WATER RESOURCES RESEARCH, VOL. 33, NO. 10, PAGES 2263–2272, OCTOBER 1997

# The minimum structure solution to the inverse problem

Peter K. Kitanidis
Stanford University, Stanford, California

**Abstract.** The inverse problem of estimating the conductivity function from head observations is generally ill posed: Many conductivity functions are consistent with the data. It is widely accepted now that a well-defined estimate can be obtained only if additional information about the function structure is introduced into the problem formulation. This work presents a method to obtain a stable and reasonable estimate that utilizes only the data and the flow or transport model with the minimum possible suppositions about the unknown function or its structure. The motivation is to develop a solution that has only characteristics that are traced directly to the data and the flow or transport model, without taking advantage of spatial continuity or other "prior information." The solution is obtained by minimizing the upper bound to the error, or, in a stochastic conceptual framework, as the most likely solution given the data. This solution, although generally not the most accurate since it neglects to utilize structural information that may be available, is of fundamental importance and may be useful as a benchmark. For example, by comparing this solution with other solutions, one can become aware of how prior information or the model of spatial structure affects the solution to the inverse problem.

## 1. Background

The equation of steady flow in heterogeneous isotropic porous media is

$$\frac{\partial}{\partial x_i}\left(K\frac{\partial \phi}{\partial x_i}\right) = -N \tag{1}$$

where $K$ is conductivity, $\phi$ is hydraulic head, $N$ represents sources and sinks, and $x_i$ is a spatial coordinate. $K$, $\phi$, and $N$ vary in space. Summation is implied over an index that appears exactly twice in a term. Consider the following inverse problem: Given $N$, appropriate boundary conditions, and sparse observations of head and perhaps conductivity, estimate the conductivity function. Reviews of the extensive literature on such inverse problems are given by *Yeh* [1986], *Ginn and Cushman* [1990], *Sun* [1994], and others.

For practical reasons we will seek not the conductivity but its logarithm (known as the log conductivity): The obvious advantage is that at the very outset we restrict our search to nonnegative conductivity functions. In most practical applications the domain is discretized in the process of solving the flow equation through finite difference or finite element methods. To fix ideas, although this discussion does not require the discretization of the domain, consider that the domain has been discretized and the conductivity function is represented through a vector. Thus we search for an $m$ by 1 vector $\mathbf{s}$ of log conductivity values, for example, the log conductivity values at the $m$ elements of a finite element model.

Similarly, the observations are arranged into an $n$ by 1 vector $\mathbf{y}$ of observations. The relation between log conductivity and observations may be represented in the following observation equations:

$$y_i = h_i(s_1, \cdots, s_m) + v_i, \qquad i = 1, \cdots, n \tag{2}$$

Copyright 1997 by the American Geophysical Union.

Paper number 97WR01619.
0043-1397/97/97WR-01619$09.00

or, in vector notation,

$$\mathbf{y} = \mathbf{h}(\mathbf{s}) + \mathbf{v} \tag{3}$$

Disregarding for a moment $\mathbf{v}$, this equation expresses that for a given $\mathbf{s}$ one may find $\mathbf{y}$; for example, the head at an observation point can be computed from the groundwater flow model for given log conductivity. The role of $\mathbf{v}$ becomes clear when we consider that a model prediction does not have to match exactly the observation on account of observation errors but also model imperfections. Thus $\mathbf{v}$ represents deviations between values measured in the field and those predicted by the mathematical model and is referred to as measurement error.

In least squares fitting we select the values $s_1, \cdots, s_m$ that minimize the sum of squares of differences between measured values and the corresponding predictions:

$$\min_{s_1, \ldots, s_m} \sum_{i=1}^{n} (y_i - h_i(s_1, \cdots, s_m))^2 \tag{4}$$

or, in vector notation,

$$\min_{\mathbf{s}} (\mathbf{y} - \mathbf{h}(\mathbf{s}))^T(\mathbf{y} - \mathbf{h}(\mathbf{s})) \tag{5}$$

where the superscript $T$ means matrix transpose. For the sake of generality, in order to account for nonuniform magnitudes and tendencies for simultaneous change (i.e., correlation) in the observation errors, a weighted least squares approach is adopted

$$\min_{\mathbf{s}} (\mathbf{y} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{h}(\mathbf{s})) \tag{6}$$

where $\mathbf{R}$ is an $n$ by $n$ positive definite matrix.

Such least squares formulations to the inverse problem are quite popular. One of the earliest applications was given by *Jahns* [1966]. In groundwater hydrology a well-known approach is the nonlinear regression method [*Cooley*, 1977]. In

this method the number $m$ of unknown variables cannot exceed the number of observations. In practice, in the nonlinear regression approach the number of observations must exceed the number of unknown parameters sufficiently for the error analysis to make sense. Such least squares problems are called algebraically overdetermined. The nonlinear regression approach is a sound and practical approach that is proper for solving algebraically overdetermined problem. In applications the challenge is in how to represent the log conductivity function through a small number of parameters. The usual approach is called "zonation" and consists of dividing the domain into a small number of homogeneous zones; zonation is typically much coarser than the subdivision imposed by the grid of the numerical model that is used for the solution of the flow equation. This sort of parameterization undoubtedly constrains the resolution of the log conductivity estimate. This method should work superbly when available information allows the arrangement of the formation into a few zones that are homogeneous for practical purposes. However, the selected zonation predisposes the solution and even suppresses information in the data about important heterogeneities when the arrangement into zones turns out to be particularly unfortunate. In addition, this approach is not intended for stochastic analysis, for example, generation of conditional realizations that may represent small-scale (subzone) variability.

In applications, it is desirable to achieve the maximum resolution of log conductivity that is possible consistent with the mathematical model. For example, in a finite element formulation it is preferable to allow every element to have a different log conductivity; in analytical formulations that involve no discretization, one may view log conductivity as a function defined over the domain occupied by the formation, which in an intuitive sense is equivalent to a vector with infinite entries (i.e., mathematically speaking, a function is a point in a Hilbert space the same way that a vector is a point in a Euclidean space). The salient and fundamental feature of this formulation, which is the focus of this work, is that the number of unknown values $m$ is much larger than the number of observations $n$, $m \gg n$. But this means that there are many solutions that reproduce the data because there are more unknowns than measurement equations that can be used for the determination of the unknowns. Such problems are known as algebraically underdetermined. Thus the criterion of (6) does not have a unique minimum, and the minimization yields a solution that depends on the initial guess, number of iterations, computer accuracy, etc. Solutions typically contain spurious features, that is, characteristics that although consistent with the data appear only in particular solutions. It is unreasonable to pick at random one of these solutions and anoint it "best estimate." Instead, common sense dictates that a best estimate should include only features that are somehow common to all mathematical solutions.

*Neuman* [1973] argued that the best estimate should be selected by considering a second objective: physical plausibility. This may be achieved by adding to the term that penalizes deviations of predictions from observations, another term that penalizes deviations of log conductivity from what other information suggests as a physically reasonable solution. *Gavalas et al.* [1976] arrived at a combined criterion of the form

$$\min_{\mathbf{s}} \{(\mathbf{y} - \mathbf{h}(\mathbf{s}))^T \mathbf{R}^{-1} (\mathbf{y} - \mathbf{h}(\mathbf{s})) + (\mathbf{s} - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\mathbf{s} - \boldsymbol{\mu})\} \quad (7)$$

Interpreting the Bayesian approach of Gavalas et al. in intuitive terms, the vector $\boldsymbol{\mu}$ represents a "prior" estimate of $\mathbf{s}$, which means based on data other than $\mathbf{y}$. The matrix $\mathbf{Q}$ represents the accuracy of this prior estimate. In the combined criterion the first term penalizes deviations of model predictions from observations and the second term penalizes deviations of log conductivity from the prior estimate. The matrix $\mathbf{Q}$ serves to weigh the second term against the first: the "larger" the $\mathbf{Q}$, the less important the second term. The new optimization problem should, if properly constructed, have a unique solution. This formulation is appealing and has been used again and again in the solution of inverse problems.

