Nonlinear inversion by stochastic relaxation with applications to residual statics estimation

Daniel Rothman

Abstract

Several data processing problems in reflection seismology are cast as general inverse problems, and are solved by maximizing the posterior probability of a model, given the observed data and a prior probability function for the model. Both this Bayesian solution and the computational techniques employed may be generally applicable: no assumptions of local linearity or Gaussian statistics are necessary. The only restriction is that the conditional probabilities of the model parameters exhibit local dependencies, or, specifically, that the model be expressible as a stochastic process called a *Markov random field*. Maximizing posterior probabilities for this relatively unrestricted class of problems is usually considered to be computationally intractable due to the existence of many local extrema. By making an analogy with statistical physics, however, it is shown that many large-scale nonlinear inverse problems that exhibit these local characteristics may be solved by a method that can yield solutions superior to previous efforts. This inversion procedure is successfully applied to the problem of residual statics estimation. The well-known problem of "cycle-skipping" is effectively attacked because no assumptions of local linearity are made. Further applications and extensions of the method are proposed.

Introduction

Many of the problems encountered in geophysical data processing are essentially problems of data inversion: data are collected on the Earth's surface, and we try to formulate a model of the Earth's interior, given these surface observations. The problem may be approached deterministically, in which we consider observed data to be error free, and implement a procedure which inverts this "exact" data to obtain a solution. Alternatively, we may view the data and the model as stochastic processes, and invert the observed data to

obtain the most probable solution, or, if possible, a probability distribution for the model.

The inversion process is viewed here from a probabilistic perspective. The proposed method incorporates a meaningful joint prior probability distribution for the model, from which the conditional posterior probability of the model, given the observed data, is derived. The posterior distribution is maximized to obtain the most probable model. This is a form of Bayesian estimation known as *maximum a posteriori* estimation.

The solution does not rely on the usual assumptions of local linearity and/or Gaussian noise. The method does, however, require that the model be expressible as a stochastic process known as a *Markov random field.* This limits us to problems in which the conditional probabilities of parameters exhibit local dependencies, but this is not a stringent restriction; many inverse problems in reflection seismology have this characteristic.

The maximization of posterior probabilities for this relatively unrestricted class of problems is a notoriously difficult optimization problem because of the existence of many local extrema. Following the recent work of Geman and Geman (1983) in image restoration, this problem is surmounted by first observing the equivalence of Markov random fields and Gibbs distributions. [Gibbs (or canonical) distributions usually occur in the description of systems in equilibrium in statistical physics.] The Markov-Gibbs equivalence leads easily to the conclusion that the posterior distribution of the model is also Gibbsian. This posterior probability is then maximized by employing a method of stochastic relaxation recently devised by Kirkpatrick et al (1983). This optimization technique implicitly assumes that the underlying probability distribution of the model is Gibbs, and effectively attacks the difficult problem of local extrema.

An application to the problem of residual statics estimation is illustrated. The statics solution is obtained without making any assumption of local linearity or Gaussian errors. Thus the solution does not suffer from the well-known problem of "cycle-skips" that conventional solutions exhibit.

Further theoretical extensions and applications of these methods in seismic data processing are proposed. Particular attention is given to the problems of frequency-dependent surface-consistent statics and missing data restoration.

The inverse problem

Consider a physical system (for example, the Earth) that is parameterized by a model, a finite set of parameters $\mathbf{M} = \{M_1, \ldots, M_s\}$ in the s-dimensional *model space.* An experiment performed in this physical system produces a finite set of data $\mathbf{D} = \{D_1, \ldots, D_r\}$ in the r-dimensional *data space.* In our notation the upper-case roman letters are (random)

variables which take on specific values $\mathbf{m} = \{m_1, \ldots, m_r\}$ or $\mathbf{d} = \{d_1, \ldots, d_s\}$.

Consider **G** to be some (nonlinear) function of the model **m** that describes the outcome of the experiment. Then we may represent the observed data **d** as

$$\mathbf{d} = \mathbf{G(m)} + \mathbf{n} \ , \tag{1}$$

where $\mathbf{n} = \{n_1, \ldots, n_r\}$ is a realization of the random noise **N**, assumed to be independent, identically distributed, and independent of **m**. This equation represents the most ambitious of geophysical inverse problems - the determination of the entire underlying model given the observed data. Smaller (though still important) problems entail solving for a more limited model, a subset $\overline{\mathbf{m}}$ of **m**. For example, $\overline{\mathbf{m}}$ could be the residual static time shifts associated with the near-surface, while **m** would include the entire velocity and reflectivity structure of a survey area. If we let $\overline{\mathbf{m}}$ be the *unknown* part of the model, and let $\hat{\mathbf{m}}$ be the remainder of the model that is assumed known, we may then rewrite (1) as

$$\mathbf{d} = \mathbf{G}(\overline{\mathbf{m}}; \hat{\mathbf{m}}) + \mathbf{n} \ . \tag{1a}$$

This second equation is more realistic, but for clarity we usually assume the form (1) throughout our discussion.

This paper primarily addresses the problem of inverting (1) when **G** is nonlinear and/or **n** is non-Gaussian (though the derived solution is equally valid for linear, Gaussian problems). Existing techniques usually rely on a linearization of **G** and the assumption of Gaussian statistics. Linearization is accomplished by providing an initial guess of **m** that is sufficiently close to the correct solution. The remaining perturbation of the model is then linearly related to the change in **d** due to the initial guess of **m**. Though often successful, these methods have serious drawbacks, most notably the necessity of a reasonable initial guess. Good accounts of linearized inversion techniques are in Aki and Richards (1980) and Parker (1977).

Tarantola and Valette (1982) developed an imaginative solution to the inverse problem (1) that is valid, in principle, for any degree of nonlinearity or any noise statistics. They stated the inverse problem as the combination of two states of information: the observed data and an estimate of its errors, along with the theoretical information contained in **G** and an estimate of its errors. Their solution is in the form of probability density functions and is essentially Bayesian in nature, but it reaches beyond Bayesian statistics because it explicitly includes both theoretical and observational errors. Their general solution presents a severe practical problem, however, because it is computationally intractable for problems with more than a just a few parameters (Rothman, 1983). The presentation here follows the spirit of the work of Tarantola and Valette, but with the additional attraction of

computability. A significant prior probability function for the model is specified, and the posterior probability function is obtained via Bayes' rule. The Bayesian approach presented here is less general than the solution of Tarantola and Valette (because theoretical errors are not incorporated), but it leads naturally to an algorithmic solution.

The Bayesian solution

We seek the most probable model, given the observed data and a prior probability distribution for the model. Thus we maximize the *posterior* probability

$$P(\mathbf{M}=\mathbf{m} \mid \mathbf{D}=\mathbf{d}) = \frac{P(\mathbf{D}=\mathbf{d} \mid \mathbf{M}=\mathbf{m})\, P(\mathbf{M}=\mathbf{m})}{P(\mathbf{D}=\mathbf{d})} . \tag{2}$$

When the model space is large (as it usually is in reflection seismology) there may be $s = 1000$ or more distinct parameters to solve for simultaneously. For nonlinear and/or non-Gaussian problems, the maximization of (2) then appears to be an overwhelming computational task. If each parameter may take on one of q states, then there are q^s possible solutions, making an exhaustive, point-by-point search for the global maximum a computational impossibility, even for the unlikely case of a binary ($q = 2$) model. Thus conventional attempts at the maximization of (2) or its equivalent have usually resorted to a linearization or a restriction to Gaussian errors. The method presented here for the maximization of the posterior distribution assumes only that the model parameters are related to each other locally, rather than globally, as described in the next section.

