Statistical Science
1986, Vol. 1, No. 4, 502–527

A Statistical Perspective on Ill-posed Inverse Problems

Finbarr O'Sullivan

Abstract. Ill-posed inverse problems arise in many branches of science and engineering. In the typical situation one is interested in recovering a whole function given a finite number of noisy measurements on functionals. Performance characteristics of an inversion algorithm are studied via the mean square error which is decomposed into bias and variability. Variability calculations are often straightforward, but useful bias measures are more difficult to obtain. An appropriate definition of what geophysicists call the Backus-Gilbert averaging kernel leads to a natural way of measuring bias characteristics. Moreover, the ideas give rise to some important experimental design criteria. It can be shown that the optimal inversion algorithms are methods of regularization procedures, but to completely specify these algorithms the signal to noise ratio must be supplied. Statistical approaches to the empirical determination of the signal to noise ratio are discussed; cross-validation and unbiased risk methods are reviewed; and some extensions, which seem particularly appropriate in the inverse problem context, are indicated. Linear and nonlinear examples from medicine, meteorology, and geophysics are used for illustration.

Key words and phrases: Averaging kernel, B-splines, cross-validation, experimental design, mean square error, reservoir engineering, stereology, satellite meteorology.

1. INTRODUCTION

Inverse problems pertain to situations where one is interested in making inferences about a phenomenon from partial or incomplete information. Accordingly, statistical estimation and model building are both inverse problems. In modern science there is an increasingly important class of inverse problems which are not amenable to classical statistical estimation procedures and such problems are termed ill-posed. The notion of ill-posedness is usually attributed to Hadamard (1923); a modern treatment of the concept appears in Tikhonov and Arsenin (1977). In an ill-posed inverse problem, a classical least squares, minimum distance, or maximum likelihood solution may not be uniquely defined. Moreover, the sensitivity of such solutions to slight perturbations in the data is often unacceptably large.

A typical example of an ill-posed inverse problem, arising in stereology, is described by Nychka et al. (1984). Here one is interested in the estimation of three-dimensional tumor size distribution in liver tissue from measurements on cross-sectional slices. A schematic for the experiment is given in Figure 1.1.

By modeling tumors as spheres randomly distributed in the tissue, an approximate integral relationship between the three-dimensional distribution of tumor radii and the two-dimensional distribution of radii observed in cross-sectional slices may be derived. Letting z_i be the observed proportion of two-dimensional slices with radii in the interval $[x_i, x_{i+1}]$ one has

$$z_i = \int K(x_i, r) f_3(r)\, dr + \varepsilon_i, \quad i = 1, 2, \cdots, m,$$

where ε_i are measurement/modeling errors, f_3 is the density of three-dimensional tumor radii, and the kernels $K(x_i, r)$ are given by

$$K(x_i, r) = \begin{cases} 0 & \varepsilon \le r < x_i, \\ \sqrt{r^2 - x_i^2} & x_i \le r < x_{i+1}, \\ \sqrt{r^2 - x_i^2} - \sqrt{r^2 - x_{i+1}^2} & x_{i+1} \le r < R, \end{cases}$$
$$i = 1, 2, \cdots, m,$$

where $\varepsilon \le x_1 < x_2 < \cdots < x_{m+1} \le R$. Physically, ε is the smallest detectable tumor radius and R is the

Finbarr O'Sullivan is Assistant Professor, Department of Statistics, University of California, Berkeley, California 94720.

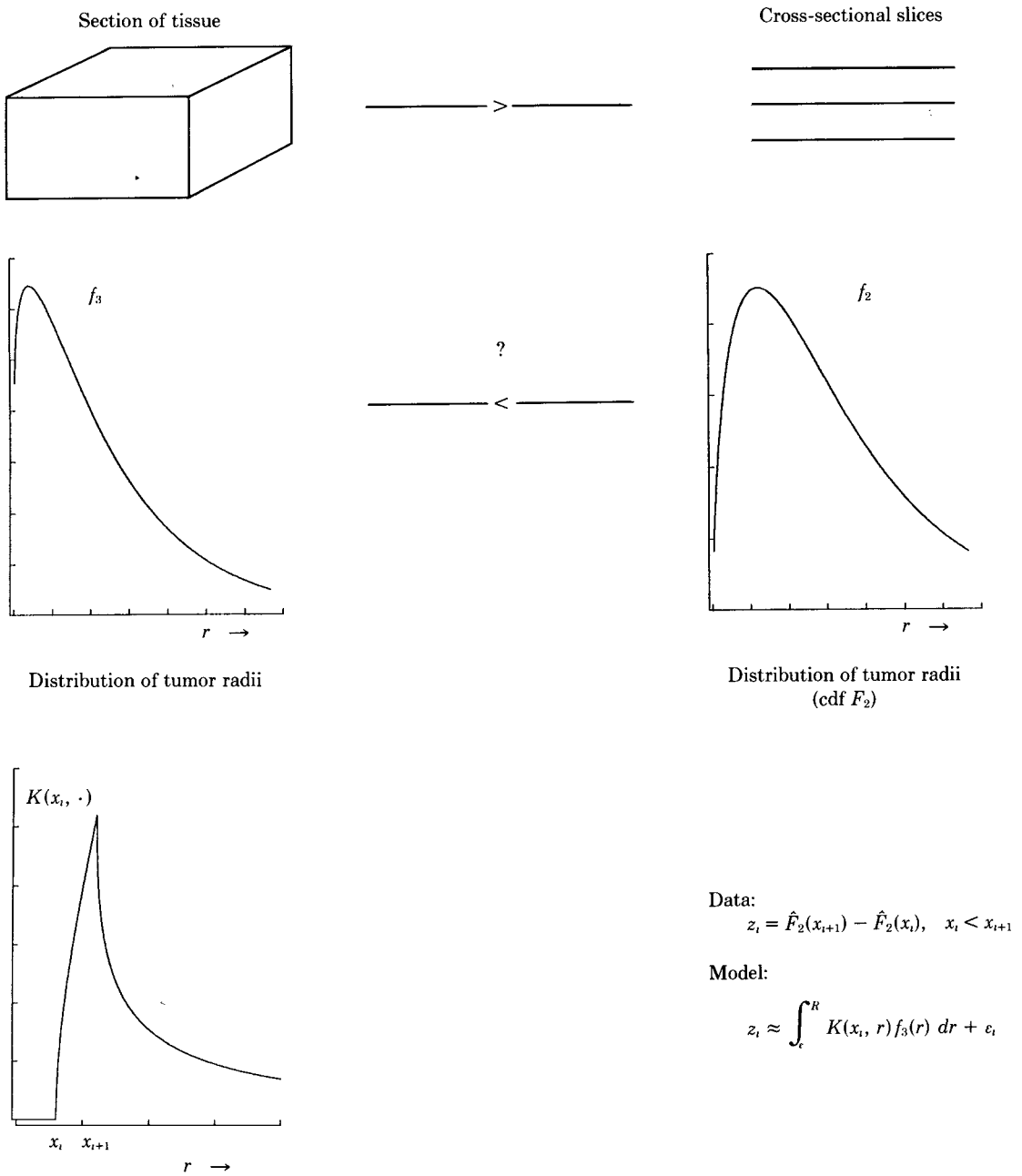

FIG. 1.1. *Reconstruction of the tumor size distribution from data on cross-sectional slices.*

largest possible tumor radius in the given section of tissue. Since $\varepsilon \leq x_1 < x_2 < \cdots < x_{m+1} \leq R$, the estimation of f_3 is restricted to the interval $[\varepsilon, R]$. It is easy to appreciate the ill-posedness of this inverse problem. The kernels $K(x_i, \cdot)$ are smooth; as a result relatively large perturbations of f_3 can give rise to very slight perturbations in the data and conversely. It follows from this that least squares, minimum χ^2, or maximum likelihood solutions will be very sensitive to slight changes in the data.

Ill-posed inverse problems have become a recurrent theme in modern science, see for example crystallography (Grunbaum, 1975), geophysics (Aki and Richards, 1980; Bolt, 1980; Jeffreys, 1976), medical electrocardiograms (Franzone, Taccardi, and Viganotti, 1977), meteorology (Smith, 1983; Smith et al., 1979), microfluoroimagery (Mendelsohn and Rice, 1982), radio astronomy (Jaynes, 1983), reservoir engineering (Kravaris and Seinfeld, 1985; Neuman and Yakowitz, 1979), and tomography (Budinger, 1980; Vardi, Shepp, and Kaufman, 1985). Corresponding to this broad spectrum of fields of application, there is a wide literature on different kinds of inversion algorithms, that is, techniques for solving the inverse problem. The basic principle common to all such methods is as follows: seek a solution that is consistent

both with the observed data and *prior* notions about the physical behavior of the phenomenon under study. Different practical problems have led to unique strategies for implementation of this principle, such as the method of regularization (Tikhonov and Arsenin, 1977), maximum entropy (Jaynes, 1983), and quasi reversibility (Lattès and Lions, 1969). Understanding the performance characteristics of a given inversion method is an important issue. First, such information has obvious intrinsic value and second, it can critically influence the choice of experimental design (see Section 2).

The primary goal of this paper is to identify some tools for assessing the finite sample performance characteristics of an inversion algorithm. These tools, most of which can be found scattered throughout the diverse inverse problem literature, are considered in the context of the following generalized nonlinear regression model. Measurements, z_i, are of the form:

$$z_i = \eta(x_i, \theta) + \varepsilon_i, \quad i = 1, 2, \cdots, m,$$

where θ is in Θ (the nominal parameter space), $\eta(x_i, \cdot)$ are linear or nonlinear functionals of θ, and the ε_i are measurement errors assumed to have mean zero. In all that I discuss, there will be an underlying assumption that the unknown true function, θ, is well approximated by a *smooth* function. Although this assumption does not allow one to talk about highly discontinuous functions, such as those that arise in typical pattern recognition problems, the model is still quite general, and includes for example linear and nonlinear integral equations of the first kind. Examples used later on include the temperature retrieval problem in satellite meteorology and the system identification problem of reservoir engineering.

1.1 Inversion Algorithms and the Method of Regularization

An inversion algorithm, S, is a mapping that takes data into parameter estimates,

$$\hat{\theta} = S\mathbf{z}.$$

When S is linear then $\hat{\theta}$ can be written as a linear combination of *impulse response functions*, s_i; i.e.,

$$\hat{\theta}(t) = \sum_{i=1}^{m} s_i(t) z_i,$$

where $s_i = S\mathbf{e}_i$ and $\mathbf{e}_i$ is the ith unit vector in R^m. In statistical terms, an inversion algorithm would be simply called an estimator, but here the terms *estimator* and *inversion algorithm* are used interchangeably.