In practice, the selection of the prior-information terms $\boldsymbol{\mu}$ and $\mathbf{Q}$ is not a trivial matter. Putting too much faith in a badly chosen $\boldsymbol{\mu}$ is bound to affect detrimentally the results. On the other hand, putting too little trust in $\boldsymbol{\mu}$ means failing to take advantage of information and may result in nonunique solutions to the minimization problem. Another issue is that in many cases there is no rational way to arrive at a $\boldsymbol{\mu}$ using prior information, but there may be information (prior or extracted from the observations that are used in the inverse problem) about the spatial continuity of log conductivity. Equation (7) requires some modification to be able to incorporate information of this type. The geostatistical approach [*Kitanidis and Vomvoris*, 1983] was developed to address such issues. An important premise of this approach, which will be described later, is that the additional information that is used in the solution of the problem is about the structure (such as spatial continuity) of the log conductivity and depends on prior information and also on the data. Similar considerations have motivated the development of the "Bayesian maximum likelihood" approach [e.g., *Carrera and Neuman*, 1986; *Loaiciga and Marino*, 1987] which however differs from the geostatistical approach (as discussed by *Kitanidis* [1996]).

From the discussion above it is clear that in both the overdetermined and the underdetermined formulations, the solution depends on the measurements and also on the parameterization. In the former case the solution depends on how the formation is subdivided into zones; in the latter formulation, the solution depends on prior information (the prior best estimate and its covariance) or on the assumed structure of the function.

The objective of this work is to develop a methodology that introduces the minimum of additional information. The potential advantage of such an approach is that it would allow the observations to "speak for themselves". Two different derivations will be presented: one based on minimization of the error norm and one based on a stochastic conceptual model.

## 2. Minimum Norm Formulation

In this section, we make no assumption about the unknown vector $\mathbf{s}$ other than that it has a finite norm, as will be discussed later. To simplify the analysis, we will start with the special case that the observation equation is linear,

$$\mathbf{h}(\mathbf{s}) = \mathbf{H}\mathbf{s} \quad (8a)$$

that is,

$$\mathbf{y} = \mathbf{H}\mathbf{s} \quad (8b)$$

where $\mathbf{s}$ is the unknown, $\mathbf{y}$ is the observations, $\mathbf{H}$ is an $n$ by $m$ known matrix, and $m > n$. Consider also that $n$ rows of $\mathbf{H}$ are

linearly independent; otherwise, some of the observations are redundant or inconsistent, and this should be rectified before proceeding.

A linear inversion scheme is a reasonable choice, that is,

$$\hat{\mathbf{s}} = \mathbf{\Lambda y} \tag{9}$$

where $\hat{\mathbf{s}}$ is the estimate of $\mathbf{s}$. The error is

$$\mathbf{e} = \mathbf{s} - \mathbf{\Lambda H s} = (\mathbf{I} - \mathbf{\Lambda H})\mathbf{s} \tag{10}$$

We may impose a number of conditions on the error vector through our selection of matrix $\mathbf{\Lambda}$. First, assume we change by $\zeta$ the datum used to quantify $\mathbf{s}$. This is equivalent to shifting the unknown vector by $\zeta\mathbf{u}$, where $\mathbf{u}$ is an $m$ by 1 vector of 1s. Obviously, the error should be unaffected by such a change:

$$(\mathbf{I} - \mathbf{\Lambda H})\mathbf{u}\zeta = 0 \tag{11}$$

Since $\zeta$ is arbitrary, the only way to guarantee that this condition will be met is by enforcing the constraint:

$$\mathbf{\Lambda H u} = \mathbf{u} \tag{12}$$

Our second concern is to make the length of the error vector, indicated by $\|\mathbf{e}\|$, as small as possible. The Euclidean (also known as Frobenius or Schur) norm of matrix $\mathbf{A}$ is by definition

$$\|\mathbf{A}\| = \{\mathrm{Tr}[\mathbf{A}^T\mathbf{A}]\}^{1/2} = \left\{ \sum_i \sum_j A_{ij}^2 \right\}^{1/2} \tag{13}$$

where "Tr" indicates matrix trace. A basic property of a norm is that $\|\mathbf{AB}\| \leq \|\mathbf{AB}\|\,\|\mathbf{B}\|$. Thus

$$\|\mathbf{e}\| = \|(\mathbf{I} - \mathbf{\Lambda H})\mathbf{s}\| \leq \|\mathbf{I} - \mathbf{\Lambda H}\|\,\|\mathbf{s}\| \tag{14}$$

We can minimize the upper bound on the error norm by selecting the $\mathbf{\Lambda}$ that minimizes $\|\mathbf{I} - \mathbf{\Lambda H}\|$.

The two requirements have thus led to the constrained minimization problem:

$$\min \mathrm{Tr}\,[(\mathbf{I} - \mathbf{\Lambda H})^T(\mathbf{I} - \mathbf{\Lambda H})] \tag{15}$$

subject to constraint (12).

The solution may be obtained readily through the method of Lagrange multipliers (see Appendix A). Finally, the estimate is

$$\hat{\mathbf{s}} = \mathbf{H}^{\mathbf{T}}\boldsymbol{\xi} + \mathbf{u}b \tag{16}$$

where the $n$ by 1 vector $\boldsymbol{\xi}$ and the scalar $b$ satisfy the system of $n + 1$ linear equations:

$$\begin{bmatrix} \mathbf{HH}^T & \mathbf{Hu} \\ (\mathbf{Hu})^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{17}$$

One may give the following intuitive interpretation to the system of equations, given the estimator of (16). The first equation in (17) is obtained by considering that the estimate must reproduce the given data,

$$\mathbf{y} = \mathbf{H}\hat{\mathbf{s}} = \mathbf{HH}^{\mathbf{T}}\boldsymbol{\xi} + \mathbf{Hu}b \tag{18}$$

To obtain the second equation, interpret $b$ as the arithmetic mean of the estimate $\hat{\mathbf{s}}$ (remember that $\mathbf{u}$ is an $m \times 1$ vector of 1s), that is,

$$b = \frac{1}{\mathbf{u}^{\mathbf{T}}\mathbf{u}}\,\mathbf{u}^{\mathrm{T}}\hat{\mathbf{s}} \tag{19}$$

But from (16),

$$\mathbf{u}^T\hat{\mathbf{s}} = \mathbf{u}^T\mathbf{H}^{\mathbf{T}}\boldsymbol{\xi} + \mathbf{u}^{\mathbf{T}}\mathbf{u}b \tag{20}$$

Since the first term equals the last term, the middle term must vanish.

We briefly mention the special case in which the constraint (12) is redundant: when the $\mathbf{u}^T$ is linearly dependent on the rows of $\mathbf{H}$. Then the minimization of the norm, (15), yields the result:

$$\hat{\mathbf{s}} = \mathbf{H}^{\mathbf{T}}\boldsymbol{\xi} \qquad (\mathbf{HH}^T)\boldsymbol{\xi} = \mathbf{y} \tag{21}$$

or, explicitly,

$$\hat{\mathbf{s}} = \mathbf{H}^{\mathbf{T}}(\mathbf{HH}^T)^{-1}\mathbf{y} \tag{22}$$

which is identical with the renowned method of *Backus and Gilbert* [1967].