Markov random fields

Stochastic processes that exhibit local interaction are often expressible as *Markov random fields*. The fundamental assumption of this paper is that the model **M** has these local characteristics. This leads to several significant computational advantages which will be discussed in detail. In this section, Markov random fields are introduced in the context of geophysical inversion.

A Markov random field is essentially a multidimensional generalization of a Markov chain. Recall that a Markov chain is a sequence in which the conditional probability of an event at time t depends only on the value of the sequence at time $t-1$. Since the event at time t is independent of all times other than $t-1$, we may write, for a random sequence $\mathbf{M} = \{M_0, M_1, \ldots, M_t\}$

$$P(M_t = m_t \mid M_{t-1} = m_{t-1}, \ldots, M_0 = m_0) = P(M_t = m_t \mid M_{t-1} = m_{t-1}) .$$

A simple generalization leads to the consideration of a two-dimensionally indexed model $\mathbf{M} = \{M_{ij}\}$. The M_{ij} determine a Markov random field if the value of each M_{ij} depends only on a *neighborhood* A_{ij} of (i,j). A_{ij} might contain only the nearest neighbors:

$$A_{ij} = \{ (i+1,j), (i-1,j), (i,j+1), (i,j-1) \} .$$

The neighborhood structure depends on the physical characteristics of the model; other, more general neighborhood structures are possible (and will be employed here). A Markov random field with an arbitrary neighborhood structure A_{ij} is stated as

$$P[M_{ij} = m_{ij} \mid M_{kl} = m_{kl}, (k,l) \neq (i,j)] = P[M_{ij} = m_{ij} \mid M_{kl} = m_{kl}, (k,l) \in A_{ij}] . \quad (3a)$$

It is additionally required that the probability of all models be positive:

$$P(\mathbf{M} = \mathbf{m}) > 0 \text{ for all } \mathbf{m} . \quad (3b)$$

We shall retain the notion of a two-dimensional model throughout our discussion, though generalization to higher dimensions is mostly a matter of notation.

The theory of Markov random fields originated with the study of systems in equilibrium in statistical physics. In the past decade, however, Markov random fields have been used to model a variety of phenomena in the physical, biological, and social sciences. An excellent review of the theory, applications, and history of Markov random fields is in the book by Kinderman and Snell (1980).

Many important problems in reflection seismology exhibit the local neighborhood dependencies characteristic of Markov random fields. In the problem of residual statics estimation, for example, individual parameters are shot and receiver statics, and the "neighborhood" comprises the shot and receiver statics located within a cablelength of each other. For the model M_{ij} and neighborhood structure A_{ij}, i would index shots and j would index receivers. Since statics are *relative* time shifts, only the statics within a cablelength of each other interact in an immediate sense. Indeed, the physical basis for a Markovian model of statics may be argued from a geologic point of view: the composition of the near-surface is the result of localized, spatial interaction due to the effects of erosion and other geologic activities. The seismic cable then provides a practical limitation for measuring this local interaction. Another, perhaps more obvious, instance of a Markov random field is the problem of missing data restoration; here it is the neighboring spatial and temporal points that determine estimates of missing data.

It is the local nature of Markov random fields that is most important in the inversion problem. The Markov model allows parameters to exhibit significant correlations over much of the model, but computations are localized and therefore realizable. Before demonstrating

this, however, we discuss in the next section how the characterization of **M** in terms of the Markov conditional probabilities (3a) leads to a simple expression for the joint prior probability distribution of the model.

Markov-Gibbs equivalence

The identification of a model as a Markov random field is based only on the existence of localized, conditional probabilities. Conditional probabilities alone do not in any obvious way determine the joint prior probability function $P(\mathbf{M} = \mathbf{m})$ needed for obtaining the Bayesian estimate (2). The existence of a Markov random field according to conditions (3) does, however, determine a joint probability measure known as a *Gibbs distribution* (Kinderman and Snell, 1980, and the references therein). Gibbs (or canonical) distributions arise in statistical physics in the study of systems in equilibrium. We say that a random field **M** is Gibbs if

$$P(\mathbf{M} = \mathbf{m}) = \frac{1}{Z} e^{\frac{-E(\mathbf{m})}{kT}} . \tag{4}$$

$E(\mathbf{m})$ is called the *energy*, and is the sum of local *potentials* $V_{A_{ij}}(\mathbf{m})$ such that

$$E(\mathbf{m}) = -\sum_{i,j} V_{A_{ij}}(\mathbf{m}) . \tag{5}$$

The $V_{A_{ij}}$ are evaluated over the same neighborhood structure A_{ij} used to specify the conditional probabilities (3a), i.e., $V_{A_{ij}} = 0$ for $i, j \not\in A_{ij}$. T is *temperature* and k is *Boltzmann's constant*. Z is the normalizing constant

$$Z = \sum_{\mathbf{m}} e^{\frac{-E(\mathbf{m})}{kT}} , \tag{6}$$

called the *partition function*. One of the pleasing aspects of the Markov-Gibbs equivalence is that the purely probabilistic notion of a Markov random field is equated to the physically based Gibbs distribution. Gibbs models describe the interaction of a macroscopic system in thermal equilibrium in the same way that spatial Markov models describe local dependencies. For our present purposes it is important to note that the Gibbs measure (4) specifies the *joint* probability distribution (describing large-scale correlations) of the model, while the conditional probabilities (3a) are generally much less useful.

The Markov-Gibbs equivalence is best intuited in view of the common neighborhood structure A_{ij}. For a Markov random field, the conditional probabilities are stated in terms of neighborhoods, while for a Gibbs distribution the energy E is the sum of potentials V

measured over the same neighborhood structure. That a Gibbs distribution implies a Markov random field is straightforward: the additive nature of the energy function (5) allows the appropriate cancellations when the conditional probabilities (3a) are expressed. The specification of a Markov random field does not, however, uniquely determine a Gibbs potential function except for special cases (Kinderman and Snell, 1980). The potential function is of paramount importance because it describes how individual parameters interact with each other; in particular, low energy signifies a preferred model. The correct choice of potential function is therefore critical for the solution of a geophysical inverse problem. In residual statics estimation, the potential function will be chosen to be the power in stacked common midpoint gathers. In the next section we show how decreasing energy (or, for statics, increasing power) is related to increasing probability.

The Gibbs posterior

We now revise our statement of the posterior probability (2). Assuming that the model **M** is a Markov random field, we may write the joint prior probability function for the model as

$$P(\mathbf{M} = \mathbf{m}) = \frac{1}{Z} e^{\frac{-E(\mathbf{m})}{T}}, \tag{7}$$

where for convenience we set $k = 1$. The choice of energy function and temperature is of course important, and will be discussed later in the context of specific applications. Our present purpose, however, is to show that the posterior probability $P(\mathbf{M} = \mathbf{m} \mid \mathbf{D} = \mathbf{d})$ is also a Gibbs distribution (Geman and Geman, 1983). This will lead naturally to the implementation of a Monte Carlo optimization procedure described by Kirkpatrick et al (1983).