One of the most useful techniques for generating inversion algorithms or estimators is the method of regularization (MOR) (see Titterington, 1985). The MOR procedure is due to Tikhonov (1963) and Tikhonov and Arsenin (1977). There are various possible implementations of the method, but they all amount to choosing $\hat{\theta}$ to be the minimizer of a weighted combination of two functionals. The first functional measures lack of fit to the observed data and the second measures physical plausibility of the estimate. For example, one might choose $\hat{\theta}$ to be the minimizer of a criterion of the form

$$(1.1) \quad \frac{1}{m} \sum_{i=1}^{m} [z_i - \eta(x_i, \theta)]^2 + \lambda J(\theta), \quad \lambda > 0.$$

The functional J is chosen so that highly *irregular* or *physically implausible* θ's have large values. Statisticians will recognize this method as a version of penalized likelihood estimation described by Good and Gaskins (1971); the sum of squares is the likelihood part and the functional J the penalty term. The method is equivalent to the method of sieves introduced by Grenander (1981). Also, if J is chosen by some information principle such as entropy, $J(\theta) = -\int \theta(t) \log \theta(t)\, dt$, then the method of regularization yields a procedure equivalent to the method of maximum entropy pioneered by Jaynes (see Jaynes, 1983, and Chapter 4 of McLaughlin, 1983).

When the functionals, $\eta(x_i, \cdot)$, are linear in θ, and J is quadratic with $J(\theta) \geq 0$ ($= 0$ for $\theta = 0$), then the solution to [1.1] is linear in the observed data. Moreover, in this case the MOR has an interesting Bayesian interpretation. To see this, first suppose that Θ is finite-dimensional, i.e., $\Theta = \text{span}_{1 \leq k \leq K}\{\phi_k\}$ with ϕ_k linearly independent. The elements of Θ can be written as $\sum_k \beta_k \phi_k$ for $\beta = (\beta_1, \beta_2, \cdots, \beta_K)'$ in R^K, so that Θ can be identified with R^K. Since J is quadratic and $J(0) = 0$, $J(\theta)$ can be expressed as a quadratic form in $\boldsymbol{\beta}$, $J(\theta) = \boldsymbol{\beta}' \Omega \boldsymbol{\beta}$, for some positive semidefinite matrix Ω, it follows that the MOR estimator is $\sum_k \hat{\boldsymbol{\beta}}_k \theta_k$ where $\hat{\boldsymbol{\beta}}$ minimizes

$$\frac{1}{m} \sum_{i=1}^{m} [z_i - X_i' \boldsymbol{\beta}]^2 + \lambda \boldsymbol{\beta}' \Omega \boldsymbol{\beta}.$$

Here X is a design matrix, $X_{ik} = \eta(x_i, \phi_k)$. Thus $\hat{\boldsymbol{\beta}}$ is given by

$$\hat{\boldsymbol{\beta}} = S\mathbf{z} \quad \text{where} \quad S = [X'X + m\lambda\Omega]^{-1} X'.$$

Obviously, the MOR estimator is linear in the observed data. A Bayesian interpretation for $\hat{\boldsymbol{\beta}}$ is obtained by specifying a Gaussian prior with mean zero and covariance matrix proportional to Ω^{-1}. Then, if the ε_i values are independent and identically distributed Gaussian random variables with mean zero, the MOR estimator, $\hat{\boldsymbol{\beta}}$ is the posterior mean of $\boldsymbol{\beta}$ given the data.

The foregoing statements carry through to more general settings. If Θ is a Hilbert space with inner

product $\langle \cdot, \cdot \rangle$, the $\eta(x_i, \cdot)$ are bounded linear functionals and J is quadratic, $J(\theta) = \langle \theta, W\theta \rangle$, where W is positive semidefinite, then it can be shown (see Cox, 1983, for example) that the MOR estimators have the form

$$\hat{\theta}_\lambda = S\mathbf{z} \quad \text{where} \quad S = [X'X + m\lambda W]^{-1}X'$$

where the *design matrix*, X is an operator derived from the functionals $\eta(x_i, \cdot)$. Thus S is a linear operator from the data space into Θ. General Bayesian interpretations for the method are also available, these are discussed by Kimeldorf and Wahba (1971). Further results on the optimality of the MOR are described in Section 2.

1.2 Outline

Performance characteristics of an inversion algorithm are studied via the mean square error which can be split into bias and variability components. Bias measures the systematic error while variability measures the random error. In Section 2, I consider linear inversion algorithms and describe some ways of measuring bias and variability. Variability is calculated in a direct manner. For bias it is convenient to introduce a generalized version of what geophysicists call the Backus-Gilbert averaging kernel (see Backus and Gilbert, 1968 and 1970). The generalized notion of the averaging kernel allows one to compute the maximum or average expected bias and also leads to some natural design criteria. These are described and illustrated in Section 2.4. Optimal inversion algorithms can be found and these turn out to be MOR procedures. The use of B-splines for obtaining numerically convenient and reliable approximations to the averaging kernel and bias is described in Section 3. Although the theory of linear inverse problems is fairly well established, the field of nonlinear inverse problems is in its infancy. There are many exciting and challenging problems that need to be tackled in this area. The performance of MOR estimators when applied to two interesting nonlinear inverse problems is discussed in Section 4. One of these problems arises in satellite meteorology and is concerned with the estimation of atmospheric temperature from upwelling radiance measurements. The second problem is of major interest in reservoir engineering. It concerns the estimation of reservoir characteristics, the ease of flow of fluid in a reservoir, from pressure-history data measured at distributed well sites. As described in Section 2, optimal inversion algorithms although MOR procedures are not fully specified without supplying the signal to noise ratio. For a given problem this will not be known and so has to be empirically determined. The final section of the paper deals with this issue; the methods of cross-validation and unbiased risk are described and some relevant extensions to ill-posed inverse problems are developed.

2. FINITE SAMPLE PERFORMANCE OF AN INVERSION ALGORITHM

The quality of an inversion algorithm at some point, t, is measured by comparing the estimate, $\hat{\theta}(t)$, to the true value, $\theta(t)$. This difference can be decomposed into systematic and random components as

$$\theta(t) - \hat{\theta}(t) = [\theta(t) - E\hat{\theta}(t)] + [E\hat{\theta}(t) - \hat{\theta}(t)].$$

The expectation is with respect to the error distribution. The average performance of the inversion at t is measured by the mean square error (MSE).

$$\begin{aligned} \text{MSE}(t) &= E[\theta(t) - \hat{\theta}(t)]^2 \\ &= [\theta(t) - E\hat{\theta}(t)]^2 + E[\hat{\theta}(t) - E\hat{\theta}(t)]^2 \\ &= \text{bias}^2(t, \theta) + \text{var}(t, \theta). \end{aligned}$$

Mean square error depends both on θ and the assumed error distribution. It is the sum of the squared bias, $\text{bias}^2(t, \theta)$, and the variance, $\text{var}(t, \theta)$. If the inversion algorithm is designed solely to minimize bias then the variance dominates the mean square error and vice versa. Thus a good inversion algorithm must balance bias and variability. Unbiasedness is not a desirable property in this context.

Mean square error performance of an inversion algorithm can, in principle, be found by Monte Carlo simulation. Modern computing resources are making this a very viable and practical approach. For linear problems one can avoid direct Monte Carlo simulation and in the process obtain some useful insights which can be applied to more complex situations.

2.1 Linear Problems

By a linear problem I shall mean that both the functionals, $\eta(x_i, \cdot)$, are linear in θ, and the inversion algorithm, S, is linear in the data.

2.1.1 Variability

Variability computations for linear inversion algorithms are very straightforward. By linearity, variability does not depend on θ and

$$\text{var}(t, \theta) \equiv \text{var}(t) = \text{var}\left[\sum_{i=1}^{m} s_i(t)\varepsilon_i\right]$$

where s_i is the impulse response function. Thus if the errors, ε_i, have covariance $\sum_\varepsilon$, then

$$\text{var}(t) = \mathbf{s}(t)' \, \Sigma_\varepsilon \, \mathbf{s}(t)$$

where $\mathbf{s}(t) = (s_1(t), s_2(t), \cdots, s_m(t))'$. In particular, if the errors are independent with constant variance σ^2,

then the variability is simply

$$\text{var}(t) = \sigma^2 \mathbf{s}(t)'\mathbf{s}(t).$$

2.1.2 Bias

Bias properties are best understood by introducing the notion of the averaging kernel. In certain circumstances the averaging kernel yields a representation of the form

$$E\hat{\theta}(t) = \int A(t, s)\theta(s)\, ds, \tag{2.1}$$

for the systematic part of the estimate. The function $A(t, \cdot)$ is known as the averaging kernel at t, and it determines the nature of the bias incurred at t. The averaging kernel is related to what engineers and astronomers call the *point-spread* function. The point-spread function at t is defined to be the solution obtained by the inversion algorithm when the true function is a Dirac δ-function at t and there is no measurement error. Thus from [2.1] if the averaging kernel at t is $A(t, \cdot)$, then the point-spread at t is the function $A(\cdot, t)$.

The representation in [2.1] is an L_2 representation for the averaging kernel. In the geophysics literature this representation is known as the Backus-Gilbert averaging kernel, after two geophysicists Backus and Gilbert (1968, 1970). Related ideas go back to Peano (1914), who used a similar representation to study the bias in numerical quadrature formulae (see also Sard, 1949). Alternative representations for the averaging kernel are also possible and these alternative representations are more useful when it comes to computing bias. The averaging kernel and its generalizations are described next.

2.2 Averaging Kernel

2.2.1 Backus-Gilbert Formulation

Backus and Gilbert worked in an integral equation context,

$$\eta(x_i, \theta) = \int K_i(s)\theta(s)\, ds, \quad i = 1, 2, \cdots m,$$

and the kernels $K_i(\cdot)$ are known smooth functions. For a linear inversion algorithm the $E\hat{\theta}(t)$ can be written as

$$E\hat{\theta}(t) = \sum_{i=1}^{m} s_i(t) \int K_i(s)\theta(s)\, ds.$$

Taking the summation inside the integral sign, this becomes:

$$E\hat{\theta}(t) = \int A(t, s)\theta(s)\, ds$$

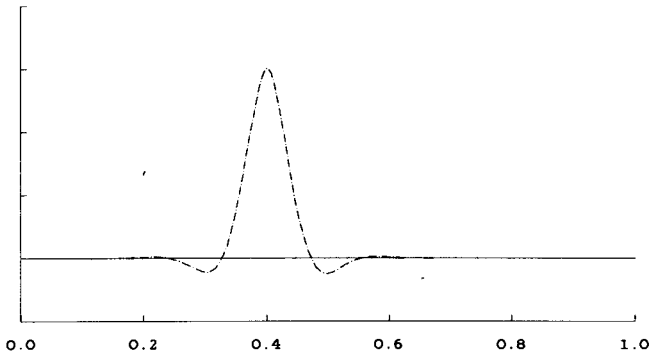

FIG. 2.1. *Sample averaging kernel for a MOR procedure applied to the tumor problem* ($r = 0.4$).

where

$$A(t, s) = \sum_{i=1}^{m} s_i(t)K_i(s).$$

The function $A(t, \cdot)$ is the Backus-Gilbert averaging kernel for the inversion algorithm S at t. For illustration, an averaging kernel corresponding to a method of regularization procedure applied to the tumor size distribution problem, described in Section 1, is given in Figure 2.1. One can see that the kernel is well centered about the point of interest, $r = 0.4$. Moreover the kernel seems fairly symmetric so that if the three-dimensional tumor radius density were locally linear in the neighborhood of this point, then the inversion algorithm would be locally unbiased. Properties of the averaging kernel can be varied by changing the regularization parameter, λ—large values of λ cause the averaging kernel to be more spread out. Techniques for empirically selecting this parameter are discussed in Section 5.