This is the basic difference between the method presented here and that of Backus and Gilbert: We enforce (unless it is redundant) the constraint (12). Because the method of Backus and Gilbert does not enforce that requirement, its estimate is affected if one changes the datum for assigning values to $\mathbf{s}$ because the estimates are always biased toward the datum value! The Backus and Gilbert solution was apparently intended for cases where the data dominate the estimate sufficiently to prevent the bias from becoming conspicuous or where it does make sense to push the estimate toward zero. This method may not perform satisfactorily under different circumstances. The Backus and Gilbert solution is the one with the smallest norm while the solution presented here has the smallest norm of fluctuations from its mean; thus the former is called smallest solution while the latter may be called most uniform solution. These designations are accurate, but it should be kept in mind that their raison d'être is that they minimize bounds on the estimation error.

The approach presented in this section can be extended to the case of nonlinear measurement equation, through successive linearizations. We will deal with this issue within the stochastic formulation where the meaning of the nonlinear estimation will be more transparent and intuitive as a weighted least squares approach.

## 3. Stochastic Formulation and Least Squares

### 3.1. Model

In the stochastic approach, information about the unknown log conductivity function $s(x)$ is represented through statistical moments. The a priori moments of $s(x)$ are deterministic functions that may involve a few unknown parameters. In the prevailing approach, only the first two moments, the mean function and the covariance function, are utilized. A convenient representation of the mean function is

$$\mu(x) = c(x) + \sum_{i=1}^{p} X_i(x)\beta_i \tag{23}$$

where $c(x)$ and $X_i(x)$, $i = 1, \cdots, p$ are known functions and $\beta_i$ are unknown parameters. For example, *Kitanidis and Vomvoris* [1983] used a constant but unknown mean: $c(x) = 0$, $p = 1$, and $X_1(x) = 1$. The covariance function is usually a function of the separation vector $q(x - x'; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ represents a few parameters that need to be adjusted from data.

It is natural and convenient to arrange the $n$ observations in a vector, $\mathbf{y}$. Similarly, it is convenient to consider that the unknown function has been discretized into an $m$ by 1 vector, $\mathbf{s}$. We are interested in situations where $m \gg n$. The function discretization is mathematically unessential but expedient because it allows us to derive solutions, even analytic ones, without recourse to Hilbert or Banach spaces. A priori, the vector $\mathbf{s}$ is random with mean

$$\boldsymbol{\mu}_s = E[\mathbf{s}] = \mathbf{c}_s + \mathbf{X}_s\boldsymbol{\beta} \tag{24}$$

(where $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters, $\mathbf{c}_s$ is a known vector of dimension $m$; $\mathbf{X}_s$ is $m$ by $p$ matrix) and covariance matrix $\mathbf{Q}_{ss}$.

### 3.2. Approach

We now give a synopsis of the stochastic approach in a Bayesian framework [*Kitanidis*, 1986] abridged for the case that the covariance is known. Bayes theorem states that the a posteriori probability distribution of $\mathbf{s}$ and $\boldsymbol{\beta}$ is proportional to the likelihood of the data given $(\mathbf{s}, \boldsymbol{\beta})$ times the a priori distributions of $\mathbf{s}$ and $\boldsymbol{\beta}$:

$$p''(\mathbf{s}, \boldsymbol{\beta}) \propto p(\mathbf{y}|\mathbf{s}, \boldsymbol{\beta})p'(\mathbf{s}, \boldsymbol{\beta}) \tag{25}$$

Since a priori $\mathbf{s}$ and $\boldsymbol{\beta}$ are independent and the prior of $\boldsymbol{\beta}$ is noninformative, that is, $p'(\boldsymbol{\beta}) \propto 1$,

$$p'(\mathbf{s}, \boldsymbol{\beta}) \propto p'(\mathbf{s})$$
$$\propto \exp[-\tfrac{1}{2}(\mathbf{s} - \mathbf{c}_s - \mathbf{X}_s\boldsymbol{\beta})^T\mathbf{Q}_{ss}^{-1}(\mathbf{s} - \mathbf{c}_s - \mathbf{X}_s\boldsymbol{\beta})] \tag{26}$$

under the assumption that $p'$ is Gaussian. The measurement-error vector $\mathbf{v}$ is modelled as random, with Gaussian probability distribution with zero mean and covariance matrix $\mathbf{R}$. The likelihood is

$$p(\mathbf{y}|\mathbf{s}, \boldsymbol{\beta}) \propto \exp[-\tfrac{1}{2}(\mathbf{y} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{h}(\mathbf{s}))] \tag{27}$$

The posterior distribution has thus been derived, although its brute force computation for an arbitrary measurement equation is generally impractical. In applications we are interested in obtaining best estimates and error estimates or bounds. The straightforward way is to compute as best estimates the values of $\mathbf{s}$ and $\boldsymbol{\beta}$ that maximize the posterior pdf, a procedure known as maximum a posteriori (MAP) estimation. This is equivalent to minimizing the negative logarithm of $p''(\mathbf{s}, \boldsymbol{\beta})$ with respect to $\mathbf{s}$ and $\boldsymbol{\beta}$,

$$C_1 = (\mathbf{y} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{h}(\mathbf{s}))$$
$$+ (\mathbf{s} - \mathbf{c}_s - \mathbf{X}_s\boldsymbol{\beta})^T\mathbf{Q}_{ss}^{-1}(\mathbf{s} - \mathbf{c}_s - \mathbf{X}_s\boldsymbol{\beta}) \tag{28}$$

This is a weighted least squares criterion, just like the one in (7). The first term represents the requirement to reproduce the observations, and the second term represents the requirement for the function estimate to be close to the prior mean. The second term thus serves to "regularize" the problem, that is, to produce an optimization problem that has a unique and well-defined solution.

Although the criterion of expression (28) is simple and intuitive, keep in mind that the objective is the estimation of just $\mathbf{s}$. Thus the relevant distribution is the marginal posterior pdf of $\mathbf{s}$:

$$p''(\mathbf{s}) = \int p''(\mathbf{s}, \boldsymbol{\beta})\, d\boldsymbol{\beta} \tag{29}$$

This integration is easy to perform:

$$p''(\mathbf{s}) \propto \exp[-\tfrac{1}{2}(\mathbf{y} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{h}(\mathbf{s}))]$$
$$\cdot \exp[-\tfrac{1}{2}(\mathbf{s} - \mathbf{c}_s)^T\mathbf{G}(\mathbf{s} - \mathbf{c}_s)] \tag{30}$$

where

$$\mathbf{G} = \mathbf{Q}_{ss}^{-1} - \mathbf{Q}_{ss}^{-1}\mathbf{X}_s(\mathbf{X}_s^T\mathbf{Q}_{ss}^{-1}\mathbf{X}_s)^{-1}\mathbf{X}_s^T\mathbf{Q}_{ss}^{-1} \tag{31}$$

Then the MAP estimate is obtained by minimizing with respect to $\mathbf{s}$ the criterion:

$$C_2 = (\mathbf{y} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{h}(\mathbf{s})) + (\mathbf{s} - \mathbf{c}_s)^T\mathbf{G}(\mathbf{s} - \mathbf{c}_s) \tag{32}$$

It is noted parenthetically that the same criterion could be obtained if the expression of (28) were minimized with respect to $\boldsymbol{\beta}$ given $\mathbf{s}$ and then this expression were substituted into (28). That is why it is also accurate to state that (32) is obtained from (28) by eliminating $\boldsymbol{\beta}$, whereas in the Bayesian interpretation one should say that (32) is obtained from (28) by averaging out $\boldsymbol{\beta}$.

