is $[X'X + m\lambda\Omega_2]^{-1}X'\varepsilon$. (This notation is taken from Section 3 in the paper.) The averaging kernel operator, A, identifies the parts of Θ which are best resolved by the inversion procedure.

Rice refers to the standard linear model, here $\lambda = 0$ and when $X'X$ is invertible, A is the identity map. This corresponds to the statement that least squares is unbiased in this situation. For unbiased estimators the Backus-Gilbert averaging kernels are uninteresting being given by rows (or columns) of the identity map. (This is not to say that the hat matrix $X[X'X]^{-1}X'$ or more generally $X[X'X + m\lambda\Omega_2]^{-1}X'$ is not important. It surely is but not as a means of studying bias. I'll come back to this later.) An important part of linear model theory is concerned with fractionated designs for which there is already a well developed understanding of bias; aliasing patterns, design resolution, and so forth. Unfortunately there is too little attention paid to this in many modern linear model texts.

Let $\{(\mu_\nu, \phi_\nu), \nu = 1, 2, 3, \ldots\}$ be the eigenvalues (arranged in decreasing order) and corresponding normalized eigenvectors obtained by principal component analysis of the range space of A. These are obtained from the singular values and right singular vectors of $X'[X'X + m\lambda\Omega_2]^{-1/2}$. Expanding elements of Θ in terms of the eigenvectors ($\theta = \sum_\nu \theta_\nu\phi_\nu$ where θ_ν is the projection of θ into ϕ_ν) we have that

$$A\theta = \sum_{\nu} \mu_\nu \theta \phi_\nu. \tag{2}$$

The eigenvalues μ_ν are bounded above by 1 and the more poorly resolved features will be associated with smaller eigenvalues. Variability in the estimation of θ_ν is

$$\sigma_\nu^2 = \sigma^2 \| X[X'X + m\lambda\Omega_2]^{-1}\phi_\nu \|_m^2. \tag{3}$$

From this analysis an interesting collection of plots would include eigenvectors corresponding to the larger eigenvalues, μ_ν versus ν (resolution) and σ_ν versus ν (standard error). Figure 1 gives the eigenvectors corresponding to the first six eigenvalues in the tumor size distribution problem. The resolution (μ_ν) and standard error (σ_ν) are given in Figure 2. Eigenvectors tend to have more detailed structure toward the right-hand side of the interval.This is consistent with the results obtained in Section 2 of the paper. The eigenvectors $\{\phi_\nu, \nu = 1, 2, 3, \ldots\}$ are norm-dependent and the Euclidean norm analysis given above must be modified to produce eigenvectors corresponding to L_2 or Sobolev norm. Figure 1 actually corresponds to the L_2 norm.

Comparing the SVD of the averaging kernel operator A for different choices of the smoothing parameter leads to a way of understanding the effect of the smoothing parameter on resolution characteristics. A canonical correlation analysis (see Greenacre, 1984 or Mardia, Kent, and Bibby, 1979) might be useful here.

1.2 Nonlinear Inversion

Extending the above methods to nonlinear problems leads to some fascinating questions. A general MOR (method of regularization) inversion method is defined by

$$\begin{aligned}\hat{\theta} &= S(data)\\ &= \underset{\theta\in C}{\operatorname{argmin}}\{l_m(data \mid \theta) + \lambda J(\theta)\} \quad \lambda > 0.\end{aligned} \tag{4}$$

$l_m(data \mid \theta)$ measures the plausibility of the observed data if the true function was θ, this might be a residual sum of squares, a negative log likelihood, or a general distance measure. The penalty functional J measures the prior plausibility of θ, this could be quadratic or nonquadratic as is the case in maximum entropy methods. The regularization parameter λ adjusts the influence of the penalty functional. The set C is a subset of the parameter space Θ representing possible constraints such as positivity, monotonicity, etc.