The center, spread, and skewness of the averaging kernel give a rough appreciation for its behavior. Assuming they exist, these are defined to be the first, second, and third moment of the absolute value of the averaging kernel when suitably normalized. Let $\bar{A}_\lambda(t, \cdot) = A_\lambda(t, \cdot)/\int |A_\lambda(t, s)|\, ds$. Then the characteristics of the averaging kernel are as follows:

Center

$$c(t) = \int |\bar{A}_\lambda(t, s)|\, s\, ds;$$

Spread

$$sp(t) = \sqrt{\int |\bar{A}_\lambda(t, s)|\, (s - c(t))^2\, ds};$$

Skewness

$$sk(t) = \int |\bar{A}_\lambda(t, s)| \left[\frac{(s - c(t))}{sp(t)}\right]^3 ds.$$

The skewness is dimensionless while the center and spread are in t units. Skewness is important since a symmetric averaging kernel ($sk(t) = 0$) will exactly pass a linear trend.

Intuitively, the bias at a point, t, is determined by how close the averaging kernel is to a Dirac δ-function at t. Backus and Gilbert tried to develop some direct measures of the nearness of the averaging kernel to a Dirac δ-function—"δ-ness of the averaging kernel." By choosing the inversion algorithm, so that the averaging kernel is as δ-like as possible, subject to some upper bound on the size of the standard error, one obtains so called Backus-Gilbert inversion algorithms. The idea seems perfectly reasonable; however, there is some degree of arbitrariness in the δ-ness criteria defined by Backus and Gilbert. Moreover, in general it is not true that the function maximizing a δ-ness criterion will necessarily be the Dirac δ-function. Problems with the δ-ness criteria really arise because the Backus-Gilbert calculus takes place in an L_2 setting where evaluation is not a continuous linear functional. By working in a space where evaluation is continuous, one can derive a more refined formulation of the averaging kernel and use a straightforward calculus to assess δ-ness. The refined definition of the averaging kernel also allows one to deal with more general linear functionals $\eta(x_i, \theta)$.

2.2.2 Refined Formulation of the Averaging Kernel

Preliminaries: Linear Functionals and Representers. The notion of a representer of a continuous linear functional will be needed. To motivate this concept, consider first the case where $\Theta = \text{span}_{1\le k\le K}\{\phi_k\}$ and the ϕ_k are linearly independent. Here, elements of Θ are identified by a K-vector of coefficients, $\boldsymbol{\beta}$ in R^K. Moreover, the usual inner product on R^K determines an inner product on Θ by

$$\langle \theta_1, \theta_2 \rangle = \boldsymbol{\beta}_1' \boldsymbol{\beta}_2$$

where $\theta_1 = \sum_{k=1}^K \boldsymbol{\beta}_{1k}\phi_k$ and $\theta_2 = \sum_{k=1}^K \boldsymbol{\beta}_{2k}\phi_k$. If $\eta(x_i, \cdot)$ is linear then for any θ in Θ

$$\eta(x_i, \theta) = X_i' \boldsymbol{\beta} \tag{2.2}$$

where $\theta = \sum_{k=1}^K \boldsymbol{\beta}_k\phi_k$ and $X_i = (\eta(x_i, \phi_1), \eta(x_i, \phi_2), \cdots, \eta(x_i, \phi_K))'$. The vector X_i determines an element, $\xi_{x_i} = \sum_{k=1}^K X_{ik}\phi_k$, in Θ with the property that for all θ in Θ

$$\eta(x_i, \theta) = \langle \xi_{x_i}, \theta \rangle.$$

In functional analysis terms, ξ_{x_i} is the *representer* of the linear functional $\eta(x_i, \cdot)$. An important linear functional is evaluation at a point. The representer in Θ of evaluation at t is given by

$$e_t = \sum_{k=1}^K \phi_k(t)\phi_k.$$

One can easily verify that

$$\theta(t) = \langle e_t, \theta \rangle,$$

for all θ in Θ.

The notion of a representer extends to more elaborate function spaces. The level of functional analysis needed to understand this is elementary and the interested reader might consult Rudin (1976). It is important to realize that, in general, the form of the representer depends on which inner product is used. Let Θ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. A linear functional l is continuous if there is a constant M such that

$$| l(\theta) | \le M \| \theta \| \quad \text{for all} \quad \theta \text{ in } \Theta.$$

Corresponding to any continuous linear functional, l, there exists a representer ξ in Θ such that

$$l(\theta) = \langle \xi, \theta \rangle \quad \text{for all} \quad \theta \text{ in } \Theta.$$

This is known as the Riesz representation theorem (Rudin, 1976). A Hilbert space of real valued functions in which evaluation is continuous is known as a reproducing kernel Hilbert space (RKHS). Reproducing kernel Hilbert spaces play an important role in applied mathematics and their role in the study of ill-posed inverse problems has been emphasised by Golomb and Weinberger (1959); see also Wahba (1984). Evaluation is not continuous in L_2, but it is continuous in the space of functions the first derivatives of which are square integrable.

Representer of the Averaging Kernel. Given the notion of a representer, the generalization of the averaging kernel is very simple. Let Θ be a real Hilbert space with inner product $\langle \cdot, \cdot \rangle$ and suppose that the functionals $\eta(x_i, \cdot)$ are continuous with representers ξ_{x_i}. Since

$$E\hat{\theta}(t) = \sum_{i=1}^m s_i(t)\eta(x_i, \theta),$$

the representer of the *averaging kernel*, $A(t)$, is given by

$$A(t) = \sum_{i=1}^m s_i(t)\xi_{x_i}.$$

Thus the averaging kernel is a linear combination of the representers of the functional $\eta(x_i, \cdot)$. The particular linear combination is determined by the impulse response functions of the inversion algorithm.

2.3 Bias Measures and Some Design Criteria

The more general formulation of the averaging kernel leads to natural ways of measuring bias and this in turn motivates some useful design criteria. Let Θ be such that evaluation at t is continuous and let e_t denote the corresponding representer. From the averaging kernel, the bias at t may be written as

$$\theta(t) - \hat{\theta}(t) = \langle e_t - A(t), \theta \rangle.$$

Average Bias. The representation, $\langle e_t - A(t), \theta \rangle$, can be used to compute the expected squared bias with respect to a prior distribution on possible θ values. Thus, if $\Theta = \text{span}_{1 \le k \le K}\{\phi_k\}$ and a prior mean and covariance for θ is specified by means of the coefficients of the ϕ_k's, $\boldsymbol{\beta}$,

$$E_\beta[\boldsymbol{\beta}] = \boldsymbol{\beta}_0, \quad \text{Var}_\beta[\boldsymbol{\beta}] = \sum_\beta,$$

then the average expected bias is

$$\begin{aligned}\overline{b^2}(t) &= E_\beta[\theta(t) - E\hat{\theta}(t)]^2 \\ &= E_\beta\langle e_t - A(t), \theta \rangle^2 \\ &= \mathbf{c}_t' \sum_\beta \mathbf{c}_t + \langle e_t - A(t), \theta_0 \rangle^2,\end{aligned}$$

where

$$\mathbf{c}_t = (\langle e_t - A(t), \phi_1 \rangle, \langle e_t - A(t), \phi_2 \rangle, \cdots, \langle e_t - A(t), \phi_K \rangle)'$$

and

$$\theta_0 = \sum_{k=1}^{K} \beta_{0k}\phi_k.$$

Maximum Bias. A less sophisticated measure of bias is the maximum bias over all functions in Θ the norm of which is less than some specified value, μ. From the averaging kernel representation for bias and the Cauchy-Schwarz inequality, this is given by

$$\begin{aligned}\sup_{\|\theta\|^2 \le \mu^2} [\theta(t) - E\hat{\theta}(t)]^2 &= \sup_{\|\theta\|^2 \le \mu^2} \langle e_t - A(t), \theta \rangle^2 \\ &= \| e_t - A(t) \|^2 \mu^2.\end{aligned}$$

Thus letting $b_M(t) = \| e_t - A(t) \|$, $b_M^2(t)$ is proportional to the maximum squared bias over any ball in Θ. Sard (1949) would probably call $b_M^2(t)$ the modulus of the bias at t.

Figure 2.2 plots the maximum bias, $b_M(t)$, and standard error, square root of variability, for a MOR inversion applied to the tumor size distribution problem. Both the bias and standard error have poor behavior for small radii, suggesting that there is difficulty in getting reliable estimates of the size distribution for small radii. This is a consequence of the simple fact that small tumors are hard to detect, see Nychka et al. (1984) for more discussion. The ripples in the plot are due to the finite sample size, $m = 50$, in this illustration. The vertical scales on the bias and variability plots are left unspecified. Only if particular values for the μ and σ are assigned can an appropriate scaling be set up.

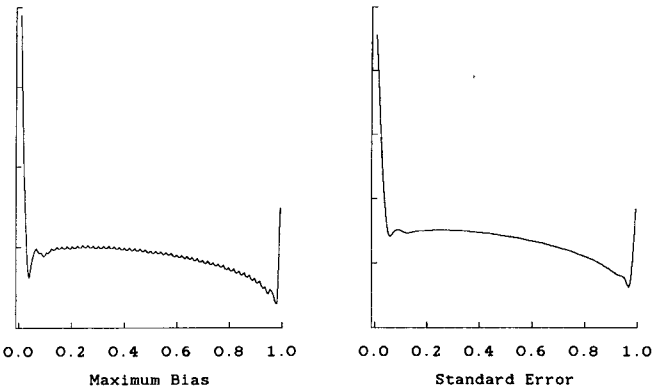

FIG. 2.2. *Maximum bias and standard error for a MOR procedure applied to the tumor problem. The ripples are due to the finite sampling* ($m = 50$).

Design Criteria. Combining bias with variability gives an overall assessment of the performance of the inversion algorithm which can be used for design purposes. The average bias measure gives rise to an average mean square error design criterion (AMSE):

$$\text{AMSE}(t) = \overline{b^2}(t) + \text{var}(t).$$

Also, from the maximum bias measure, one obtains the maximum mean square error design criterion (MMSE):

$$\text{MMSE}(t) = \mu^2 b_M^2(t) + \text{var}(t).$$

2.4 Optimal Inversion Algorithms and Experimental Design

Since both the average and maximum mean square errors depend on the inversion algorithm, algorithms can be selected which perform best with respect to either of these criteria. Interestingly, the solution one obtains in each case is a MOR procedure. This analysis can be carried out in a very general setting. However, the structure of these results is most transparent when Θ is finite-dimensional and the measurement errors are mean zero uncorrelated with constant variance σ^2.

Minimizing the Average Mean Square Error. Let $\mathbf{s}(t) = (s_1(t), s_2(t), \cdots, s_K(t))$. For simplicity suppose that the prior for $\boldsymbol{\beta}$ has zero mean and covariance $\tau^2 \sum_\beta$. The average mean square error can be written as

$$\begin{aligned}\text{AMSE}(t) = \tau^2 \left[X_t - \sum_{i=1}^{m} s_i(t) X_i \right]' \Sigma_\beta \left[X_t - \sum_{i=1}^{m} s_i(t) X_i \right] \\ + \sigma^2 \mathbf{s}(t)'\mathbf{s}(t),\end{aligned}$$

where $X_t = (\phi_1(t), \phi_2(t), \cdots, \phi_K(t))'$ and the ith row

of X is given in equation [2.2]. Minimizing with respect to $\mathbf{s}(t)$ gives:

$$\left[X\Sigma_\beta X' + \frac{\sigma^2}{\tau^2} I\right]\hat{s}(t) = X\Sigma_\beta X_t'.$$

This holds for any t. If Σ_β is invertible, then, after some algebra, the optimal inversion algorithm can be expressed as

$$\hat{S} = \left[X'X + \frac{\sigma^2}{\tau^2}\Sigma_\beta^{-1}\right]^{-1} X'.$$

Comparing this with the forms given in Section 1.1, it can be seen that the optimal algorithm corresponds to a MOR estimator with $\lambda = (\sigma^2/m\tau^2)$ and $J(\theta) = \boldsymbol{\beta}\Sigma_\beta^{-1}\boldsymbol{\beta}$. λ^{-1} is interpreted as the signal to noise ratio.