$P(\mathbf{D} = \mathbf{d})$ is assumed to be uniformly distributed over the data space, so by substituting (7) into (2),

$$P(\mathbf{M} = \mathbf{m} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} P(\mathbf{D} = \mathbf{d} \mid \mathbf{M} = \mathbf{m}) \, e^{\frac{-E(\mathbf{m})}{T}} \tag{8}$$

where Z is now a new constant. The noise $\mathbf{N} = \{N_1, \ldots, N_r\}$ is assumed to be independent, identically distributed, and independent of the model. The prior distribution of the noise has little bearing in principle, but for both clarity and generality the distribution of the noise is assumed to be zero-mean and expressible as the generalized exponential function

$$P(\mathbf{N} = \mathbf{n}) \propto e^{-\frac{1}{2}\left[\frac{\|\mathbf{n}\|_p}{\sigma}\right]^p} \tag{9}$$

where $\| \cdot \|_p$ is the L^p norm such that $(\| \mathbf{n} \|_p)^p = \sum_i^r n_i^p$. Note that if $p = 2$ the noise is

Gaussian and if $p = 1$ the noise is exponential. We now solve for the posterior. We rewrite (8) as

$$P(\mathbf{M} = \mathbf{m} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} P[\mathbf{D} = G(\mathbf{m}) + \mathbf{n} \mid \mathbf{M} = \mathbf{m}]\, e^{\frac{-E(\mathbf{m})}{T}} \tag{10}$$

$$= \frac{1}{Z} P[\mathbf{N} = \mathbf{d} - G(\mathbf{m}) \mid \mathbf{M} = \mathbf{m}]\, e^{\frac{-E(\mathbf{m})}{T}} . \tag{11}$$

Since the noise is independent of the model,

$$P(\mathbf{M} = \mathbf{m} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} P[\mathbf{N} = \mathbf{d} - G(\mathbf{m})]\, e^{\frac{-E(\mathbf{m})}{T}} \tag{12}$$

and by inserting (9)

$$P(\mathbf{M} = \mathbf{m} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z} e^{\frac{-E(\mathbf{m})}{T} - \frac{1}{2}\left[\frac{\|\mathbf{n}\|_p}{\sigma}\right]^p} \tag{13}$$

where Z is again a new normalizing constant. The noise term in the exponent is just a constant which may also be absorbed into Z, so

$$P(\mathbf{M} = \mathbf{m} \mid \mathbf{D} = \mathbf{d}) = \frac{1}{Z}\, e^{\frac{-E(\mathbf{m})}{T}} , \tag{14}$$

establishing that the posterior distribution is also Gibbs with energy function $E(\mathbf{m})$. Geman and Geman derive some additional results showing that the posterior neighborhood structure is slightly modified to include "second-order" neighbors (i.e., neighbors of neighbors). For computational purposes, however, we assume that the prior and posterior neighborhood structures are approximately equal.

The model which best fits the data, in the sense of Bayesian inference, is determined by maximizing (14). This is *maximum a posteriori* estimation. The posterior probability is maximized when energy is minimized - this is analogous to the situation of thermal equilibrium in statistical physics, where the most probable molecular configurations occur at the lowest energies. For the case of residual statics, it will be shown that the most probable statics model occurs when the *negative* stack power is minimized. Maximizing (14) by conventional gradient techniques is virtually impossible for all but the simplest models, because of the existence of many local extrema. Recently, however, Kirkpatrick et al (1983) devised a method of stochastic relaxation that globally maximizes functions of the form (14). Their method is reviewed in the following section.

Stochastic relaxation

In many important geophysical problems, the energy function in (14) will be a nonlinear, highly variable, and possibly non-differentiable function of the model. In residual statics estimation, for example, the model comprises the surface-consistent time shifts, and energy is expressed as the power in the stack of time-shifted traces. If the time shifts are large relative to the dominant period of the data, or if the signal-to-noise ratio is low, then extremalization of power is not a straightforward task. Many local extrema of power will exist, but simply locating a local extremum may cause incorrect alignment of traces - this is the "cycle-skipping" problem. Thus the *global* extremum must be found.

Kirkpatrick observed that many large-scale optimization problems can be modeled in analogy with the behavior of macroscopic systems in statistical physics. He noted that the problem of finding a global extremum when many local extrema exist is akin to the procedure of chemical *annealing*. Annealing is the method by which crystals are grown - a substance is first melted, and then cooled very slowly until the crystal is formed. The cooling schedule is crucial, because a non-crystalline, metastable glass may form instead. Kirkpatrick called his optimization algorithm "simulated annealing" - he viewed the occurrence of a crystal as analogous to finding the global extremum, and the occurrence of a glass as the counterpart to wrongly selecting a local extremum.

Our characterization of an inverse problem in terms of Markov random fields and Gibbs distributions lends itself naturally to Kirkpatrick's analogy with statistical physics. "Energy" corresponds to an objective function (i.e., power) and "annealing" corresponds to lowering the temperature T, causing the largest peak in the posterior probability function (14) to become progressively more exaggerated. T is expressed in the same units as the objective function, and it controls the rate of convergence.

Kirkpatrick's technique is essentially a variant of a Monte Carlo procedure due to Metropolis et al (1953). Metropolis addressed the problem of randomly sampling from a Gibbs distribution at constant temperature, thereby simulating the behavior of a physical system in thermal equilibrium. The Metropolis algorithm proceeds as follows. For each element M_{ij} of a model **M**, a random perturbation is made, and the change in energy, ΔE, is computed. If $\Delta E \leq 0$, the perturbation is accepted. If ΔE is positive then the perturbation is accepted with probability

$$P(\Delta E) = e^{\frac{-\Delta E}{T}} . \tag{15}$$

This conditional acceptance is easily implemented by choosing a random number α uniformly distributed between 0 and 1. If $\alpha \leq P(\Delta E)$ then the perturbation is accepted; otherwise

the existing model is retained. Random perturbation according to these rules eventually causes the system to reach equilibrium, or a condition of *maximum entropy*,† in which configurations **m** are realized with the Gibbs probability (14).

Kirkpatrick's optimization technique slowly lowers the temperature T during execution of the Metropolis algorithm. If the system is cooled sufficiently slowly and equilibrium conditions are maintained, the model eventually converges to a (ground) state of minimum energy, or maximum a posteriori probability. The essential characteristic of this optimization procedure is that local extrema may exist, but the algorithm can escape from these local extrema and find the global extremum. Thus, in the Bayesian estimation problem here, perturbations which *lower* probability are accepted in accordance with (15), but the final result yields the model associated with *maximum* probability.

Application to residual statics

The conventional residual statics model (Wiggins et al, 1976; Taner et al, 1974) expresses observed static time deviations t_{ij} of normal moveout corrected traces associated with the ith shot and jth receiver as

$$t_{ij} = s_i + r_j + g_k + c_k x_{ij}^2 \ . \tag{16}$$

The unknown surface-consistent time shifts due to near-surface velocity anomalies underneath the ith shot and jth receiver are denoted by s_i and r_j, respectively. The unknown subsurface-consistent part of the time shift due to variations in geologic structure at the kth midpoint is given by g_k. Finally, the last term represents the component of the time shift due to residual normal moveout; c_k is called the residual normal moveout coefficient, and x_{ij} is the distance between shot i and receiver j. The c_k are included in an attempt to account for the usually imperfect stacking velocities used prior to measuring t_{ij}.