Obviously, inversion characteristics cannot be analyzed without some topological structure, i.e., one

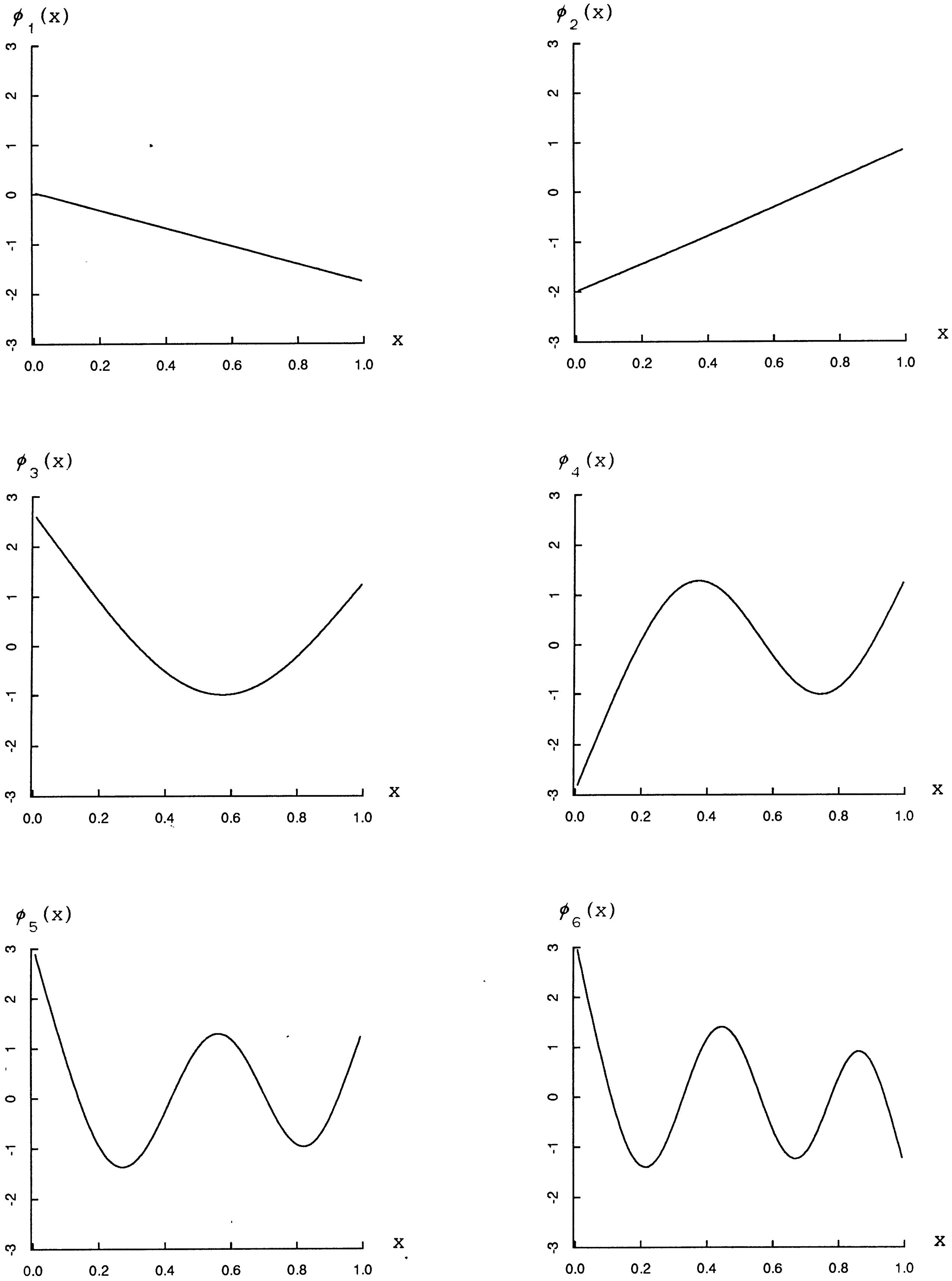

FIG. 1. *The first six eigenvectors* (ϕ_ν, $\nu = 1, 2, \ldots, 6$) *of the averaging kernel operator in the tumor size distribution problem. Eigenvectors tend to have more detailed structure toward the right-hand side of the interval.*