Minimizing the Maximum Mean Square Error. The maximum mean square error in a ball of radius μ is given by

$$\text{MMSE}(t) = \mu^2\left[X_t - \sum_{i=1}^m s_i(t)X_i\right]'\left[X_t - \sum_{i=1}^m s_i(t)X_i\right] + \sigma^2\mathbf{s}(t)'\mathbf{s}(t).$$

Again minimizing with respect to $\mathbf{s}(t)$, the optimal vector satisfies

$$\left[XX' + \frac{\sigma^2}{\mu^2} I\right]\hat{\mathbf{s}}(t) = XX_t',$$

so the optimal inversion algorithm is given by

$$\hat{S} = \left[X'X + \frac{\sigma^2}{\mu^2} I\right]^{-1} X'.$$

This is a MOR inversion algorithm with $\lambda = (\sigma^2/m\mu^2)$ and $J(\theta) = \|\theta\|^2$. Here λ^{-1} is again interpreted as a signal to noise ratio.

Versions of the above results have appeared at several times in the literature (see Foster, 1961; Strand and Westwater, 1968; and Weinreb and Crosby, 1972). The optimal inversion algorithm from the point of view of average mean square error is called the minimum rms solution, while the optimal solution for maximum mean square error is the minimum information solution. Generalizations of these results have appeared in the statistics literature. Kimeldorf and Wahba (1971) show that the minimum rms solution is sometimes interpretable as an optimal Bayesian procedure. Minimum information solutions have also been termed minimax. Results on the minimaxity of the MOR are given by Li (1982) and also Speckman (1979). In practice the signal to noise ratio is not known so that the parameter λ needs to be set empirically. Some statistical methods available for doing this are given in Section 5.

Experimental Design. The MMSE and the AMSE are also functions of the design points, x_i, $i = 1, 2, \cdots, m$, and an optimal experimental design can be defined as the design making the MMSE or the AMSE minimum. Since one may not be interested in performance at just a single point, t, an integrated AMSE or MMSE over some region of interest is often more appropriate. Weinreb and Crosby (1972) have carried out a program of this kind in connection with the selection of spectral wavelengths for satellite radiometers. Their criterion is reduced to a simple trace criterion (see equation [10] of their paper). More recently, Wahba (1983a) has also discussed the design issue. She presents two design criterion: one of which is akin to Weinreb and Crosby's criterion, see their equation [13], and a second which is termed the "degrees of freedom for signal" criterion. The use of integrated mean square error as a design criterion is not new to statisticians. Box and Draper (1959) proposed this as a design criterion for *model-robust* response surface designs. Further discussion of this literature can be found in Section 5 of the recent paper by Steinberg and Hunter (1984).

3. NUMERICAL APPROXIMATION OF THE AVERAGING KERNEL WITH B-SPLINES

Exact computation of averaging kernels requires the manipulation of the representers of the functionals $\eta(x_i, \cdot)$ for $i = 1, 2, \cdots, m$. In reproducing kernel Hilbert spaces there are theoretical formulas available for the evaluation of representers, see Nychka et al. (1984) for example. However, direct evaluation of averaging kernels by means of such formulas is extremely inefficient, computationally, so other approaches are needed. Approximating the elements of Θ by simpler forms, it is possible to obtain highly efficient methods for evaluating the averaging kernels at a negligible loss in accuracy. To illustrate this I consider one very popular MOR procedure for estimating a one-dimensional real valued function, θ defined on an interval $[a, b]$: $\hat{\theta}$ is the minimizer over the Sobolev space $W_2^2[a, b] = \Theta$ of

$$\frac{1}{m}\sum_{i=1}^m [z_i - \eta(x_i, \vartheta)]^2 + \lambda\int_a^b [\ddot{\theta}(t)]^2\,dt, \quad \lambda > 0.$$

$W_2^2[a, b]$ is an infinite-dimensional Hilbert space with inner product

$$\langle\theta, \phi\rangle = \int \theta(t)\phi(t)\,dt + \int \ddot{\theta}(t)\ddot{\phi}(t)\,dt \quad \theta, \phi \text{ in } W_2^2[a, b].$$

Sobolev spaces are discussed at length in the book by Adams (1975). It is well known that the elements of Θ can be approximated to an arbitrarily high degree of accuracy by cubic B-splines. We take good

advantage of this in the computation of the averaging kernels. Before describing this let me pause briefly to describe cubic B-splines. The standard reference on B-splines is the book by DeBoor (1978). Throughout this section continue to assume that the functionals $\eta(x_i, \cdot)$ are linear.

3.1 B-splines

Consider a set of distinct knot points $[t_1 < t_2 < \cdots < t_{K+4}]$ with $t_3 \le a$ and $b \le t_{K+2}$. (Multiple knots are also allowed; see DeBoor (1978) for more details.) A cubic spline on $[a, b]$ is a cubic polynomial between successive knot points joined up at the knots so as to have continuous second derivative over the interval $[a, b]$. By increasing the knot density in $[a, b]$, functions in $W_2^2[a, b]$ can be very closely approximated by cubic splines. Cubic splines can be expressed as a linear combination of basis elements $\{B_k\}_{k=1}^K$. The elements, B_k, of the basis are called B-splines and the entire basis is called the B-spline basis. Each element of the basis is non-negative and has very local support; in fact, B_k is zero outside of the interval $[t_k, t_{k+4}]$. Over $[t_k, t_{k+4}]$, B_k is proportional to the probability density for the sum of four independent uniform random variables; U_i, $i = 1, 2, 3, 4$, where U_i is defined on $[t_{k+i-1}, t_{k+i}]$. The local support property of B-splines can be used to great advantage in computations. DeBoor (1978) has developed a set of Fortran programs for manipulating B-splines and these are now available in most modern mathematical software libraries.

3.2 Computation of the Averaging Kernel with B-splines

Let $\Theta_K = \{B_k\}_{k=1}^K$. The MOR estimate $\hat{\theta}_\lambda$ is approximated as $\hat{\theta}_\lambda \approx \sum_{k=1}^K \hat{\beta}_k B_k$ where

$$\hat{\boldsymbol{\beta}} = [X'X + m\lambda\Omega_2]^{-1}X'\mathbf{z} \tag{3.1}$$

and $X_{ik} = \eta(x_i, B_k)$ and $\Omega_2(j, k) = \int_a^b \ddot{B}_j(t)\ddot{B}_k(t)\,dt$.

Three inner products on Θ_K will now be defined. Let $\theta = \sum_{k=1}^K \theta_k B_k$ and $\phi = \sum_{k=1}^K \phi_k B_k$ in Θ.

1. Euclidean Inner Product:

$$\langle \theta, \phi \rangle_E = \sum_{k=1}^K \theta_k\phi_k = \boldsymbol{\theta}'\boldsymbol{\phi}$$

where

$$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_K)'$$

and

$$\boldsymbol{\phi} = (\phi_1, \phi_2, \cdots, \phi_K)',$$

2. L_2 Inner Product:

$$\langle \theta, \phi \rangle_2 = \int \theta(t)\phi(t)\,dt = \boldsymbol{\theta}'\Omega_0\boldsymbol{\phi}$$

where $\Omega_0(j, k) = \int B_j(t)B_k(t)\,dt$,

3. Sobolev Inner Product:

$$\langle \theta, \phi \rangle_S = \int \theta(t)\phi(t)\,dt + \int \ddot{\theta}(t)\ddot{\phi}(t)\,dt$$

$$= \boldsymbol{\theta}'[\Omega_0 + \Omega_2]\boldsymbol{\phi}.$$

Corresponding to each inner product there is an approximate representation for the averaging kernel at t. If these be denoted $A_E(t)$, $A_2(t)$, and $A_S(t)$, then

$$E\hat{\theta}_\lambda(t) \approx \langle A_E(t), \theta\rangle_E = \langle A_2(t), \theta\rangle_2 = \langle A_S(t), \theta\rangle_S.$$

$A_E(t)$ is most easily computed; from [3.1], its B-spline coefficients are given by

$$\mathbf{a}_e(t) = X'X[X'X + m\lambda\Omega_2]^{-1}\mathbf{e}(t)$$

where $\mathbf{e}(t) = (B_1(t), B_2(t), \cdots, B_K(t))'$ are the B-spline coefficients of the representer of evaluation at t (with respect to the Euclidean inner product). The B-spline coefficients, $\mathbf{a}_2(t)$ and $\mathbf{a}_s(t)$, of $A_2(t)$ and $A_S(t)$ are directly obtained from the $\mathbf{a}_e(t)$.

$$\mathbf{a}_2(t) = \Omega_0^{-1}\mathbf{a}_e(t); \quad \mathbf{a}_s(t) = [\Omega_0 + \Omega_2]^{-1}\mathbf{a}_e(t).$$

A word of caution. The matrices Ω_0 and $[\Omega_0 + \Omega_2]$ are poorly conditioned for K large. As a result a very stable method such as a singular value decomposition (see Dongarra et al., 1979) should be used to compute the inverses. From the L_2 representation, it follows that the Backus-Gilbert averaging kernel at t is approximately given by

$$A_2(t, s) \approx \sum_{k=1}^K a_{2k}(t)B_k(s).$$

The Sobolev representation for the averaging kernel is

$$A_s(t, s) \approx \sum_{k=1}^K a_{sk}(t)B_k(s).$$

Bias Computations. The average and maximum bias can be approximated directly in terms of the Euclidean representation for the averaging kernel in $\{B_k\}_{k=1}^K$. For example, the *maximum bias* in a ball of radius μ in $\Theta = W_2^2[a, b]$ is $\mu^2 b_M^2(t)$ where

$$b_M^2(t) \approx [\mathbf{e}(t) - \mathbf{a}_e(t)]'[\Omega_0 + \Omega_2]^{-1}[\mathbf{e}(t) - \mathbf{a}_e(t)].$$

4. APPLICATION TO NONLINEAR INVERSE PROBLEMS

The techniques described in Sections 2 and 3 will now be applied to the study of two nonlinear ill-posed inverse problems taken from satellite meteorology and reservoir engineering. In both cases the function, θ, of interest is restricted to be one-dimensional and the following MOR procedure is considered: θ is the

minimizer in $W_2^2[a, b]$ of

$$\frac{1}{m}\sum_{i=1}^{m}[z_i - \eta(x_i, \theta)]^2 + \lambda\int_a^b [\ddot{\theta}(t)]^2\,dt, \quad \lambda > 0.$$

In order to study performance characteristics, linearize the functionals $\eta(x_i, \cdot)$ about some value θ_0,

$$\eta(x_i, \theta) \approx \eta(x_i, \theta_0) + \nabla_\theta\eta(x_i, \theta_0)(\theta - \theta_0),$$

and consider the properties of a modified MOR procedure: θ is the minimizer of

$$\frac{1}{m}\sum_{i=1}^{m}[z_i^* - \nabla_\theta\eta(x_i, \theta_0)\theta]^2 + \lambda\int_a^b [\ddot{\theta}(t)]^2\,dt, \quad \lambda > 0.$$

where $z_i^* = z_i - \eta(x_i, \theta_0) + \nabla\eta(x_i, \theta_0)\theta_0$. The averaging kernel calculus can be applied to the modified MOR procedure and the results of this *linearized* analysis are presented below. A rigorous justification for the linearization is not attempted. (This is a very challenging problem and even an asymptotic analysis seems to be quite difficult. The theory presented in Cox and O'Sullivan (1985) may provide a starting point for further investigation of this topic.) It is assumed that the analysis will give reliable results whenever the degree of nonlinearity in the functionals η is low.