We will now focus on criterion $C_2$, equation (32), which is more general than expression (7): expression (7) is obtained from expression (32) by eliminating terms that involve $\mathbf{X}_s$. The MAP estimate of $\mathbf{s}$ is obtained by minimizing $C_2$. There is another important advantage in using criterion (32) as starting point. Consider for example that a solution that meets certain criteria of flatness or smoothness is sought. One may then set $\mathbf{c}_s = 0$ and adopt an appropriate $\mathbf{G}$ (using the concept of generalized covariance functions) that will have the desired effect.

### 3.3. Minimum Structure Model

In contrast with the geostatistical approach that strives to identify the most appropriate structure and thus incorporate useful information into the estimation, here we will use the least informative or minimum-structure model: that is, the model that supposes the least about the unknown function. What is this model? A rigorous treatment of this subject would require a serious incursion into information theory, although through common sense and intuition the answer may be already obvious to many readers. We introduce the following postulate:

For a random function with mean square value, the minimum-structure model is one where $s(x)$ is Gaussian-distributed with mean

$$\mu(x) = \beta \tag{33}$$

where $\beta$ is unknown and covariance function

$$q(x - x') = \begin{cases} \theta & x = x' \\ 0 & x \neq x' \end{cases} \tag{34}$$

where $\theta \to \infty$. That is, the mean is constant but unknown, and the covariance is, in the geostatistical terminology, nugget effect with arbitrarily large variance.

For this model

$$\mathbf{Q}_{ss} = \theta\mathbf{I} \tag{35}$$

where $\mathbf{I}$ is the $m$ by $m$ identity matrix. The $\mathbf{X}$ matrix is written for convenience:

$$\mathbf{X}_s = \vartheta\mathbf{u} \tag{36}$$

where $\vartheta$ is an arbitrary scalar (which may be chosen purely for convenience in numerical computations) and $\mathbf{u}$ is an $m$ by 1 vector consisting of 1s.

$$\mathbf{G} = \frac{1}{\theta}[\mathbf{I} - \mathbf{u}(\mathbf{u}^T\mathbf{u})^{-1}\mathbf{u}^T] = \frac{1}{\theta}\mathbf{G}_0 \tag{37}$$

where $\mathbf{G}_0$ is a symmetric matrix with $m - 1$ eigenvalues equal to unit and 1 eigenvalue equal to zero. Thus our objective is to minimize:

$$C = (\mathbf{y} - \mathbf{h}(\mathbf{s}))^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{h}(\mathbf{s})) + \frac{1}{\theta}\mathbf{s}\mathbf{G}_0\mathbf{s} \tag{38}$$

A cursory look at (38) might suggest that for infinite $\theta$ there is no unique solution, because the second term that regularizes the solution vanishes, as expected since this is the no-prior-information case. However, in the next section a more careful examination will demonstrate that a unique and well-defined solution does exist. This is achieved by developing the solution in an appropriate form for finite $\theta$ and then considering the limit when $\theta \to \infty$.

## 4. Obtaining the Estimate

The MAP estimate $\hat{\mathbf{s}}$ must satisfy the equation obtained by setting the derivative of $C$ with respect to $\mathbf{s}$ equal to zero,

$$-(\mathbf{y} - \mathbf{h}(\hat{\mathbf{s}}))^T\mathbf{R}^{-1}\mathbf{H} + \frac{1}{\theta}\hat{\mathbf{s}}^T\mathbf{G}_0 = 0 \tag{39}$$

where $\mathbf{H}$ is $n$ by $m$ matrix with $H_{ij} = \partial h_i/\partial s_j$. Also, for convenience, we will set $\vartheta$ equal to $\theta$.

### 4.1. Linear Case

In the special case that the observation equation is linear, equation (8), the solution might be found from the linear vector equation:

$$\left[\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \frac{1}{\theta}\mathbf{G}_0\right]\hat{\mathbf{s}} = \mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} \tag{40}$$

However, this direct approach led to a system of linear equations that is unsatisfactory because (1) it involves solving a system of order of $m$, where $m$ can be an arbitrarily large number, and (2) as $\theta \to \infty$ the matrix of coefficients tends to become singular because its rank tends to become $n$, which is less than its order, $m$. Thus the solution to this system tends to become unstable and nonunique.

We can rework the result using matrix identities in order to obtain what is widely known as kriging-cokriging equations. As shown in Appendix B,

$$\hat{\mathbf{s}} = \mathbf{u}b + \mathbf{H}^T\boldsymbol{\xi} \tag{41}$$

where the $n$ by 1 matrix $\boldsymbol{\xi}$ and the scalar $b$ are found from the solution of a system of $n + 1$ equations with $n + 1$ unknowns.

$$\begin{bmatrix} \mathbf{H}\mathbf{H}^T + \frac{1}{\theta}\mathbf{R} & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\xi} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{42}$$

At the limit, the $(1/\theta)\mathbf{R}$ term vanishes.

### 4.2. Nonlinear Case

The observation equation $\mathbf{h}(\mathbf{s})$, equation (38), is typically nonlinear and must be solved using an iterative method. Let $\tilde{\mathbf{s}}$ be the most recent guess, which may be the starting guess or the result of a previous iteration. The essence of the celebrated Gauss-Newton method is to linearize the observation function about the most recent guess,

$$-(\mathbf{y} - \mathbf{h}(\tilde{\mathbf{s}}) - \mathbf{H}(\hat{\mathbf{s}} - \tilde{\mathbf{s}}))^T\mathbf{R}^{-1}\mathbf{H} + \frac{1}{\theta}\hat{\mathbf{s}}^T\mathbf{G}_0 = 0 \tag{43}$$

or,

$$-((\mathbf{y} - \mathbf{h}(\tilde{\mathbf{s}}) + \mathbf{H}\tilde{\mathbf{s}}) - \mathbf{H}\hat{\mathbf{s}})^T\mathbf{R}^{-1}\mathbf{H} + \frac{1}{\theta}\hat{\mathbf{s}}^T\mathbf{G}_0 = 0 \tag{44}$$

Thus the problem has been reduced to the linear case. Thus the guess is updated through equations of extended kriging-cokriging:

$$\hat{\mathbf{s}} = \mathbf{u}b + \mathbf{H}^T\boldsymbol{\xi} \tag{45}$$

where the $\boldsymbol{\xi}$ and $\mathbf{b}$ coefficients are found by solving a single linear system of $n + 1$ equations:

$$\begin{bmatrix} \mathbf{H}\mathbf{H}^T + \frac{1}{\theta}\mathbf{R} & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix}\begin{bmatrix} \boldsymbol{\xi} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} - \mathbf{h}(\tilde{\mathbf{s}}) + \mathbf{H}\tilde{\mathbf{s}} \\ 0 \end{bmatrix} \tag{46}$$

Next, select $\hat{\mathbf{s}}$ as the new $\tilde{\mathbf{s}}$ and repeat the procedure until convergence is achieved. The important practical issue of whether this approach converges will be discussed later.

It is repeated for emphasis that in this manner the weighted least squares criterion $C$, equation (38), is minimized with respect to the $m$ dimensional vector $\mathbf{s}$, although a much smaller system of $n + 1$ equations is solved thus achieving great computational savings.

## 5. Noisy Data

The solution developed in the previous sections aims to reproduce the data exactly and effectively treats the measurement error term as zero. However, in applications, the measurement error term is anything but negligible because it represents (1) errors in the collection and transmission of measurements; (2) conceptual errors or limitations in the model, such as neglecting the effects of pumping wells, recharge, or leakage; (3) scale discrepancies, such as when the measured quantity is defined over a small volume in contrast to the predicted quantity that is defined over a large volume; and (4) computational errors in the mathematical model, which are always present and can be significant in some cases.

In extreme cases there may be no solution that reproduces the data. In all cases errors in the data affect the predictions, especially if there are many observations. In fact, it is in the nature of the problem of estimating conductivity from head that as the density of head observations increases, the estimates become more sensitive to perturbations in the observations.