Among the many well-known deficiencies of the model (16), we specifically address only one - the problem of measuring t_{ij}. This is usually performed by crosscorrelating traces against some reference model. If s_i and r_j are small relative to the dominant period in the data and if the signal-to-noise ratio of the data is reasonably good, then the weighted least squares approach of Wiggins et al is usually successful at solving for the parameters on the right side of (16). Serious problems may occur, however, when either of these two conditions are not met. The problem is that the measured t_{ij} may be grossly in error - this is the

†Gibbs distributions may be derived in statistical physics by maximizing entropy (Reif, 1965). Clearly, the proven capabilities of maximum entropy estimation techniques (Shore and Johnson, 1980) may yield much theoretical insight into this procedure.

"cycle-skipping" problem. These errors are highly non-Gaussian, so the least squares approach may fail severely. Donoho (1979) noted this, and employed methods of robust estimation that diminished the degrading influence of these non-Gaussian errors on the least squares solution.

The problem of non-Gaussian errors in t_{ij} is approached here by skipping the correlation-time picking step entirely. Let the observed data **d** be the seismic traces *instead* of the time deviations t_{ij}. Assume that the entire Earth model, source waveforms, etc., are known *except* for the near-surface timing delays. Then if **s** and **r** represent the unknown shot and receiver statics, respectively, and if $\hat{\mathbf{m}}$ represents the known part of the Earth model, we may write

$$\mathbf{d} = \mathbf{G}(\mathbf{s}, \mathbf{r}; \hat{\mathbf{m}}) + \mathbf{n} \ . \tag{17}$$

G represents a time shift, and the equation is nonlinear (and potentially highly variable) with respect to **s** and **r**. Note that the traveltime model (16) is a *linearization* of (17), and thus suffers the usual problems of linearized, nonlinear inversion; in this instance it is the inability to handle non-Gaussian errors. Note also that (17) does not contain the subsurface-consistent terms g_k and c_k - this will be explained shortly.

To justify the application of the Markov-Gibbs model to invert (17), we first make the following observation. If the traces in a common midpoint gather are identical, except for a time shift and uncorrelated noise, then the power in the stack of these traces is maximized when the time shifts are all equal. Thus for a given shot static s_i, we seek the shift which maximizes (with respect to s_i) the power in the stacks of the midpoint gathers associated with shot i. The value for s_i depends of course on the *other* shot and receiver statics associated with the remaining traces in this set of midpoint gathers, and maximization of stack power depends *only* on these neighboring statics in an immediate sense. Note that the estimates of neighboring statics interact with one another, so in general the s_i and r_j cannot be systematically chosen so that power is always maximized in their respective midpoint gathers - this could lead to a local maximum with respect to power (the "cycle-skipping" problem). However, these nearby statics make up "neighborhoods" of s_i and r_j in the sense of a Markov random field - the range of s_i and r_j in the neighborhood is fixed by the seismic cablelength. Thus, given the Markov random field, we need only choose an energy function in order to specify the joint Gibbs distribution for the statics. We choose to equate local stack power with the potential function V in (5), and seek the maximum power that is attainable with surface-consistent statics corrections.

The optimization problem is formally stated as follows. Let each moveout-corrected seismic trace be characterized according to the shot i from which the source energy

originated, the receiver j at which the wavefield was recorded, the midpoint y (the point midway between shot i and receiver j), and the offset h (the distance between the shot and receiver). The data are expressed as $d_{yh}(t - s_{i(y,h)} - r_{j(y,h)})$, where t is the true reflection time (assuming no static delay). The subscripts i and j are both functions of y and h. Stacking is expressed as a sum over h for each y. Each shot static s_i influences (in an immediate sense) the stack power of only a subset Y_{s_i} of all midpoints y; likewise each receiver static r_j affects a subset Y_{r_j}. The portion of the stack power influenced by shot static s_i is given by

$$V_{s_i}(\mathbf{s},\mathbf{r}) = \sum_{y \in Y_{s_i}} \sum_t \left[\sum_h d_{yh}(t - s_{i(y,h)} - r_{j(y,h)}) \right]^2 . \tag{18a}$$

Similarly, the portion of the stack power influenced by receiver static r_j is

$$V_{r_j}(\mathbf{s},\mathbf{r}) = \sum_{y \in Y_{r_j}} \sum_t \left[\sum_h d_{yh}(t - s_{i(y,h)} - r_{j(y,h)}) \right]^2 . \tag{18b}$$

Note that, due to the restriction on y, i and j vary essentially over a cablelength. To maximize the total power in the stack, we seek the **s** and **r** that minimize the energy in equation (5), now restated as

$$E(\mathbf{s},\mathbf{r}) = -\sum_i V_{s_i}(\mathbf{s},\mathbf{r}) - \sum_j V_{r_j}(\mathbf{s},\mathbf{r}) . \tag{19}$$

Use of the Kirkpatrick algorithm to minimize (19) necessitates continual perturbations of each s_i and r_j, each time computing V_{s_i} or V_{r_j}.

As noted earlier, the subsurface-consistent terms g_k and c_k are not included in this approach. The g_k represent timing differences due to geologic variation from midpoint to midpoint, and are useful only with models like (16) that depend on measurements of trace-to-trace time deviations. The power computations in (18a,b) are performed within midpoint gathers, however, where the g_k are constant, and therefore irrelevant. This is an important point. The statics solution presented here does *not* attempt a decomposition of structural and near-surface variations; thus the algorithm exhibits no inherent ambiguity between structure and statics. (Poorly determined long wavelength statics, however, may still be confused with structural variations). Residual normal moveout, although ignored, is still an important parameter. Estimation of the c_k in conjunction with a function like (18) is cumbersome, however. It is of course possible, but experience with conventional residual statics estimation shows that the residual normal moveout term is largely inconsequential. One reason is that the c_k represent time-averaged residual moveout, since calculations are usually performed over a window containing several reflections.

Implementation

The relaxation method is an iterative technique which continually creates samples from a Gibbs distribution while slowly decreasing the temperature parameter T. A convenient, data-dependent method for choosing the initial temperature T_0 is to compare the average input stack power p_0 with the average stack power p_r computed after applying uniformly random shot and receiver statics. For a given β, T_0 is then chosen such that

$$\beta = e^{-\frac{p_0 - p_r}{T_0}}.$$

β determines the degree of "melting" prior to "annealing." It is chosen between 0 and 1 and represents the probability of accepting a decrease in power by the amount $p_0 - p_r$ [see equation (15)]. Cooling then proceeds in any of a number of ways. The most consistently good results were attained with a logarithmic cooling function $T_k = T_0 / \log k$, where k represents the number of complete sweeps over the parameter space (one sweep incorporates one attempted perturbation for each shot and receiver static). Geman and Geman (1983) proved convergence for a cooling function of this form. Practically, the most important aspect of any cooling function is that it be slow, especially near the "critical temperatures" where convergence is rapid. The successful choice of an annealing schedule requires experience; ideally, the procedure would be interactive. I was usually able to produce a successful result after two or three trials. I found it best to set β very close to 1 for the initial run. Then, after a preliminary determination of the critical temperatures for the data, T_0 may be chosen much lower, in addition to choosing a k_0 such that $T_k = T_0 / \log(k_0 + k)$. A large k_0 substantially decreases the cooling rate.

Once started, the next question to resolve is when to stop. In my tests I collected run statistics every iteration, which is defined here to be twenty sweeps (twenty attempted perturbations per parameter). The algorithm then simply stops after an iteration in which few or no perturbations are permitted.