To compute averaging kernels, bias, and variability for the linearized problem, I use a B-spline basis and follow the development in Section 3. The number of basis elements is chosen so that any plots of averaging kernels and bias and variability characteristics are visually unchanged by the addition of extra basis elements. The design matrix, X, has the form

$$X_{ik} = \nabla_\theta\eta(x_i, \theta_0)B_k, \quad i = 1, 2, \cdots, m,$$

where B_k is the kth element of the B-spline basis. Two sets of system libraries were employed: computations for the satellite meteorology example were carried out on a DEC VAX 11/750 machine using the B-spline and Linpack routines which are part of the publicly available CMLIB; computations for the reservoir engineering example, required repeated numerical solutions of a diffusion equation, and these were carried out on a Boeing Computer Services Cray 1 machine using routines available in the BCSLIB.

4.1 Temperature Retrieval from Satellite Radiance Data

Intensities of radiation measured by modern meteorological satellites provide information about atmospheric characteristics such as temperature and moisture. This new database is becoming an increasingly important tool in the process of *nowcasting*, i.e., specifying the current state of the atmosphere, and *forecasting*, i.e., describing the future states of the atmosphere. Smith et al. (1979) and Smith (1983) describe the basic features of these measurement systems. A typical climatological temperature profile is given in Figure 4.1. The vertical axis is decreasing in pressure while the horizontal axis gives temperature. It is standard practice for meteorologists to plot things in this way because moving up the vertical axis then corresponds to going higher up in the atmosphere. The plot shows a temperature inversion near 20 mb; the temperature is generally increasing as one moves away from this point. Inversions are a characteristic feature of atmospheric temperature profiles. The location of this upper atmosphere temperature inversion is known as the tropopause height.

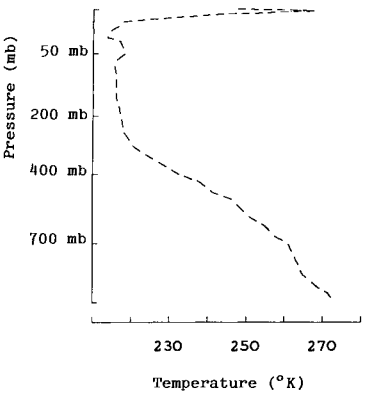

FIG. 4.1. *Typical climatological profile, T_0. The vertical scale is pressure in kappa (pressure$^{(5/8)}$) units. Note the temperature inversion high up in the atmosphere.*

The processing of radiance data to get temperature estimates involves the solution of an interesting inverse problem. The radiative transfer equations, Liou (1979), are used to model how the intensity of radiation, z_i, at frequency, ν_i, depends on the temperature profile, T, temperature as a function of pressure, in the column beneath the satellite;

$$z_i = R_{\nu_i}(T) + \varepsilon_i, \quad i = 1, \cdots, m,$$

where

$$R_\nu(t) = B_\nu(T(x_s))\tau_\nu(x_s) - \int_{x_0}^{x_s} B_\nu[T(x)]\dot{\tau}_\nu(x)\,dx$$

and x is some monotone transformation of pressure p; x_s corresponds to the surface and x_0 corresponds to the top of the atmosphere. Meterologists usually work in kappa units, i.e., $x(p) = p^{5/8}$, because atmospheric variations are believed to be slowly varying in this scale. $\tau_\nu(x)$ is the transmittance of the atmosphere above x at wave number ν, and B_ν is Planck's function given by:

$$B_\nu[T] = c_1\nu^3/[\exp(c_2\nu/T) - 1]$$

where

$$c_1 = 1.1906 \times 10^{-5}\text{erg cm}^2\text{ sec}^{-1}$$

and

$$c_2 = 1.43868\text{ cm/deg } (K).$$

The measurement errors, ε_i, are roughly mean zero and uncorrelated. However, different channels have different noise levels so by dividing through by relative noise levels in channels results in a generalized nonlinear regression model as in Section 1. The TIROS-N system, see Smith et al. (1979), has 15 channels ($m = 15$). Linearizing the $R_{\nu_i}(T)$ about the climatological profile, T_0, given in Figure 4.1 results in:

$$\nabla_T R_{\nu_i}(T_0)T = k_s(\nu_i)T(x_s) + \int_{x_s}^{x_0} k(\nu_i, x)T(s)\,dx,$$

where $k_s(\nu) = \dot{B}_\nu[T_0(x_s)]\tau_\nu(x_s)$ and $k(\nu, x) = -\dot{B}_\nu[T_0(x)]\dot{\tau}_\nu(x)$. Since the linearized functionals have an explicit form, the linearized design matrix is very easy to compute by numerical integration:

$$X_{ik} = \nabla_T R_{\nu_i}(T_0)B_k$$

$$= k_s(\nu_i)B_k(x_s) + \int_{x_s}^{x_0} k(\nu_i, x)B_k(x)\,dx,$$

where B_k is the kth element of the B-spline basis.

An averaging kernel at 700 mb for the MOR inversion is given in Figure 4.2. The corresponding bias and variability characteristics are given in Figure 4.3. Again these plots correspond to a particular value for the regularization parameter λ. Larger values for the smoothing parameter result in broader averaging kernels (more bias and less variance). Notice that the averaging kernel has sharp behavior near the surface. This is attributable to the microwave channels. The data obtained from these are nearly direct measurements of the surface *skin* temperature, $T(x_s)$. As a result the L_2 representers of the functionals corresponding to these channels are very spiked at the surface and since the averaging kernel is a linear combination of these representers, the behavior at the surface is to be expected. The effect of the microwave channels on the bias is also quite dramatic. The variability profile indicates regions near 600 mb and 200 mb where sampling density might be improved. However, there are physical constraints on selection of spectral wavelengths which make it difficult to get good coverage throughout the atmosphere (see Liou (1979), page 250 and following). The operating characteristics given in Figure 4.3 relate to maximum bias. In a meteorological setting, where there is a huge database of prior information on atmospheric variation, an average bias measure with respect to a climatological prior would be more appropriate. In spite of the fact that there is a dependence on the initial climatology, T_0, the retrieval characteristics predicted by the averaging kernel calculus is largely in agreement with those found by Monte Carlo simulation in O'Sullivan and Wahba (1985).

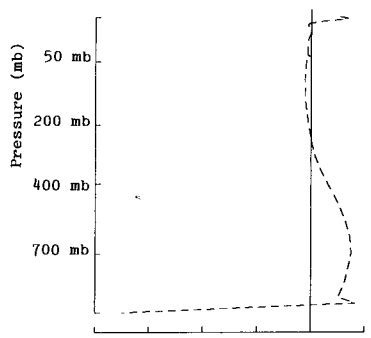

FIG. 4.2. *Averaging kernel at 700 mb for the temperature retrieval problem. The sharp behavior near the surface is attributable to the microwave channels.*

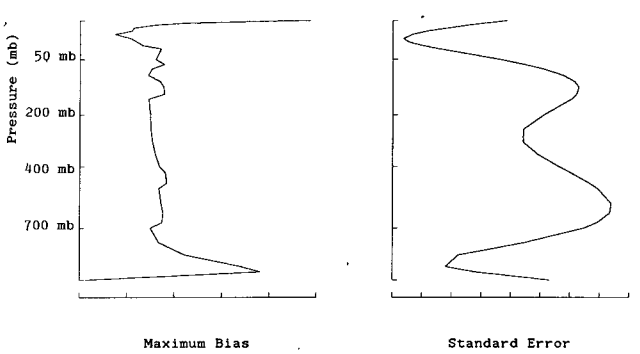

FIG. 4.3. *Inversion characteristics of a MOR procedure applied to the temperature retrieval problem.*

4.2 The History Matching Problem of Reservoir Engineering

The dynamic flow of fluid through a porous medium is usually modeled by a diffusion equation

$$\frac{\partial u(\mathbf{x}, t)}{\partial t} - \frac{\partial}{\partial \mathbf{x}}\left\{a(\mathbf{x})\frac{\partial u}{\partial x}(\mathbf{x}, t)\right\} = q(\mathbf{x}, t),$$

$$\mathbf{x} \text{ in } \Omega, \quad t \text{ in } [0, T],$$

subject to prescribed initial and boundary conditions. Here u is pressure, q accounts for the withdrawal or injection of fluid into the region Ω, and a is the transmissivity or conductance which determines the ease with which fluid flows through the medium. The initial condition is $u(\mathbf{x}, 0) = u_0(\mathbf{x})$ and a typical boundary condition is no fluid flow across the boundary of the region, i.e., $(\partial u/\partial w) = 0$, where w represents the direction normal to the boundary. The history matching problem arises as one tries to use scattered well data on $u(\mathbf{x}_i, t_j)$ and $q(\mathbf{x}_i, t_j)$, $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, l$, to infer the diffusion parameter, a; see Cooley (1982, 1983), Kravaris and Seinfeld (1985), and Neuman and Yakowitz (1979). This problem is an example of a broad class of inverse problems which arise in connection with partial differential equations. Such problems have attracted an amount of pure mathematical interest. See Anger (1979),

Lions (1971), McLaughlin (1983), and especially Payne (1975).

For a simplified version of the history matching problem, consider the situation discussed by Kravaris and Seinfeld (1985). Let $\Omega = [0, 1]$ and $T = 1$, assume there is no injection or withdrawal of fluid and that there is no flow across the boundary. Suppose that there are 10 measurement sites at $x_i = (5i - 3)/49$, $i = 1, 2, \cdots, 10$, and readings on u are made for $t_j = 0.0(.007)0.5$. These data are modeled by a nonlinear regression

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

where the errors are mean zero with constant variance.