To make the methodology more useful, the estimates must become robust to perturbations in the data. This can be achieved easily by selecting a finite variance parameter $\theta$. The solution then loses its probity as the most likely solution given only the data and the model; a finite $\theta$ means infusing information into the solution which may countervail information in

the data. However, the approach supposes no spatial continuity so that in a less absolute sense this is still a "minimum structure" solution. The benefit of suppressing the effects of errors in the data outweighs any disadvantages associated with the infusion of limited outside information.

The value of $\theta$ may be selected by the modeler through a trial-and-error procedure or if sufficient observations are available through a systematic cross-validation method. Here we will describe how to apply the general parameter estimation (i.e., restricted maximum likelihood or cross validation) method described by *Kitanidis* [1995] to estimate $\theta$ from the data.

The method is iterative. We start with an initial estimate of $\theta$ and we proceed to find a best estimate $\hat{\mathbf{s}}$. About this value, linearize the $\mathbf{h}$ function with respect to $\mathbf{s}$, that is,

$$h(\mathbf{s}) = h(\hat{\mathbf{s}}) + \mathbf{H}(\mathbf{s} - \hat{\mathbf{s}}) \tag{47}$$

where

$$\mathbf{H} = \left.\frac{\partial \mathbf{h}}{\partial \mathbf{s}}\right|_{s=\hat{\mathbf{s}}} \tag{48}$$

Compute

$$\mathbf{\Sigma} = \mathbf{H}\mathbf{H}^T + \frac{1}{\theta}\mathbf{R} \tag{49}$$

$$\mathbf{y}_0 = \mathbf{y} - \mathbf{h}(\hat{\mathbf{s}}) + \mathbf{H}\hat{\mathbf{s}} \tag{50}$$

$\mathbf{y}_0$ and $\mathbf{H}$ are treated as constant in the estimation of $\theta$.

Then perform matrix inversion

$$\begin{bmatrix} \mathbf{\Sigma} & \mathbf{Hu} \\ (\mathbf{Hu})^T & 0 \end{bmatrix}^{-1} = \begin{bmatrix} \mathbf{\Xi} & \mathbf{\Psi} \\ \mathbf{\Psi}^T & \Omega \end{bmatrix}$$

where $\mathbf{\Xi}$ is an $n$ by $n$ matrix.

We seek $\theta$ that satisfies the scalar equation

$$g = \mathrm{Tr}\,[\mathbf{\Xi}\mathbf{H}\mathbf{H}^T] - \frac{1}{\theta}\mathbf{y}_0^T(\mathbf{\Xi}\mathbf{H}\mathbf{H}^T\mathbf{\Xi})\mathbf{y}_0 = 0 \tag{51}$$

Because the equation is nonlinear, it is solved iteratively. A scheme that usually works is the Gauss-Newton method (for restricted maximum likelihood, not to be confused with the Gauss-Newton method for determination of $\mathbf{s}$.)

$$\theta_{l+1} = \theta_l - \left[\frac{\mathrm{Tr}[\mathbf{\Xi}\mathbf{H}\mathbf{H}^T] - \frac{1}{\theta}\mathbf{y}_0^T(\mathbf{\Xi}\mathbf{H}\mathbf{H}^T\mathbf{\Xi})\mathbf{y}_0}{\frac{1}{\theta}\mathrm{Tr}[\mathbf{\Xi}\mathbf{H}\mathbf{H}^{\mathbf{T}}\mathbf{\Xi}\mathbf{H}\mathbf{H}^T]}\right]_l \tag{52}$$

When the iterative procedure has converged, a new value of $\theta$ has been obtained and the procedure of estimating $\mathbf{s}$ needs to be repeated.

## 6. Implementation and Examples

In weakly nonlinear cases, characterized by mild and gradual variability, the iterative Gauss-Newton method was found to converge relatively rapidly and to the global minimum. The estimate does not include unwarranted bumps and is invariant to the initial guess. However, for large and abrupt changes in log conductivity the basic Gauss-Newton method was found to be inadequate: It would often fail to converge, or it would converge to a local minimum. This experience is different from that of *Kitanidis* [1995, 1996], for whom the Gauss-Newton method performed quite well. It is clear that in methods that utilize information about the structure (particularly by requiring that the solution varies continuously or smoothly) and where the large measurement error is taken into account, the objective function satisfies the conditions required for Gauss-Newton to perform well. However the minimum supposition solution relies only on the data and thus the nonlinearity in the measurement equation may become crucial.

Finding the solution in truly nonlinear least squares is a notoriously difficult problem. Although the nonlinear least squares problem is one of the most studied problems, there is unfortunately no method that is infallible or guaranteed to work in all cases. The following quote from *Draper and Smith* [1981, p. 471] sums up the situation: "... given a particular method a problem can be constructed to defeat it. Alternatively, given a particular problem and a suggested method, ad hoc modifications can often provide quicker convergence ...."

The performance of the Gauss-Newton method generally improves by including line searches. That is, in each iteration, we obtain from an initial estimate $\tilde{\mathbf{s}}_0$ an updated estimate $\tilde{\mathbf{s}}_1$. The line connecting the two points is

$$\mathbf{s} = \tilde{\mathbf{s}}_0 + \rho(\tilde{\mathbf{s}}_1 - \tilde{\mathbf{s}}_0) \tag{53}$$

For a number of $\rho$ values (including $\rho = 0$ and 1), solutions are obtained and evaluated by computing the objective function. The solution that minimizes the objective function is selected as the most up-to-date estimate instead of always using the value that corresponds to $\rho = 1$. Disadvantages of the approach is that it increases the cost of computations and also is not a panacea in strongly nonlinear cases.

We will present representative examples and will discuss the methods that were found to work well.

### 6.1. A Weakly Nonlinear Case

Consider one-dimensional flow without sources or sinks between two boundaries. The flow satisfies

$$\frac{d}{dx}\left(K\frac{d\phi}{dx}\right) = 0 \tag{54}$$

where $K$ is conductivity, $\phi$ is hydraulic head, and $x$ is a spatial coordinate. The hydraulic head at the left boundary and the specific discharge $q$ are given (not the hydraulic head at the other boundary). The problem has been rendered into a dimensionless form by selecting the size of the domain as unit length for measuring $x$, selecting the hydraulic head at $h = 0$ as the unit hydraulic length, and the specific discharge $q$ as the unit for measuring conductivities. There is one log conductivity observation at $x = 0.5$ and six head observations at $x = 0.1$, 0.2, 0.3, 0.4, 0.7, and 0.9. For the numerical computation and plotting, the domain was subdivided into 100 equal segments with uniform conductivity.

First, consider the solution obtained through engineering judgment. Since the discharge is given, application of Darcy's law in the segment between two consecutive head observations yields an estimate to the effective conductivity in that segment. This is the best estimate one could obtain without using additional information about the structure of the function.

The minimum-structure inverse methodology was then applied, starting with error-free data. To account for the finite numerical of the computer, the variance of the measurement error was taken equal to $10^{-14}$, that is,

$$\mathbf{R} = 10^{-14}\mathbf{I} \tag{55}$$

where **I** is the $n$-dimensional identity matrix. The estimate obtained with the method that depends solely on the data (for $1/\theta \to 0$), which is shown in Figure 1, is exactly equal to our engineering guess. Note that in the segment [0.9, 1] Darcy's law cannot be applied because the head is measured only at the one end and the methodology yields as best estimate the mean of the other segments. The same final result was obtained for many different starting estimates (needed because of the iterative nature of the methodology), but convergence seemed the fastest when starting with a constant log conductivity close to the measured log conductivity. Also, using any value $\theta > 10^{-6}$ gave practically the same final estimate of log conductivity. By the way, a $\theta$ larger than, say, 100 is an extremely large value for a log conductivity variance.