Synthetic data example - residual statics

The Markov-Gibbs assumption and the Kirkpatrick relaxation algorithm were tested on synthetic data that exhibit a severe surface-consistent statics problem. The data simulate the results of a survey conducted with a 12-trace cable, off-end shooting with a two receiver group gap, and with shots and receiver groups evenly spaced. There are 100 6-fold common midpoint gathers. The sampling rate is 4 msec. and the data contain frequencies between 5 and 60 Hz. The data, *without* statics, are shown in Figures 1a and 1b. Figure 1a shows four representative "moveout-corrected" common midpoint gathers, and Figure

1b is the common midpoint stack. The cablelength extends over 24 stacked traces. The signal-to-noise ratio (the total power of the signal divided by the total power of the noise) after stack is approximately 2.0. The entire dataset is scaled to an rms amplitude of 100. The "signal" is identical for all traces, except for the bulk time shift simulating a fault. These data represent the desired solution for the test illustrated in the following figures.

Figure 2 shows the statics model for shots and receivers; this model is a sample from a Markov random field. (The two one-dimensional functions **s** and **r** form the "random field" for the statics model). Using the data of Figure 1, the Markov random field was generated by executing the Metropolis algorithm at a high temperature (so that more than 90% of all perturbations were accepted) and stopping after one iteration. The perturbations were limited to vary between ±40 msec., in 4 msec. increments. A full range of statics was attained, so the combined effect of shot and receiver statics varies over ±80 msec.

Figure 3a shows the same common midpoint gathers of Figure 1a, but now with the traces shifted in accordance with the statics model in Figure 2. Figure 3b is the common midpoint stack after the model statics were applied. Due to the severity of the statics, virtually no indications of reflection events can now be observed. The data in Figures 3a,b are the input to the statics estimation algorithm.

Figures 4a-e illustrate the results of applying the statics algorithm. Three stages of the algorithm's execution are depicted; the stack after 241 iterations (4a), after 308 iterations (4b), and the final solution, after 454 iterations (4c), which closely resembles the desired solution in Figure 1b. Figure 4d shows the same four common midpoint gathers from Figure 3b; they are now depicted after the statics solution has been applied. For this example, $T_k = T_0 / \log(k_0 + k)$, with $T_0 = 4500$, $k_0 = 5000$, and k the number of sweeps (there are twenty sweeps per iteration). Allowable perturbations for shot and receiver statics fell within ±40 msec., in 4 msec. increments. Figure 4e shows a graph of average stack power versus iteration. Note that there is very little change in power until after approximately 230 iterations. After 300 iterations, an abrupt increase in power occurs. Sudden changes similar to this are analogous to "phase transitions" in statistical physics, and were repeatedly observed in virtually all tests. By the time iteration 308 was reached, the statics algorithm essentially completed its most important work - solving for the shorter wavelength statics, leaving only long wavelength residuals. The longer wavelengths are the most poorly determined components of the solution - this is as true for the linearized technique of Wiggins et al (1976) as it is here. By iteration 454 (the final solution), only a slight long wavelength residual remains. Although we observe here, as elsewhere, the fundamental ambiguity between long wavelength statics and structure, it is important to note that the drastic structural variation implied by the artificial fault does not influence the solution.

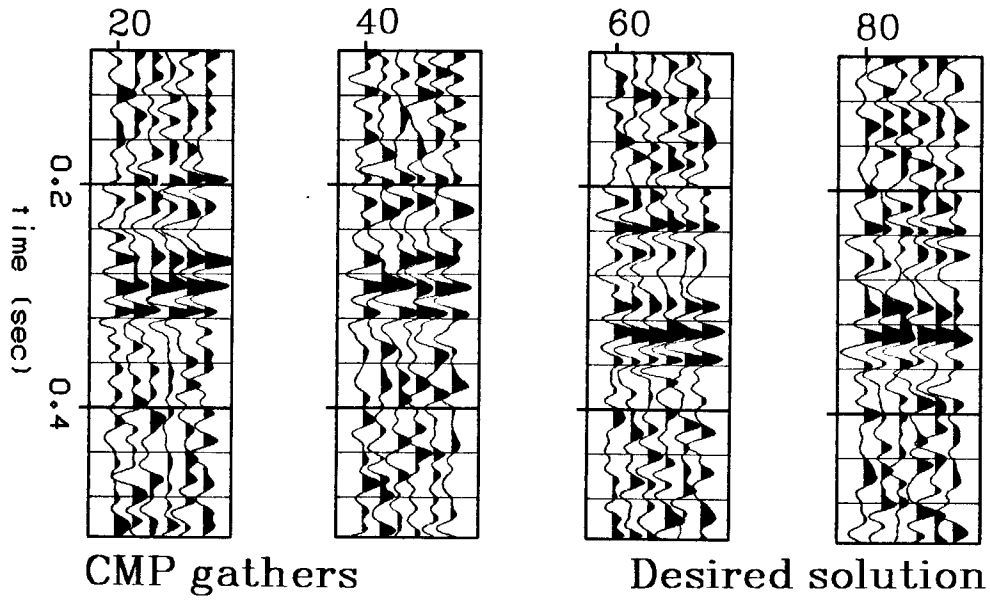

FIG. 1a. Four "moveout-corrected" common midpoint gathers. The gathers are shown without static shifts; there are 6 offsets in each gather. This correct alignment of traces represents the desired solution for pre-stack data.

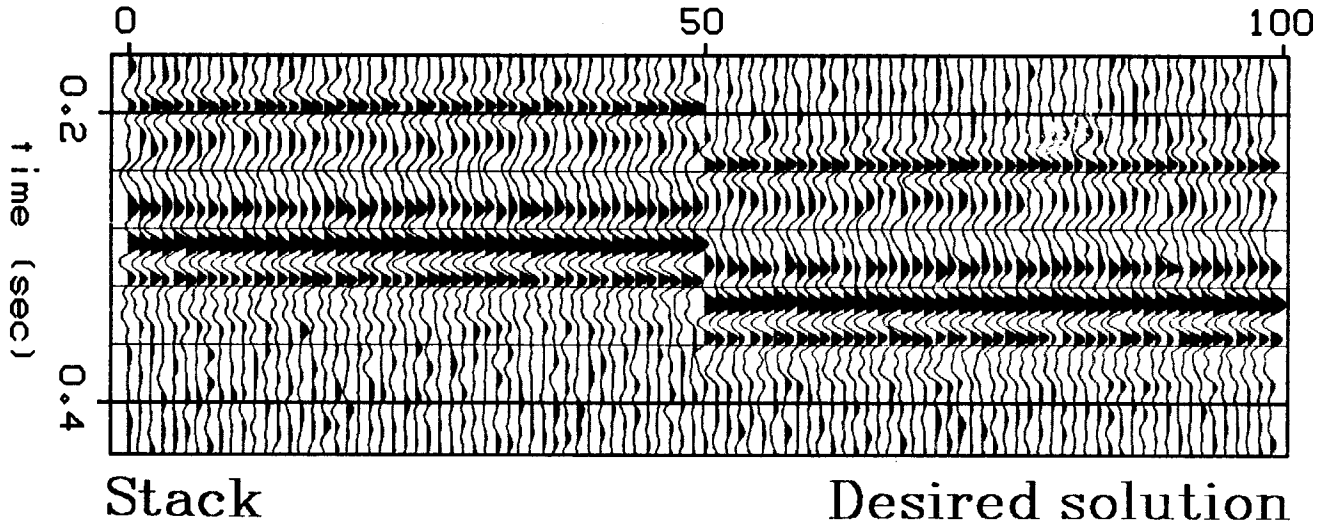

FIG. 1b. Common midpoint stack prior to the introduction of static shifts. The cablelength extends over 24 midpoints; there are 100 midpoints in total. The signal-to-noise ratio is approximately 2. This represents the desired solution for stacked data.