The dependence of $u(x_i, t_j; a)$ on a is again nonlinear. By linearizing $u(x_i, t_j; a)$ about a plausible transmissivity profile, a_0, such as the one given in Figure 4.4, one can compute averaging kernels, etc. The true pressure history plays a significant role in determining the information recovered about transmissivity. Roughly speaking, gradients in the pressure history generate information about the transmissivity. A pressure history corresponding to the transmissivity profile in Figure 4.4 is given in Figure 4.5. This pressure history is driven by an initial pressure distribution which ranges from 10 at $x = 0.$ to 100 at $x = 1.$

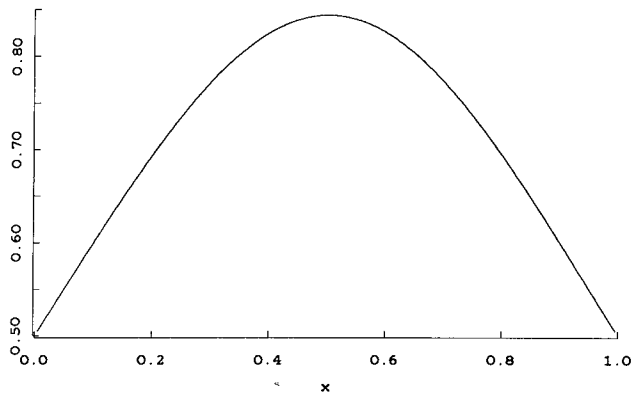

FIG. 4.4. *A transmissivity distribution, a_0, for the history matching problem.*

Computation of the Linearized Design Matrix. Unlike the temperature retrieval problem, there is not an explicit analytical representation for the observed functionals, and this makes the computation of the design matrix a bit more complicated.

$$X_{ijk} = \nabla_a u(x_i, t_j, a_0)B_k = \frac{\partial \hat{u}}{\partial a_k}(x_i, t_j; a_0),$$

where the gradient is taken in the direction of functions of the form $a(x) = \sum_{k=1}^{K} a_k B_k(x)$ and $\hat{u}$ is the solution to

$$\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left\{a(x)\frac{\partial u}{\partial x}\right\} = q(x, t) \quad \text{with} \quad (x, t) \quad \text{in} \quad [0, 1] \times [0, .5],$$

$$\text{subject to} \begin{cases} u(x, 0) = u_0(x), \\ \dfrac{\partial u(x, t)}{\partial x} = 0 \quad \text{for} \quad x = 0, 1. \end{cases}$$

Let $D(a)$: $U \to Q \times U_0 \times B_0 \times B_1$ denote the mapping

$$D(a)u = \left[\frac{\partial u}{\partial t} - \frac{\partial}{\partial x}\left\{a(x)\frac{\partial u}{\partial x}\right\}, u(\cdot, 0), \frac{\partial u(0, \cdot)}{\partial x}, \frac{\partial u(1, \cdot)}{\partial x}\right],$$

which takes pressure histories in U into the product space of forcing functions Q, initial pressure distributions U_0, and the $t = 0$ and $t = 1$ boundary value functions B_0 and B_1. Under regularity, the implicit function theorem implies that the inverse of $D(a)$,

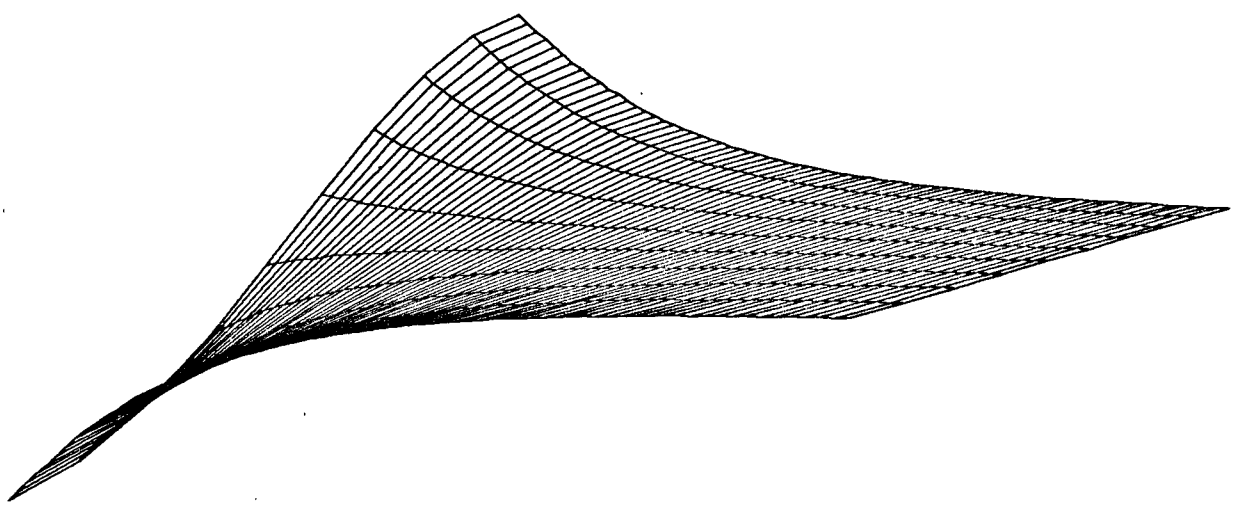

FIG. 4.5. *True pressure history. The initial distribution ranges from* 10–100 *units. Notice how the distribution flattens out in time. Data accumulated on pressure at later times will tend to be less informative about transmissivity.*

denoted $G(a)$, exists and is differentiable in a neighborhood of a_0.

$$G(a)[q, u_0, 0, 0] = \hat{u}(\cdot, \cdot; a).$$

The relation $G(a)D(a) = I$ may be used to obtain $\partial\hat{u}(x_i, t_j; a_0)/\partial a_k$. Differentiating,

$$\frac{\partial}{\partial a_k}\{G(a)D(a)\} = \frac{\partial G(a)}{\partial a_k} D(a) + G(a) \frac{\partial D(a)}{\partial a_k} = 0.$$

Thus

$$\frac{\partial G(a)}{\partial a_k} = -G(a) \frac{\partial D(a)}{\partial a_k} G(a),$$

which implies

$$\frac{\partial \hat{u}}{\partial a_k} = -G(a)\left\{\frac{\partial D(a)}{\partial a_k} \hat{u}\right\}.$$

But

$$\frac{\partial D(a)}{\partial a_k} u = \left[-\frac{\partial}{\partial x}\left\{B_k(x) \frac{\partial u}{\partial x}\right\}, 0, 0, 0\right],$$

and the expression simplifies to

$$\frac{\partial \hat{u}(x_i, t_j; a_0)}{\partial a_k} = G(a_0)\left[\frac{\partial}{\partial x}\left\{B_k(x) \frac{\partial u}{\partial x}\right\}, 0, 0, 0\right].$$

Hence X_{ijk} can be found by solving the original diffusion equation with the forcing term, q, replaced by $(\partial/\partial x)\{B_k(x)(\partial u/\partial x)\}$ and the initial pressure distribution, u_0, replaced by the constant 0.

By this method, the computation of the entire linearized design matrix requires K separate numerical solutions of the diffusion equation (K being the number of basis elements). Although this is a generally applicable technique, it is rather inelegant. Since the solution corresponding to B_k is likely to be "near" the solution corresponding to B_{k+1}, it may be possible to improve computational efficiency by using some form of relaxation. This is currently being investigated. For the time invariant problem, Neuman and Yakowitz (1979) use properties of a particular finite difference scheme to develop a fast method for computing the analogue of the design matrix. A further approach to this problem, relying on an optimal control formulation, is employed by Kravaris and Seinfeld (1985).

Linearized Averaging Kernels and Retrieval Characteristics. Figures 4.6 and 4.7 give linearized averaging kernels and retrieval characteristics. One would suspect that information about a should depend critically on the true pressure history. If the initial pressure were constant, then, since there would be no pressure gradients, there would be no lateral flow. However, the functional parameter is a transmissivity and information about transmissivity can only be generated by lateral flow. The averaging kernel and the bias and variability characteristics show that for the given pressure history, the greatest detail on transmissivity is recovered near the middle of the x range. Changing the pressure history results in different retrieval properties—the bias and variability calculus seems to make very good physical sense.

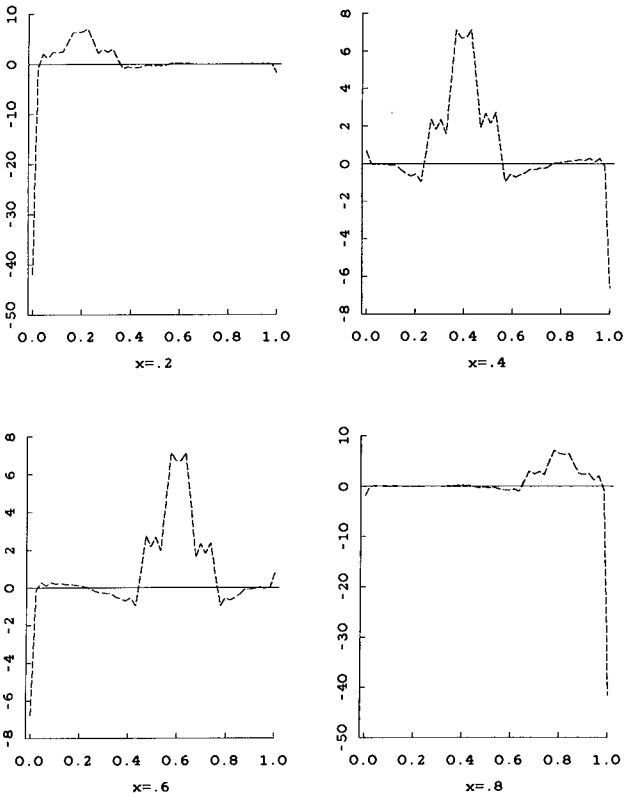

FIG. 4.6. *Sample averaging kernels for a MOR procedure applied to the history matching problem. The resolution is better near the center. The assumed true pressure history causes the averaging kernels to have sharp behavior near the boundary.*

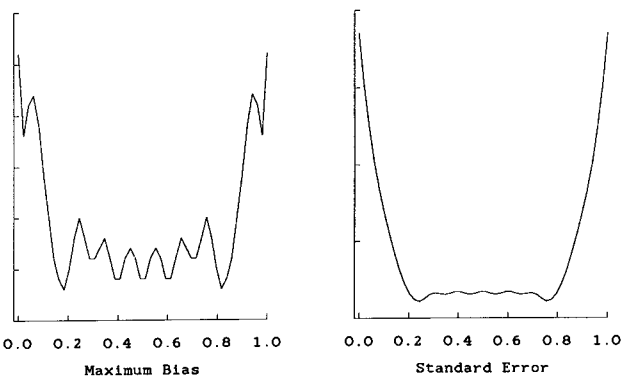

FIG. 4.7. *Inversion characteristics of a MOR procedure applied to the history matching problem. Retrieval properties are best near the middle of the x range.*

5. EMPIRICAL SELECTION OF SMOOTHING PARAMETERS

Nearly all inversion algorithms, and in particular the MOR, have explicit or implicit smoothing/tuning

parameters, corresponding to the signal to noise ratio which, in a practical setting, have to be selected by the user. Two of the more popular techniques to emerge from the statistical literature on this problem are the methods of cross-validation and unbiased risk estimation. Typically, these techniques try to find that value of the smoothing parameter which minimizes the predictive mean square error (PMSE),

$$\text{PMSE}(\lambda) = \frac{1}{m}\sum_{i=1}^{m}[\eta(x_i, \theta) - \eta(x_i, \hat{\theta}_\lambda)]^2,$$

where θ and $\hat{\theta}_\lambda$ are the true and estimated functions. (Throughout this section λ will be used to denote the smoothing/tuning parameters of the inversion algorithm.) The predictive mean square (PMSE) is a convenient criterion but not necessarily a criterion of real intrinsic interest. Inverse problems are focused on a particular function, so it maybe more appropriate to phrase a loss directly in terms of that function. Thus it might be of interest to consider a loss of the form

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

where the t_i values are some points of direct interest to the investigator. For certain loss functions, which are *estimable* in the sense described below, it is possible to develop refined versions of the cross-validation and unbiased risk assessments. It should be pointed out that the PMSE has a certain robustness property; it is often the case that the best choice of the smoothing parameter, from the point of view of PMSE loss, is very nearly optimal from the point of view of other loss functions also (see Cox, 1983; Lukas, 1981; Ragozin, 1981; and Wahba, 1979). Thus the practical need for refined procedures will only arise in situations where the PMSE and the loss function of interest have very different minimizers.