Next, the observations were modified by introducing observation error, pseudorandom Gaussian with zero mean and variance 0.02 and consistent with that,

$$\mathbf{R} = 4 \times 10^{-4}\mathbf{I} \tag{56}$$

For the case of infinite $\theta$ the methodology strives to reproduce the data exactly and yields the estimate shown in Figure 2. In this case the estimates are less accurate particularly where head observations are the closest. (The reasons are obvious if one considers that the estimate is $K = q\Delta L/|\Delta\phi|$ in every segment between head observations, obtained from Darcy's law.) Next, when we use the method for noisy data, we obtain a finite $\theta = 0.7$, and the estimates are shown in Figure 3. The estimated log conductivity has become considerably more uniform because of the lack of information in the data, but the effect of the observations errors has been suppressed effectively. The estimate when introduced into the flow equation reproduces the data only approximately, as appropriate given the observation error.

### 6.2. A Strongly Nonlinear Case

We will consider a log conductivity function that has 10 times the spread of the function seen in the previous one. In this case the Gauss-Newton method would not converge to the best estimate for an arbitrary initial condition, as is known to often be the case in strongly nonlinear least squares. Nevertheless, a method was devised which worked in this as well as in other cases examined so far:

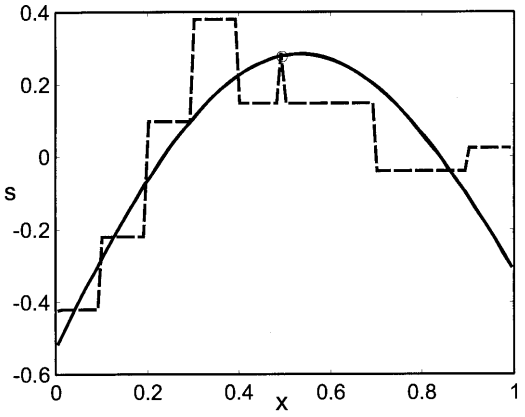


**Figure 2.** Actual (solid line) and estimate (dashed line) when noise in the data is not taken into account.

1. Start with a constant initial estimate of **s** as close to the actual mean as the data allow.
2. Utilize a $\theta$ that is small enough to balance the two terms in (38). The idea is to make the objective function nearly quadratic in the neighborhood that contains the estimate of **s**. The initial value of $\theta$ may need to be extremely small.
3. Apply Gauss-Newton. If the new estimates of **s** are uneven anywhere except at observation points, the Gauss-Newton method has gone astray and one needs to return to step 2 and reduce the value of $\theta$. Otherwise allow the Gauss-Newton to converge to a new estimate **s**.
4. Increase $\theta$ and return to step 3 for a Gauss-Newton search with a new initial estimate **s**.
5. Continue until the results do not change any more.

A slightly different but more systematic approach to accomplish the same task is the following:

1. Start with a constant initial estimate of **s** as close to the actual mean as possible.
2. Apply the method for noisy data starting with a sufficiently small value of $\theta$.
3. After the method converges to a final $\theta$ and **s** estimates, increase $\theta$ as much as desired.

The second method actually performed impressively in the example problem. For no noise in the data and **R** as in (55), an initial $\theta = 10^{-12}$ was used. The iterative method updated the estimate of $\theta$ to the value 20.4 at the first iteration and to 71.6

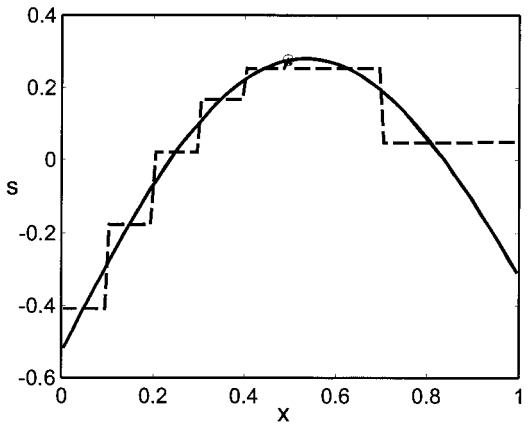


**Figure 1.** Actual (solid line) and estimate (dashed line) for weakly nonlinear noise-free case.

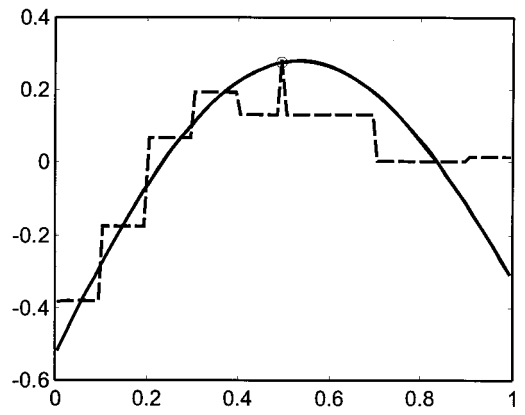


**Figure 3.** Actual (solid line) and estimate (dashed line) when accounting for error in the observations.

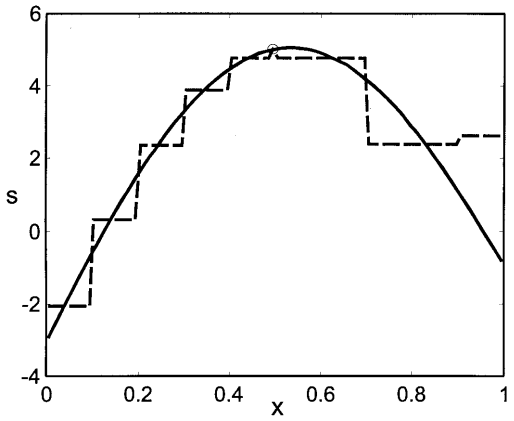


**Figure 4.** Actual (solid line) and estimated (dashed line) log conductivity for strongly nonlinear case with error-free data.

at the second iteration. At that point the method converged and the best estimate is given in Figure 4. Increasing the value of $\theta$ to any value larger than that gave estimates of **s** that are indistinguishable from those shown in Figure 4.

Next error was added to the data, pseudorandom Gaussian with zero mean and variance 0.02 and the methodology was applied with

$$\mathbf{R} = 4 \times 10^{-4}\mathbf{I} \tag{57}$$

The estimate for infinite $\theta$ is depicted in Figure 5. When the methodology for noisy data was used, an estimate of $\theta$ equal to 53 was found and the corresponding estimate is given in Figure 6. By putting less weight on the data, the estimate becomes much more uniform.

## 7. Discussion

The presented method has advantages and disadvantages:

1. The minimum-structure solution is of theoretical as well as practical interest. From a theoretical standpoint the method produces the most likely solution given the data. From a practical standpoint it is most useful as a benchmark or extreme case, when one wants to get an estimate that relies only on the data. However, in practice, better estimates may be obtained by introducing some information about spatial structure, as done for example in the geostatistical approach [e.g., *Kitanidis and Vomvoris*, 1983; *Dagan*, 1985; *Hoeksema and Kitanidis*, 1984, 1985, 1989; *Rubin and Dagan*, 1987a, b; *Wagner and Gorelick*, 1989; *Hoeksema and Clapp*; 1990, *Kitanidis*, 1995; *Yeh et al.*, 1995].

2. The minimum-structure solution tries to reproduce the data exactly. Although this is a definite advantage when the observations are free of errors, it produces artifacts when there are observation errors. In practice, observation errors are quite large and are expected to have a detrimental effect on the solution; the method would then do definitely better if it reproduced the data less faithfully. This applies to all inverse methods. Although inverse methods are often promoted on the basis that they reproduce the data, exact reproduction of imprecise data is a dubious accomplishment.