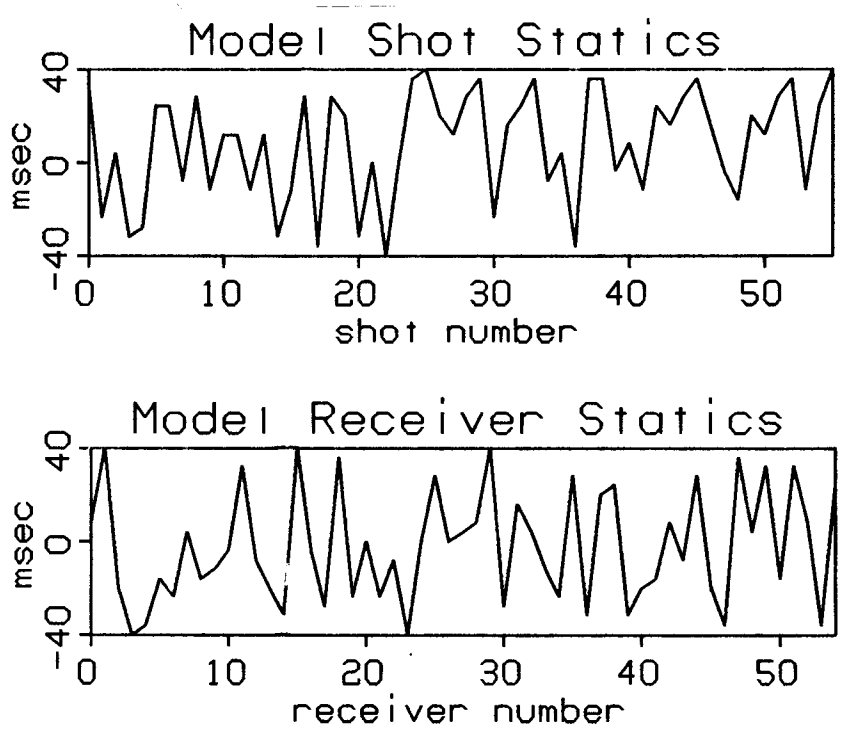

FIG. 2. The shot statics (above) and receiver statics used to generate the test data in Figures 3a,b. The statics model was created by taking a sample from a Markov random field. Statics range between ±40 msec. for both shots and receivers.

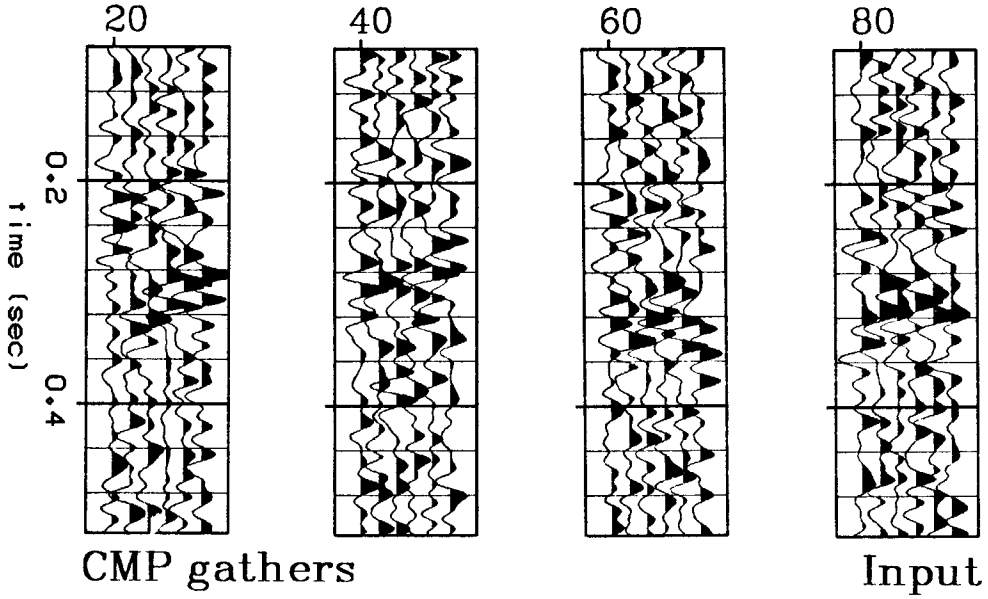

FIG. 3a. The common midpoint gathers of Figure 1a after the application of the static shifts in Figure 2. Note how the statics have degraded the appearance of the data.

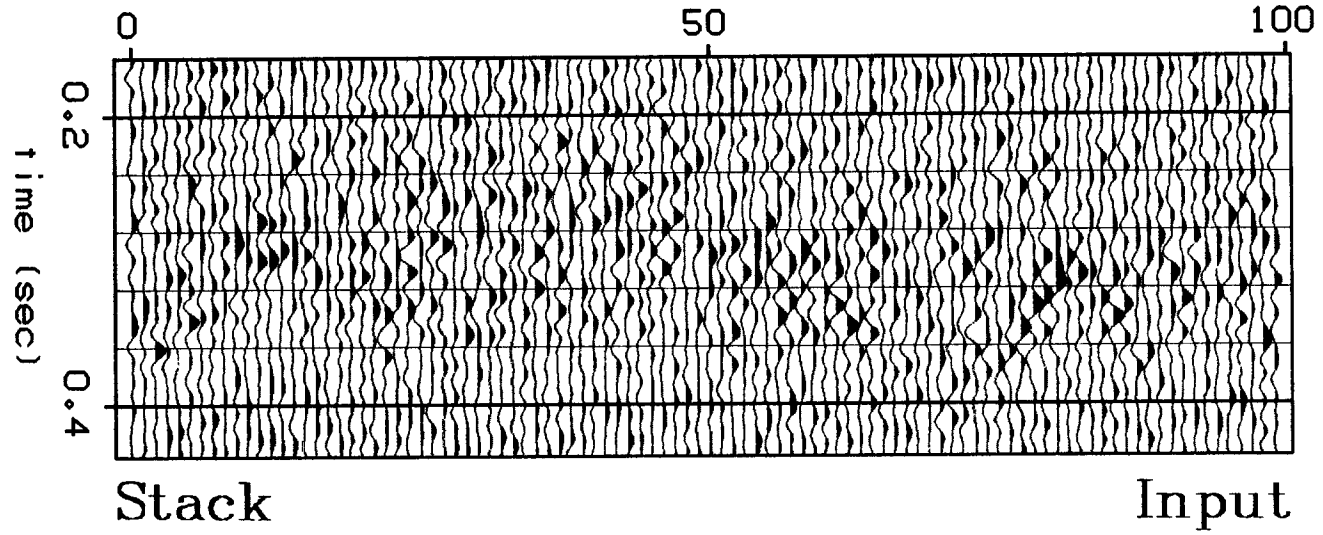

FIG. 3b. Common midpoint stack after application of the statics in Figure 2. Since the shifts are as much as 160 msec. apart, virtually no indication of the reflection events in Figure 1b can now be observed. The data in Figure 3a and 3b are the input to the statics estimation algorithm.