It is assumed throughout that inversion algorithms under consideration are linear and that the functionals $\eta(x_i, \theta)$ are themselves linear in θ. For some extensions to nonlinear inversion algorithms see O'Sullivan and Wahba (1985), O'Sullivan, Yandell, and Raynor (1986), Villalobos and Wahba (1983), and Wahba (1984). A brief review of cross-validation and unbiased risk in the standard predictive mean square error context is given next. For a more detailed discussion of these methods see the recent review of Titterington (1985).

5.1 Empirical Assessment of PMSE

The *Hat matrix*, $H(\lambda)$, is defined to be the matrix that maps data $\mathbf{z}$ into predictions $\hat{\mathbf{z}}$,

$$\hat{\mathbf{z}} = H(\lambda)\mathbf{z}.$$

For a linear inversion algorithm the Hat matrix is obtained from the impulse response functions, $s_i^{(\lambda)}$, via

$$H_{ij}(\lambda) = \eta(x_i, s_j^{(\lambda)}).$$

Here the dependence of the inversion on the smoothing/tuning parameters λ is highlighted—the impulse response functions are functions of λ. The trace of the Hat matrix occurs in both the cross-validation and unbiased risk assessments of the PMSE. I begin with the unbiased risk method which is easier to describe.

Unbiased Risk. This procedure assumes that one has a reliable estimate of the noise level σ. Given σ, the procedure uses the residual sum of squares (RSS) to construct an unbiased estimate of the PMSE risk. The basic steps are as follows:

$$\begin{aligned}\text{PMSE}(\lambda) &= \frac{1}{m}\sum_{i=1}^{m}[\eta(x_i, \theta) - \hat{z}_i]^2 \\ &\equiv \frac{1}{m}\|\eta(\theta) - \eta(\hat{\theta}_\lambda)\|_m^2,\end{aligned}$$

i.e., $\eta(\cdot) = (\eta(x_1, \cdot), \eta(x_2, \cdot), \cdots, \eta(x_m, \cdot))'$. By linearity, the expected value of the predictive mean square error is

$$\begin{aligned}mE\{\text{PMSE}(\lambda)\} &= \|[I - H(\lambda)]\eta(\theta)\|_m^2 \\ &\quad + \sigma^2\text{tr}[H(\lambda)'H(\lambda)] \\ &= \text{BIAS}^2(\lambda) + \sigma^2\text{tr}[H(\lambda)'H(\lambda)].\end{aligned}$$

Meanwhile,

$$\text{RSS}(\lambda) = \sum_{i=1}^{m}[z_i - \hat{z}_i]^2 \equiv \|[I - H(\lambda)]\mathbf{z}\|_m^2.$$

So the expected value is

$$\begin{aligned}E\{\text{RSS}(\lambda)\} &= \|[I - H(\lambda]\eta(\theta)\|_m^2 \\ &\quad + \sigma^2\text{tr}[[I - H(\lambda)]'[I - H(\lambda)]] \\ &= \text{BIAS}^2(\lambda) + \sigma^2\text{tr}[[I - H(\lambda)]'[I - H(\lambda)]].\end{aligned}$$

Combining these formulae gives that

$$\widehat{\text{PMSE}}(\lambda) = \frac{1}{m}\text{RSS}(\lambda) - \sigma^2 + 2\sigma^2\frac{\text{tr}[H(\lambda)]}{m}$$

is an unbiased estimate of the predictive mean square error. In the standard regression context $H = X(X'X)^{-1}X'$ and $\widehat{\text{PMSE}}$ reduces to the C_p statistic of Mallows (1973).

Cross-validation. In cross-validation one considers a leave-out-one prediction, $\hat{z}_{-i}$, which is defined to be the prediction of $\eta(x_i, \theta)$ from an estimator constructed from data with the ith data point, z_i, omitted. The idea being that if the prediction rule is really good (λ well chosen), then $\hat{z}_{-i}$ should be reasonably close to z_i on average. Ordinary cross-validation or Allen's predictive sum of squares (PRESS) (see Allen, 1974)

is defined to be

$$V_0(\lambda) = \frac{1}{m}\sum_{i=1}^{m}[z_i - \hat{z}_{-i}]^2.$$

For the MOR estimators in Section 1.1, where $H(\lambda) = X[X'X + m\lambda W]^{-1}X'$, a rank one update formula gives that

$$\hat{z}_{-i} = z_i - \frac{(z_i - \hat{z}_i)}{(1 - h_{ii}(\lambda))}$$

where $h_{ii}(\lambda)$ is the ith diagonal element of the Hat matrix, also known as the ith leverage value. It follows that the ordinary cross-validation assessment is a weighted residual sum of squares

$$V_0(\lambda) = \frac{1}{m}\sum_{i=1}^{m}\left[\frac{z_i - \hat{z}_i}{1 - h_{ii}(\lambda)}\right]^2.$$

If instead of dividing residuals by $1 - h_{ii}(\lambda)$, one divides by the mean value, $1 - \bar{h}_{ii}(\lambda)$, then the generalized cross-validation (GCV) of Craven and Wahba (1979) is obtained. The GCV score is usually written as

$$V(\lambda) = \frac{(1/m)\sum_{i=1}^{m}[z_i - \hat{z}_i]^2}{[1 - (1/m)\mathrm{tr}\, H(\lambda)]^2}.$$

Since

$$\frac{V(\lambda)}{m} = \frac{\mathrm{RSS}(\lambda)}{[1 - (1/m)\mathrm{tr}\, H(\lambda)]^2},$$

in a regression context, where $H = X(X'X)^{-1}X'$, the GCV score is proportional to the residual mean square divided by the degrees of freedom for error. Thus the GCV score reduces to a model selection statistic proposed by Anscombe (1967) and simplified by Tukey (see Mosteller and Tukey, 1977, page 386 and following).

A great deal is known about the asymptotic behavior of the above empirical assessment methods. The typical result says that the minimizer of the empirical assessment tends to minimize the PMSE, in large samples; Monte Carlo simulation results show that a similar property tends to hold in finite samples, see for example Craven and Wahba (1979), Golub, Heath, and Wahba (1979), Nychka et al. (1984), Rice (1984), and Speckman (1982).

5.2 Empirical Assessment of Estimable Losses

Borrowing from the terminology of the standard linear model define an estimable functional as follows:

DEFINITION. *A continuous linear functional* $\langle \xi, \cdot \rangle$ *is estimable if there exists* $\mathbf{c}$ *in* R^m *such that* $E\mathbf{c}'\mathbf{z} = \langle \xi, \theta \rangle$ *for all* θ *in* Θ.

Clearly, since $\eta(x_i, \cdot)$ are continuous linear functionals, $\langle \xi, \cdot \rangle$ is estimable if and only if there is some $\mathbf{c}$ in R^m such that

$$\xi = \sum_{i=1}^{m} c_i \xi_{x_i}$$

where ξ_{x_i} are representers of $\eta(x_i, \cdot)$ in Θ. With this a loss function is *estimable* if it is defined in terms of *estimable* functionals. A referee has pointed out that asymptotic estimability could also be of interest in certain circumstances. For a discussion of a particular instance of this see Rice (1986).

To illustrate how to empirically assess estimable losses consider a particular estimable loss of the form

$$L(\lambda) = \sum_{j=1}^{T}\langle \xi_j, \theta - \hat{\theta}_\lambda \rangle^2$$

where ξ_j are all estimable. Let $\mathbf{c}_j$ be such that

$$\xi_j = \sum_{j=1}^{m} c_{ji}\xi_{x_i}, \quad j = 1, \cdots, T.$$

A simple modification of the cross-validation and unbiased risks techniques can be formulated to directly assess the loss $L(\lambda)$.

Unbiased Risk Assessment of $L(\lambda)$. By a development similar to that used in Rice (1986), the expected value of $L(\lambda)$ is

$$\begin{aligned} E[L(\lambda)] &= \sum_{j=1}^{T}\langle \xi_j, \theta - E\hat{\theta}_\lambda \rangle^2 \\ &\quad + \sigma^2 \sum_{j=1}^{T} \mathbf{c}_j'[H(\lambda)'][H(\lambda)]\mathbf{c}_j \\ &= \mathrm{BIAS}^2(\lambda) + \sigma^2 \sum_{j=1}^{T} \mathbf{c}_j'[H(\lambda)'][H(\lambda)]\mathbf{c}_j. \end{aligned}$$

The sum of squares,

$$\mathrm{SS} = \sum_{j=1}^{T}[\mathbf{c}_j'\mathbf{z} - \mathbf{c}_j'\hat{\mathbf{z}}]^2,$$

has expected value

$$E[\mathrm{SS}] = \mathrm{BIAS}^2(\lambda) + \sigma^2 \sum_{j=1}^{T} \mathbf{c}_j'[I - H(\lambda)'][I - H(\lambda)]\mathbf{c}_j.$$

Thus

$$\hat{L}(\lambda) = \mathrm{SS} - \sigma^2 \sum_{j=1}^{T} \mathbf{c}_j'\mathbf{c}_j + 2\sigma^2 \sum_{j=1}^{T} \mathbf{c}_j' H(\lambda)'\mathbf{c}_j$$

is an unbiased estimate of $L(\lambda)$. Again note that a reliable estimate of σ is necessary in order to be able to use this assessment.