For the above reasons a modification of the basic method was introduced that accounts for the need to "filter out" the effects of observation errors: A finite variance $\theta$ is used, estimated using a cross-validation technique. In this case the method is not the truly minimum structure solution, but it still supposes no spatial continuity. This method seems to work well.

1. The methodology seeks the most uniform function that is consistent with the data but the solution does not have to be either flat or smooth. In particular, we must make a distinction between a uniform function, that (limiting attention to the one-dimensional case) has a small $\int (s(x) - \bar{s})^2\, dx$, where $\bar{s}$ is the mean of $s(x)$, and a flat solution that has a small $\int (ds/dx)^2\, dx$. (The integral is over the whole domain.) An advantage of the no-structure solution is that it adjusts nicely to sharp changes, if such changes are indicated by the data, whereas other methods that assume spatial continuity tend to smoothen the transition. However, the same characteristic may be a disadvantage in that random errors may cause sharp ups and downs.

2. A distinction is to be made between cases of weakly nonlinear least squares where the Gauss-Newton method ("extended cokriging") is robust and effective and cases of strongly nonlinear least squares which involve challenging nonlinear optimization problems. Of course, this is a concern with all inverse methods that attempt to solve strongly nonlinear least squares problems, such as when the variance is large. A method was presented to solve such problems but no procedure is expected to be applicable to all cases.

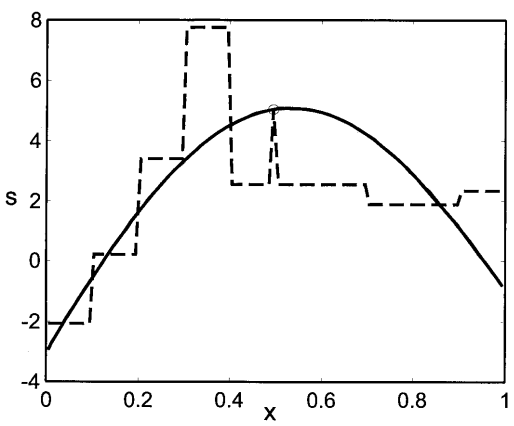


**Figure 5.** Actual (solid line) and estimated (dashed line) log conductivity for strongly nonlinear case with errors in the data (which are neglected by method).

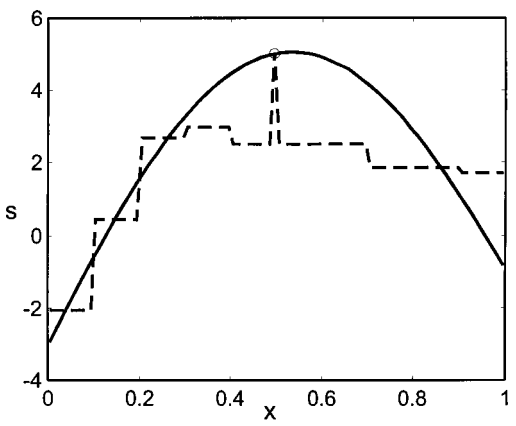


**Figure 6.** Actual (solid line) and estimated (dashed line) log conductivity for strongly nonlinear case with errors data, using methodology for noisy data.

## Appendix A: Lagrange Multipliers

The solution may be obtained readily through the method of Lagrange multipliers. Introduce the Lagrangian,

$$\mathrm{Tr}\,[(\mathbf{I} - \mathbf{\Lambda H})^T(\mathbf{I} - \mathbf{\Lambda H})] + 2\ \mathrm{Tr}\,[\boldsymbol{\nu}^T(\mathbf{\Lambda H u} - \mathbf{u})] \tag{58}$$

where $\boldsymbol{\nu}$ is an $m$ by 1 vector of Lagrange multipliers

$$\begin{aligned}\mathrm{Tr}[(\mathbf{I} - \mathbf{\Lambda H})^T(\mathbf{I} - \mathbf{\Lambda H})] &+ 2\ \mathrm{Tr}[\boldsymbol{\nu}^T(\mathbf{\Lambda H u} - \mathbf{u})] \\ &= m - \mathrm{Tr}[\mathbf{H}^T\mathbf{\Lambda}^T + \mathbf{\Lambda H}] + \mathrm{Tr}[\mathbf{H}^T\mathbf{\Lambda}^T\mathbf{\Lambda H}] \\ &\quad + 2\ \mathrm{Tr}[(\mathbf{u}^T\mathbf{H}^T\mathbf{\Lambda}^T - \mathbf{u}^T)\boldsymbol{\nu}]\end{aligned} \tag{59}$$

Take derivative [see *Schweppe*, 1973, p. 509] with respect to $\mathbf{\Lambda}$ and set equal to zero:

$$-2\mathbf{H} + 2\mathbf{H}\mathbf{H}^T\mathbf{\Lambda}^T + 2\mathbf{H}\mathbf{u}\boldsymbol{\nu}^T = 0 \tag{60}$$

Combine with constraint:

$$\begin{bmatrix} \mathbf{H}\mathbf{H}^T & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{\Lambda}^T \\ \boldsymbol{\nu}^T \end{bmatrix} = \begin{bmatrix} \mathbf{H} \\ \mathbf{u}^T \end{bmatrix} \tag{61}$$

Or

$$[\mathbf{\Lambda} \quad \boldsymbol{\nu}] = [\mathbf{H}^T \quad \mathbf{u}] \begin{bmatrix} \mathbf{H}\mathbf{H}^T & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix}^{-1} \tag{62}$$

Next multiply by vector $\begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix}$

$$[\mathbf{\Lambda} \quad \boldsymbol{\nu}] \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} = [\mathbf{H}^T \quad \mathbf{u}] \begin{bmatrix} \mathbf{H}\mathbf{H}^T & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{63}$$

Introduce the $n$ by 1 vector $\boldsymbol{\xi}$ and the scalar $b$ that satisfy the system of $n + 1$ equations with $n + 1$ unknowns:

$$\begin{bmatrix} \mathbf{H}\mathbf{H}^T & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{64}$$

Then

$$\hat{\mathbf{s}} = \mathbf{\Lambda y} = \mathbf{H}^{\mathbf{T}}\boldsymbol{\xi} + \mathbf{u}b \tag{65}$$

## Appendix B: Map Estimation

Equation (40) will be transformed into a more convenient form. Define

$$\mathbf{\Phi} = \mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \frac{1}{\theta}\mathbf{I} \tag{66}$$

$$\mathbf{\Sigma} = \mathbf{H}\mathbf{H}^T + \frac{1}{\theta}\mathbf{R} \tag{67}$$

From matrix identity [see *Schweppe*, 1973, p. 496],

$$\mathbf{\Phi}^{-1} = \theta\mathbf{I} - \theta\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H} \tag{68}$$

and also from

$$\begin{aligned}\mathbf{\Phi}^{-1}\mathbf{H}^T\mathbf{R}^{-1} &= (\theta\mathbf{I} - \theta\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H})\mathbf{H}^T\mathbf{R}^{-1} \\ &= \theta\mathbf{H}^T\mathbf{\Sigma}^{-1}\left(\mathbf{H}\mathbf{H}^T + \frac{1}{\theta}\mathbf{R}\right)\mathbf{R}^{-1} - \theta\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{H}^T\mathbf{R}^{-1} \\ &= \theta\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{H}^T\mathbf{R}^{-1} + \mathbf{H}^T\mathbf{\Sigma}^{-1} - \theta\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{H}^T\mathbf{R}^{-1} \\ &= \mathbf{H}^T\mathbf{\Sigma}^{-1}\end{aligned}$$