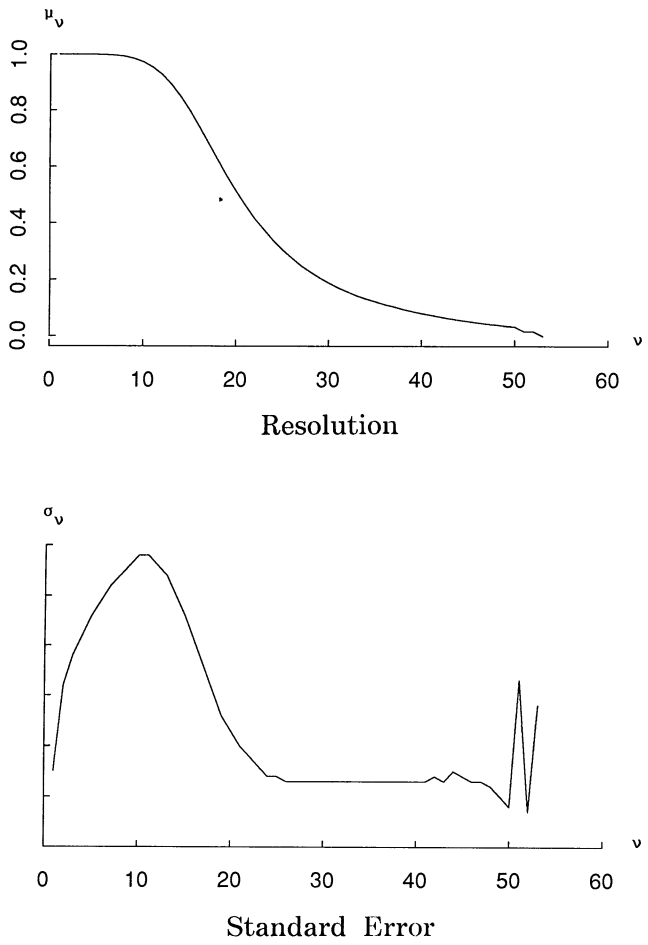

FIG. 2. *Resolution (μ_ν versus ν) and standard error (σ_ν versus ν). These characteristics are obtained from an SVD of the averaging kernel operator. The vertical scale on the standard error plot depends on the assumed noise level σ so it is left unspecified.*

needs to be able to evaluate when elements of the parameter space are close. In image restoration contexts (Besag, 1986) the definition of a meaningful distance can be a difficult problem.

The range of the nonlinear inversion method in (4) is some subset of C and in studying retrieval characteristics the local and global geometry of this set (both for noise-free and noise-contaminated data) are of interest. Modern computing environments make it possible to repeatedly solve (4) thereby generating elements in the range space. We need to identify procedures (algorithms, simulation experiments) which will generate useful information about the range of inversion in a small number of trials. Lanczos algorithms (Cullum and Willoughby, 1985; Golub and Van Loan, 1983; Parlett, 1980) are based on Krylov information. Starting with some initial vector $\theta \in C$, a Krylov sequence is $\{\theta, M(\theta), M(M(\theta)), M(M(M(\theta))), \ldots\}$. It would be interesting to know something about the ability of Krylov sequences to sample range spaces in nonlinear situations.

Further problems will arise in describing the geometry (local and global) of range spaces from sampled information. Part of this will involve generalizations of techniques like factor analysis, multidimensional scaling, and principal components. Some recent work in these areas is discussed in Hastie (1984) and Koyak (1985) and the references cited therein.

2. PRIORS

All discussants touched upon the role of prior information. Cullum (1979) has an interesting discussion of the choice of norm in regularization (the situation is more sensitive than Titterington suggests). Titterington and Wahba highlight the relationship between quadratic MOR penalty functionals and Gaussian prior distributions for the signal. In the paper I concentrated on situations where the quantity of interest was a smooth unconstrained curve. An interesting approach to the estimation of nonsmooth signals has been proposed by Kitagawa (1986). The method, which is shown to work very well, involves replacing Gaussian priors by longer tailed alternatives. Given the existence of boundary layer phenomena in the atmosphere I wonder if historical weather records are more accurately modeled by nonGaussian distributions? Atmospheric temperature profiles can have sharp changes in the first derivative so nonGaussian priors might lead to more satisfactory inversion methods.

3. SELECTION OF SMOOTHING PARAMETERS

The selection of smoothing parameters has received considerable attention in the statistical literature. There are several theoretical results (many of which are due to the discussants) describing how cross-validation or unbiased risk methods provide good choices for smoothing parameters, at least from the point of view of the predictive mean square error (PMSE).

The power of the PMSE to discriminate between different choices for the smoothing parameter is clearly limited by the span of the representers. The more ill-posed the problem the smaller the effective rank of the representers (see Wahba, 1980) and the more difficult it is to discriminate between different solutions on the basis of PMSE. The examples in Rice (1986) are surely a good illustration of this phenomenon.

Wahba notes that transformations to other loss functions cannot change the basic fact that data-based selection of smoothing parameters is limited by the information (visible) in the data (representers). I agree; estimable losses can only serve to redistribute the weight attached to these representers. In particular, if the matrix C is unstable in the sense described by Wahba then one would imagine that data-dependent estimates of associated loss though relatively unbiased will have very high variance. There is a bias-variance tradeoff here.