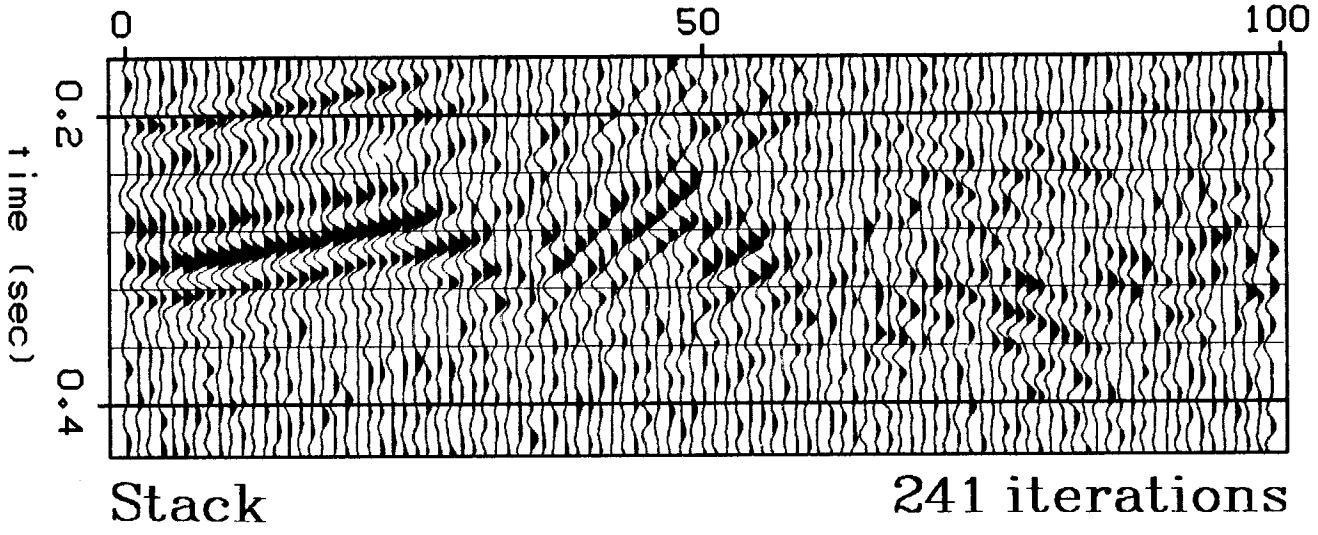

FIG. 4a. Common midpoint stack after 241 iterations of the statics estimation algorithm. Good convergence already appears on the left, though the remainder of the section exhibits the effects of misaligned traces.

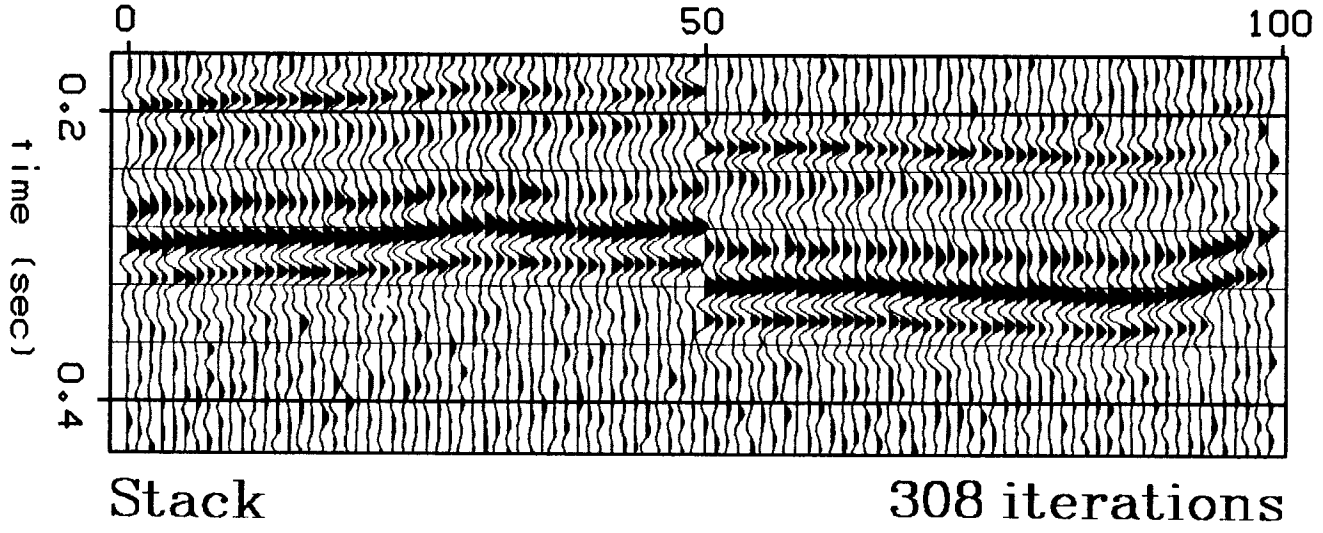

FIG. 4b. Common midpoint stack after 308 iterations. Although long wavelength statics remain to be resolved, the bulk of the algorithm's work is completed. Note that, despite the ambiguity between structure and long wavelength statics, the artificial fault at trace 50 is properly resolved.

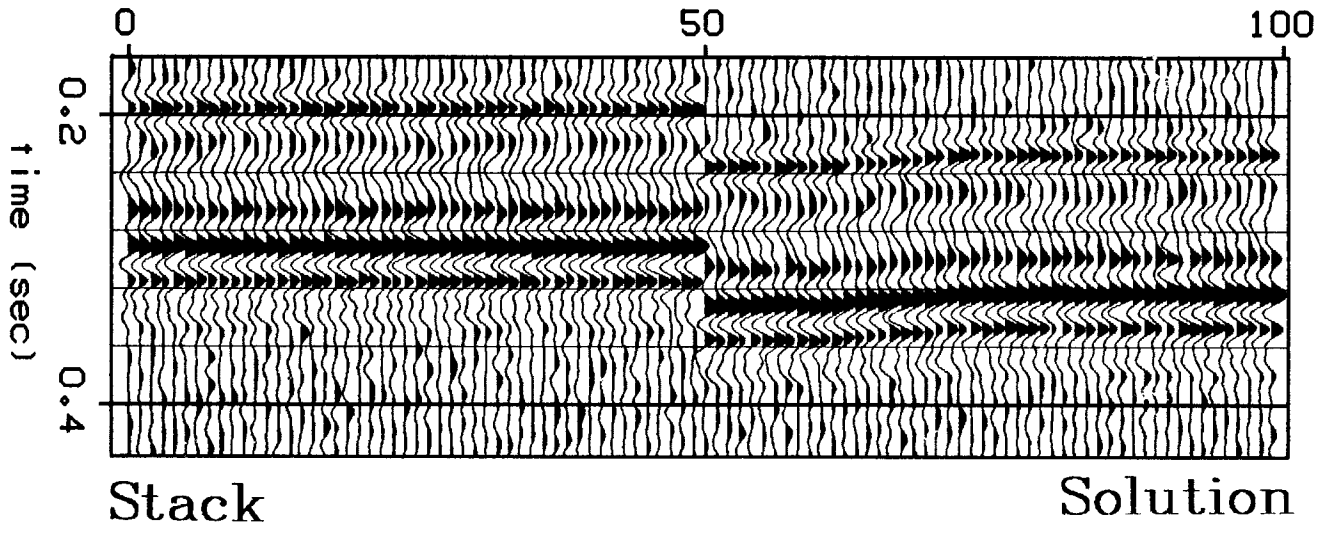

FIG. 4c. Common midpoint stack after 454 iterations. This is the final solution, and should be compared with the input (Figure 3b) and the known, desired solution (Figure 1b). The 8 msec. rise on the right half of the section is a result of poorly resolved long wavelength statics, due mostly to the noise contamination in the data.