Then

$$\begin{aligned}\hat{\mathbf{s}} &= \left[\mathbf{H}^T\mathbf{R}^{-1}\mathbf{H} + \frac{1}{\theta}(\mathbf{I} - \mathbf{u}(\mathbf{u}^T\mathbf{u})^{-1}\mathbf{u}^T)\right]^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} \\ &= \left[\mathbf{\Phi} - \frac{1}{\theta}\mathbf{u}(\mathbf{u}^T\mathbf{u})^{-1}\mathbf{u}^T\right]^{-1}\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} \\ &= [\mathbf{\Phi}^{-1} + \mathbf{\Phi}^{-1}\mathbf{u}[\theta\mathbf{u}^T\mathbf{u} - \mathbf{u}^T\mathbf{\Phi}^{-1}\mathbf{u}]^{-1}\mathbf{u}^T\mathbf{\Phi}^{-1}]\mathbf{H}^T\mathbf{R}^{-1}\mathbf{y} \\ &= \mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{y} + \mathbf{\Phi}^{-1}\mathbf{u}[\mathbf{u}^T\mathbf{u} - \mathbf{u}^T\mathbf{\Phi}^{-1}\mathbf{u}]^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{y} \\ &= \mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{y} + (\mathbf{I} - \mathbf{H}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{H})\mathbf{u}[\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}]^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{y} \\ &= \mathbf{H}^T[\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}[\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}]^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}]\mathbf{y} \\ &\quad + \mathbf{u}[\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}]^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{y}\end{aligned}$$

We write

$$\hat{\mathbf{s}} = \mathbf{H}^T\boldsymbol{\xi} + \mathbf{u}b \tag{69}$$

where

$$\boldsymbol{\xi} = [\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}[\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}]^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}]\mathbf{y}$$

$$b = [\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}]^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{y}$$

Note that

$$\begin{bmatrix} \mathbf{\Sigma} & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} =$$

$$\begin{bmatrix} \mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}(\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u})^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1} & \mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u}(\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u})^{-1} \\ (\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u})^{-1}\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1} & -(\mathbf{u}^T\mathbf{H}^T\mathbf{\Sigma}^{-1}\mathbf{H}\mathbf{u})^{-1} \end{bmatrix}$$

$$\cdot \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\xi} \\ b \end{bmatrix} \tag{70}$$

Thus the $\boldsymbol{\xi}$ and $b$ coefficients are found by solving a single system of $n + 1$ linear equations:

$$\begin{bmatrix} \mathbf{H}\mathbf{H}^T + \frac{1}{\theta}\mathbf{R} & \mathbf{H}\mathbf{u} \\ (\mathbf{H}\mathbf{u})^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\xi} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{71}$$

**Acknowledgments.** Funding for this publication was provided by the office of Research and Development, U.S. Environmental Protection Agency, under agreement *R*-815738-01 through the Western Region Hazardous Substance Research Center. The content of this publication does not necessarily represent the views of the agency. Useful comments by the reviewers are appreciated.

## References

Backus, G. E., and J. F. Gilbert, Numerical application of a formalism for geophysical inverse problems, *Geophys. J. R. Astron. Soc.*, *13*, 247–276, 1967.

Carrera, J., and S. P. Neuman, Estimation of aquifer parameters under transient and steady state conditions, 1, Maximum likelihood method incorporating prior information, *Water Resour. Res.*, *22*(2), 199–210, 1986.

Cooley, R. L., A method for estimating parameters and assessing reliability for models of steady state groundwater flow, 1, Theory and numerical properties, *Water Resour. Res.*, *13*(2), 318–324, 1977.

Dagan, G., Stochastic modeling of groundwater flow by unconditional and conditional probabilities: The inverse problem, *Water Resour. Res.*, *21*(1), 65–73, 1985.

Draper, N., and H. Smith, *Applied Regression Analysis*, 709 pp., Wiley Intersci., New York, 1981.

Gavalas, G. R., P. C. Shah, and J. H. Seinfeld, Reservoir history matching by Bayesian estimation, *Soc. Pet. Eng. J.*, *16*, 337–350, 1976.

Ginn, T. R., and J. H. Cushman, Inverse methods for subsurface flow: A critical review of stochastic techniques, *Stochastic Hydrol. Hydraul.*, *4*, 1–26, 1990.

Hoeksema, R. J., and R. B. Clapp, Calibration of groundwater flow models using Monte Carlo simulations and geostatistics, in *Model-CARE 90: Calibration and Reliability in Groundwater Modelling IAHS Publ. 195*, pp. 33–42, 1990.

Hoeksema, R. J., and P. K. Kitanidis, An application of the geostatistical approach to the inverse problem in two-dimensional groundwater modeling, *Water Resour. Res.*, *20*(7), 1003–1020, 1984.

Hoeksema, R. J., and P. K. Kitanidis, Comparison of Gaussian conditional mean and kriging estimation in the geostatistical solution of the inverse problem, *Water Resour. Res.*, *21*(6), 825–836, 1985.

Hoeksema, R. J., and P. K. Kitanidis, Prediction of transmissivities, heads, and seepage velocities using mathematical models and geostatistics, *Adv. Water Resour.*, *12*(2), 90–102, 1989.

Jahns, H. O., A rapid method for obtaining a two-dimensional reservoir description from well pressure response data, *Soc. Pet. Eng. J.*, 315–327, 1966.

Kitanidis, P. K., Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resour. Res.*, *22*(4), 499–507, 1986.

Kitanidis, P. K., Quasilinear geostatistical theory for inversing, *Water Resour. Res.*, *31*(10), 2411–2419, 1995.

Kitanidis, P. K., On the geostatistical approach to the inverse problem, *Adv. Water Resour.*, *19*(6), 333–342, 1996.

Kitanidis, P. K., and E. G. Vomvoris, A geostatistical approach to the inverse problem in groundwater modeling (steady state) and one-dimensional simulations, *Water Resour. Res.*, *19*(3), 677–690, 1983.

Loaiciga, H. A., and M. A. Marino, The inverse problem for confined aquifer flow: Identification and estimation with extensions, *Water Resour. Res.*, *23*(1), 92–104, 1987.

Neuman, S. P., Calibration of distributed groundwater flow models viewed as a multiple-objective decision process under uncertainty, *Water Resour. Res.*, *9*(4), 1006–1021, 1973.

Rubin, Y., and G. Dagan, Stochastic identification of transmissivity and effective recharge in steady groundwater flow, 1, Theory, *Water Resour. Res.*, *23*(7), 1185–1192, 1987a.

Rubin, Y., and G. Dagan, Stochastic identification of transmissivity and effective recharge in steady groundwater flow, 2, Case study, *Water Resour. Res.*, *23*(7), 1193–1200, 1987b.

Schweppe, F. C., *Uncertain Dynamic Systems*, 563 pp., Prentice-Hall, Englewood Cliffs, N. J., 1973.

Sun, N.-Z., *Inverse Problems in Groundwater Modeling*, 352 pp., Kluwer, Norwell, Mass., 1994.

Wagner, B. J., and S. M. Gorelick, Reliable aquifer remediation in the presence of spatially variable hydraulic conductivity: From data to design, *Water Resour. Res.*, *25*(10), 2211–2225, 1989.

Yeh, T.-C. J., A. L. Gutjahr, and M. Jin, An iterative cokriging-like technique for groundwater flow modeling, *Ground Water*, *33*(1), 33–41, 1995.

Yeh, W. W.-G., Review of parameter identification procedures in groundwater hydrology: The inverse problem, *Water Resour. Res.*, *22*(1), 95–108, 1986.

P. Kitanidis, Department of Civil Engineering, Stanford University, Terman Engineering Center, Stanford, CA 94305-4020. (e-mail: pkk@ce.stanford.edu)

(Received January 14, 1997; revised May 6, 1997;
accepted June 2, 1997.)