Titterington's comparison of the methods he denotes by (1) and (2) is most interesting. The class of estimation criteria suggested by Gilbert are motivated by a philosophy similar to (2) and an illuminating discussion of this appears in Titterington (1985). While selection procedures based on (2) are biased (a tendency to oversmooth) the practical significance of this bias needs to be understood. For smooth inversion methods the computation of cross-validatory scores is usually not too problematic. However many of the algorithms used in image restoration and pattern recognition (see Besag, 1986; Geman and Geman, 1984; Mendel, 1983) are such that it is hard to identify an efficient way of implementing cross-validation. Alternative techniques are needed. Selection rules based on methods like (2) may prove to be very convenient.

4. THE HAT MATRIX

The importance of the hat matrix in nonparametric regression is emphasized by Eubank (1984). The rows (or columns) of the hat matrix constitute the equivalent kernel treated by Silverman (1984). A time series analogy pinpoints the distinction between the averaging kernel and the equivalent kernel: transfer functions built up from response characteristics to impulse patterns in the *signal process* generate the *averaging kernel*; transfer functions built up from response characteristics to impluse patterns in the *raw data* generate the *equivalent kernel.*

The hat matrix continues to be important in general inverse problems. For example, generalized cross-validation and unbiased risk procedures treat all points in the design space equally. This may or may not be a good thing. The sensitivity of cross-validation to remote points in the design space seems to be worth investigating. These points are identified by large diagonal elements in the hat matrix (leverage values).

ACKNOWLEDGMENT

This research was supported by the National Science Foundation under Grant MCS-84-03239.

ADDITIONAL REFERENCES

ANDREWS, H. C. and HUNT, B. R. (1977). *Digital Image Restoration.* Prentice-Hall, Englewood Cliffs, N. J.

BESAG, J. (1986). On the statistical analysis of dirty pictures (with discussion). To appear in *J. R. Statist. Soc. Ser. B* **48**.

CULLUM, J. (1979). The effective choice of the smoothing norm in regularization. *Math. Comp.* **33** 149–170.

CULLUM, J. (1980). Ill-posed deconvolutions: regularization and singular value decompositions. *Proc. IEEE Conf. Decision Contr.* **19** 29–35.

CULLUM, J. K. and WILLOUGHBY, R. A. (1985). *Lanczos Algorithms for Large Symmetric Eigenvalue Computations Col. 1 Theory.* Birkhäuser, Boston, Mass.

EUBANK, R. L. (1984). The hat matrix for smoothing splines. *Statist. Probab. Lett.* **2** 9–14.

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

GOLUB, G. H. and VAN LOAN, CH. F. (1983). *Matrix Computations.* North Oxford Academic, New York.

GREENACRE, M. J. (1984). *Theory and Applications of Correspondence Analysis.* Academic, London.

HASTIE, T. J. (1984). Principal curves and surfaces. Ph.D. dissertation, Dept. Statist., Stanford Univ.

KITAGAWA, G. (1986). Non-Gaussian state space modeling of nonstationary time series (with discussion). To appear in *J. Amer. Statist. Assoc.*

KOYAK, R. (1985). Optimal transformations for multivariate linear reduction analysis. Ph.D. dissertation, Dept. Statist., Univ. California, Berkeley.

MARDIA, K. V., KENT, J. T. and BIBBY, J. M. (1979). *Multivariate Analysis.* Academic, New York.

MENDEL, J. M. (1983). *Optimal Seismic Deconvolution: An Estimation-Based Approach.* Academic, New York.

PARLETT, B. N. (1980). *The Symmetric Eigenvalue Problem.* Prentice-Hall, Englewood Cliffs, N.J.

RICE, J. (1986). Choice of smoothing parameter in deconvolution problems. To appear in *Contemp. Math.*

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

TITTERINGTON, D. M. (1985). General structure of regularization procedures in image reconstruction. *Astronom. and Astrophys.* **144** 381–387.

WAHBA, G. (1980). Ill-posed problems: Numerical and statistical methods for mildly, moderately and severely ill-posed problems with noisy data. Technical Report 595, Dept. Statist., Univ. Wisconsin-Madison. (To appear in *Proceedings of the International Conference on Ill-Posed Problems*, M. Z. Nashed, ed.).