Cross-validation Assessment of $L(\lambda)$. The cross-validation procedure is more complicated to derive. Instead of leaving out one data point, one now omits the $\mathbf{c}_j$th component of the data and uses the remaining data to develop a prediction, $\widehat{\mathbf{cz}}_{-j}$, of $\langle \xi_j, \theta \rangle$. The

cross-validation assessment then compares $\mathbf{c}_j'\mathbf{z}$ to $\widehat{\mathbf{cz}}_{-j}$, i.e.,

$$V_0(\lambda) = \sum_{j=1}^{T} [\mathbf{c}_j'\mathbf{z} - \widehat{\mathbf{cz}}_{-j}]^2.$$

In the context of MOR estimators the situation becomes clearer. To make the notation less complicated, suppose the $\mathbf{c}_j$ are normalized so that $\mathbf{c}_j'\mathbf{c}_j = 1$ for $j = 1, 2, \cdots, T$. Let

$$P_j = I - \mathbf{c}_j\mathbf{c}_j'.$$

P_j is a projection onto the space orthogonal to $\mathbf{c}_j$. The estimator obtained by removing the $\mathbf{c}_j$th component minimizes

$$\frac{1}{m}[P_j(\mathbf{z} - \eta(\theta))]'P_j^-[P_j(\mathbf{z} - \eta(\theta))] + \lambda \langle \theta, W\theta \rangle$$

where P_j^- is the generalized inverse of P_j. Since $P_jP_j^-P_j = P_j$, it follows that

$$\widehat{\mathbf{cz}}_{-j} = \mathbf{c}_j'[X'P_jX + m\lambda W]^{-1}X'P_j\mathbf{z}.$$

Again, using a rank one update formula, it follows after some algebra that

$$\widehat{\mathbf{cz}}_{-j} = \mathbf{c}_j'\mathbf{z} - \frac{\mathbf{c}_j'\mathbf{z} - \mathbf{c}_j'\hat{\mathbf{z}}}{[1 - \mathbf{c}_j'H(\lambda)c_j]}$$

where $H(\lambda) = X[X'X + m\lambda W]^{-1}X'$. Thus the ordinary cross-validation score is

$$V_0(\lambda) = \sum_{j=1}^{T}\left[\frac{\mathbf{c}_j'\mathbf{z} - \mathbf{c}_j'\hat{\mathbf{z}}}{1 - \mathbf{c}_j'H(\lambda)\mathbf{c}_j}\right]^2,$$

and the GCV extension would be

$$V(\lambda) = \frac{1/T\sum_{j=1}^{T}[\mathbf{c}_j'\mathbf{z} - \mathbf{c}_j'\hat{\mathbf{z}}]^2}{[1 - \overline{\mathbf{c}'H(\lambda)\mathbf{c}}]^2},$$

where

$$\overline{\mathbf{c}'H(\lambda)\mathbf{c}} = \frac{1}{T}\sum_{j=1}^{T}\mathbf{c}_j'H(\lambda)\mathbf{c}_j.$$

The advantage of the cross-validation assessment over the unbiased risk assessment is that the cross-validation method does not require knowledge of σ.

ACKNOWLEDGMENTS

I was very fortunate to learn about inverse problems from Professor Grace Wahba at the University of Wisconsin. The motivation for this article arose out of a series of lectures I gave in a seminar on "Inverse Problems and Statistical Signal Models" jointly organized with David L. Donoho at Berkeley in the Spring of 1985. I would like to thank the participants in the seminar and also D. L. Banks, B. A. Bolt, D. R. Brillinger, J. A. Rice, G. Wahba, two referees, and especially Morris H. DeGroot, for many useful comments which have lead to substantial improvements in the paper. This work was supported by the National Science Foundation under Grant MCS-8403239.

REFERENCES

ADAMS, R. (1975). *Sobolev Spaces.* Academic, New York.

AKI, K. and RICHARDS, G. (1980). *Quantitative Seismology: Theory and Methods.* Freeman, San Francisco.

ALLEN, D. M. (1974). The relationship between variable selection and data augmentation and a method of prediction. *Technometrics* **16** 125–127.

ANGER, G., ed. (1979). *Inverse and Improperly Posed Problems in Differential Equations: Proceedings of the Conference on Mathematical and Numerical Methods.* Akademie-Verlag, Berlin.

ANSCOMBE, F. J. (1967). Topics in the investigation of linear relations fitted by least squares (with discussion). *J. Roy. Statist. Soc. Ser. B* **29** 1–52.

BACKUS, G. and GILBERT, F. (1968). The resolving power of gross earth data. *Geophys. J. Roy. Astronom. Soc.* **266** 169–205.

BACKUS, G. and GILBERT, F. (1970). Uniqueness in the inversion of inaccurate gross earth data. *Philos. Trans. Roy. Soc. London Ser. A* **266** 123–192.

BOLT, B. A. (1980). What can inverse problems do for applied mathematics and the sciences? *Search* **11** 6.

BOX, G. E. P. and DRAPER, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Assoc.* **54** 622–653.

BUDINGER, T. F. (1980). Physical attributes of single-photon tomography. *J. Nucl. Med.* **21** 6.

COOLEY, R. L. (1982). Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 1. Theory. *Water Resour. Res.* **18** 965–976.

COOLEY, R. L. (1983). Incorporation of prior information on parameters into nonlinear regression groundwater flow models, 2. Applications. *Water Resour. Res.* **19** 662–676.

COX, D. D. (1983). Approximation of the method of regularization estimators. Technical Report 723, Statistics Dept., Univ. Wisconsin-Madison.

COX, D. D. and O'SULLIVAN, F. (1985). Analysis of penalized likelihood type estimators with application to generalized smoothing in Sobolev spaces. Technical Report 51, Statistics Dept., Univ. California-Berkeley.

CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.* **31** 377–403.

DEBOOR, C. (1978). *A Practical Guide to B-Splines.* Springer, New York.

DONGARRA, J. J., BUNCH, J. R., MOLER, C. B. and STEWART, G. W. (1979). *Linpack User's Guide.* SIAM, Philadelphia.

FOSTER, M. R. (1961). An application of the Weiner–Kolmogorov smoothing theory to matrix inversion. *J. SIAM* **9** 387–392.

FRANZONE, P. C., TACCARDI, B. and VIGANOTTI, C. (1977). An approach to inverse calculation of epi-cardial potentials from body surface maps. *Adv. Cardiol.* **21** 167–170.

GOLOMB, M. and WEINBERGER, H. (1959). Optimal approximation and error bounds. In *On Numerical Approximation* (R. Langer, ed.). Univ. Wisconsin Press, Madison.

GOLUB, G., HEATH, M. and WAHBA, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21** 215–223.

GOOD, I. J. and GASKINS, R. A. (1971). Non-parametric roughness penalties for probability densities. *Biometrika* **58** 255–277.

GRENANDER, U. (1981). *Abstract Inference.* Wiley, New York.

GRUNBAUM, F. A. (1975). Remark on the phase problem in crystallography. *Proc. Nat. Acad. Sci. U. S. A.* **72** 1699–1701.

HADAMARD, J. (1923). *Lectures on Cauchy's Problem.* Yale Univ. Press, New Haven, Conn.

JAYNES, E. T. (1983). *Papers on Probability, Statistics and Statistical Physics.* Synthese Library.

JEFFREYS, H. (1976). *The Earth.* Cambridge Univ. Press.

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

KRAVARIS, C. and SEINFELD, J. H. (1985). Identification of parameters in distributed parameter systems by regularization. *SIAM J. Control Optim.* **23** 217–241.

LATTES, R. and LIONS, J. L. (1969). *The Method of Quasi-reversibility, Applications to Partial Differential Equations.* Elsevier, New York.

LI, K. C. (1982). Minimaxity of the method of regularization on stochastic processes. *Ann Statist.* **10** 937–942.

LIONS, J. L. (1971). *Optimal Control of Systems Governed by Partial Differential Equations.* Springer, Berlin.

LIOU, K. (1979). *Introduction to Atmospheric Radiation.* Academic, London.

LUKAS, M. (1981). Regularization of linear operator equations. Unpublished Ph.D. thesis, Australian National Univ.

MALLOWS, C. L. (1973). Some comments on C_p. *Technometrics* **15** 661–675.

MCLAUGHLIN, D. W. (1983). *Inverse Problems: Proceedings of a Symposium in Applied Mathematics.* Amer. Math. Soc., Providence, R. I.

MENDELSOHN, J. and RICE, J. (1982). Deconvolution of microfluorometric histograms with B-splines. *J. Amer. Statist. Assoc.* **77** 748–753.

MOSTELLER, F. and TUKEY, J. W. (1977). *Data Analysis and Regression.* Addison-Wesley, Reading, Mass.

NEUMAN, S. P. and YAKOWITZ, S. (1979). A statistical approach to the inverse problem of aquifer hydrology, 1. Theory. *Water Resour. Res.* **15** 845–860.

NYCHKA, D., WAHBA, G., GOLDFARB, S. and PUGH, T. (1984). Cross-validated spline methods for the estimation of three-dimensional tumor size distributions from observations on two-dimensional cross-sections. *J. Amer. Statist. Assoc.* **79** 832–846.

O'SULLIVAN, F. and WAHBA, G. (1985). A cross-validated Bayesian retrieval algorithm for non-linear remote sensing experiments. *J. Comput. Phys.* **59** 441–455.

O'SULLIVAN, F., YANDELL, B. and RAYNOR, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *J. Amer. Statist. Assoc.* **81** 96–104.

PAYNE, L. E. (1975). *Improperly Posed Problems in Partial Differential Equations.* SIAM, Philadelphia.

PEANO, G. (1914). Residuo in formulas de quadratura. *Mathesis* **4** 5–10.

RAGOZIN, D. (1981). Error bounds for derivative estimates based on spline smoothing of exact or noisy data. Technical Report GN-50, Statistics Dept., Univ. Washington.

RICE, J. A. (1984). Bandwidth choice for nonparametric regression. *Ann. Statist.* **12** 1215–1230.

RICE, J. A. (1986). Bandwidth choice for nonparametric differentiation. *J. Mult. Anal.*, in press.

RUDIN, W. (1976). *Principles of Mathematical Analysis.* McGraw-Hill, New York.

SARD, A. (1949). Best approximate integration formulas; best approximation formulas. *Amer. J. Math.* **71** 80–91.

SMITH, W. (1983). The retrieval of atmospheric profiles from VAS geostationary radiance observations. *J. Atmospheric Sci.* **40** 2025–2035.

SMITH, W. L., WOOLF, H. M., HAYDEN, C. M., WARK, D. Q. and MCMILLIN, L. M. (1979). The TIROS-N operational vertical sounder. *Bull. Amer. Meteor. Soc.* **10** 1177–1187.

SPECKMAN, P. (1979). Minimax estimates of linear functionals in Hilbert space. Dept. Mathematics, Univ. Oregon.

SPECKMAN, P. (1982). Efficient nonparametric regression with cross-validated smoothing splines. Technical Report 45, Statistics Dept., Univ. Missouri-Columbia.

STEINBERG, D. M. and HUNTER, W. G. (1984). Experimental design: review and comment (with discussion). *Technometrics* **26** 71–130.

STRAND, O. N. and WESTWATER, E. R. (1968). Minimum-rms estimation of the numerical solution of a Fredholm integral equation of the first kind. *SIAM J. Numer. Anal.* **5** 287–295.

TIKHONOV, A. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.* **5** 1035–1038.

TIKHONOV, A. and ARSENIN, V. (1977). *Solutions of Ill-Posed Problems.* Wiley, New York.

TITTERINGTON, D. M. (1985). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev.* **53** 141–170.

VARDI, Y., SHEPP, L. A. and KAUFMAN, L. (1985). A statistical model for positron emission tomography (with discussion). *J. Amer. Statist. Assoc.* **80** 8–37.

VILLALOBOS, M. and WAHBA, G. (1983). Multivariate thin plate spline estimates for the posterior probabilities in the classification problem. *Commun. Statist. A* **12** 1449–1479.

WAHBA, G. (1979). Smoothing and ill-posed problems. In *Solution Methods for Integral Equations with Applications* (M. Goldberg, ed.) 183–194. Plenum, New York.

WAHBA, G. (1983a). Design criteria and eigensequence plots for satellite computed tomography. Technical Report 732, Statistics Dept., Univ. Wisconsin-Madison.

WAHBA, G. (1984). Cross-validated spline methods for the estimation of multivariate functions from data on functionals. In *Statistics: An Appraisal* (H. A. David and H. T. David, eds.) 205–233. Iowa State Univ. Press, Ames.

WEINREB, M. P. and CROSBY, D. S. (1972). Optimization of spectral intervals for remote sensing of atmospheric temperature profiles. *Remote Sens. Environ.* **2** 193–201.