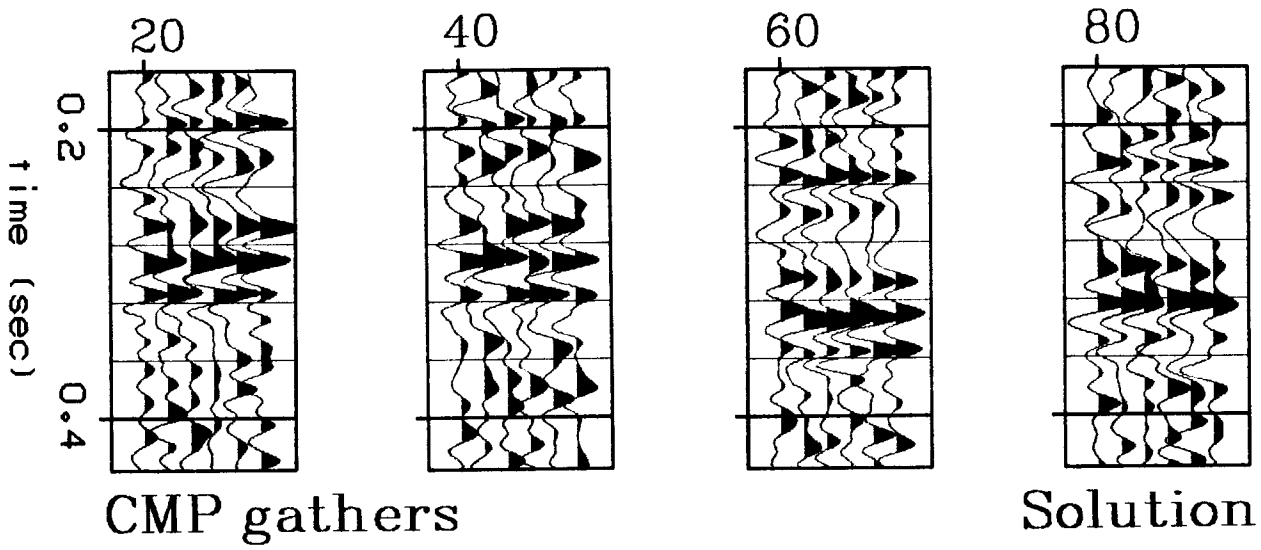

FIG. 4d. Common midpoint gathers after the statics solution has been applied. This should be compared with the input (Figure 3a) and the desired solution (Figure 1a). CMP 60 exhibits a slight error due to the poorly resolved long wavelength. The traces extend over less time now because the application of statics creates zeroes at early and late times.

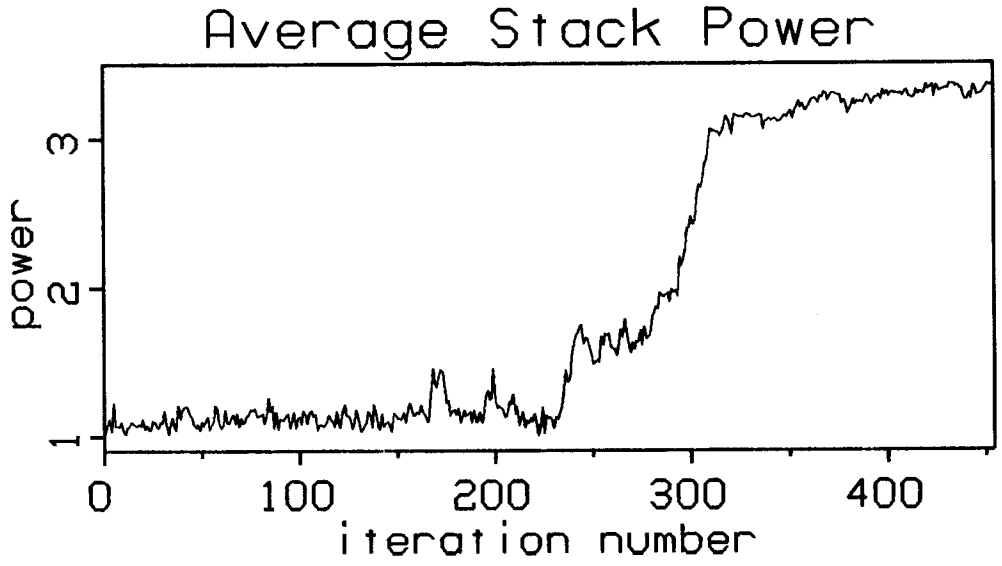

FIG. 4e. Graph of stack power versus iteration number for the test leading to the result in Figure 4c. The input stack power is normalized to 1. One iteration is defined to be 20 attempted perturbations per shot and receiver static. The final solution yields a stack power of 3.354, which is short of the true solution by 1.3%. Note the sudden increase in power after 300 iterations. This is analogous to a "phase transition" in statistical physics.

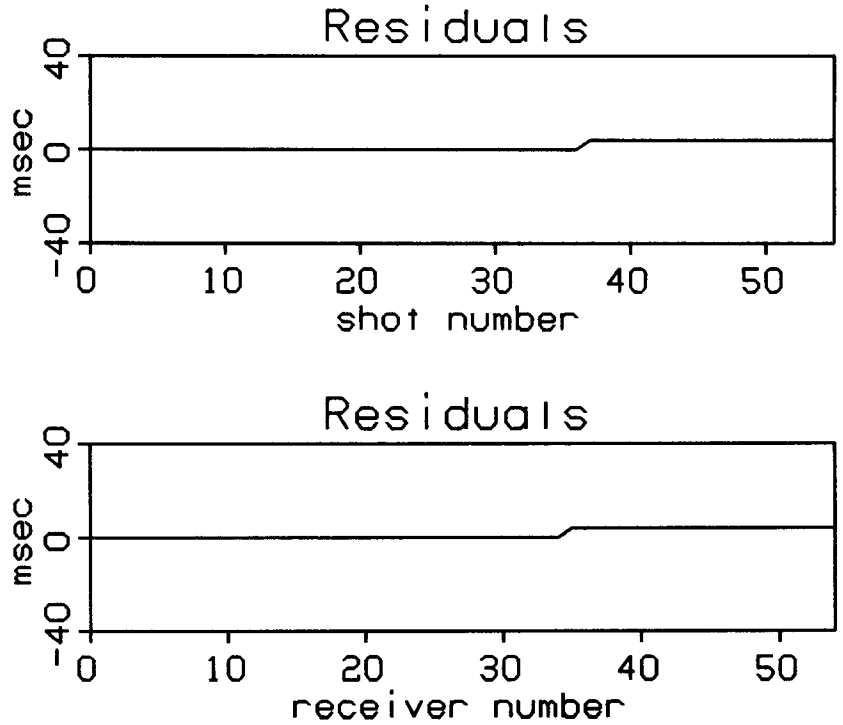

FIG. 5. Difference between the estimated statics and the true statics for the result in Figure 4c. The 8 msec. rise in the right half of Figure 4c is the result of the constant 4 msec. error for approximately the last 20 shot and receiver statics. The permitted perturbations for statics fell within ±40 msec., in 4 msec. (1 sample) increments. The degree of noise contamination for this test was too strong for the long wavelength residual to be resolved.

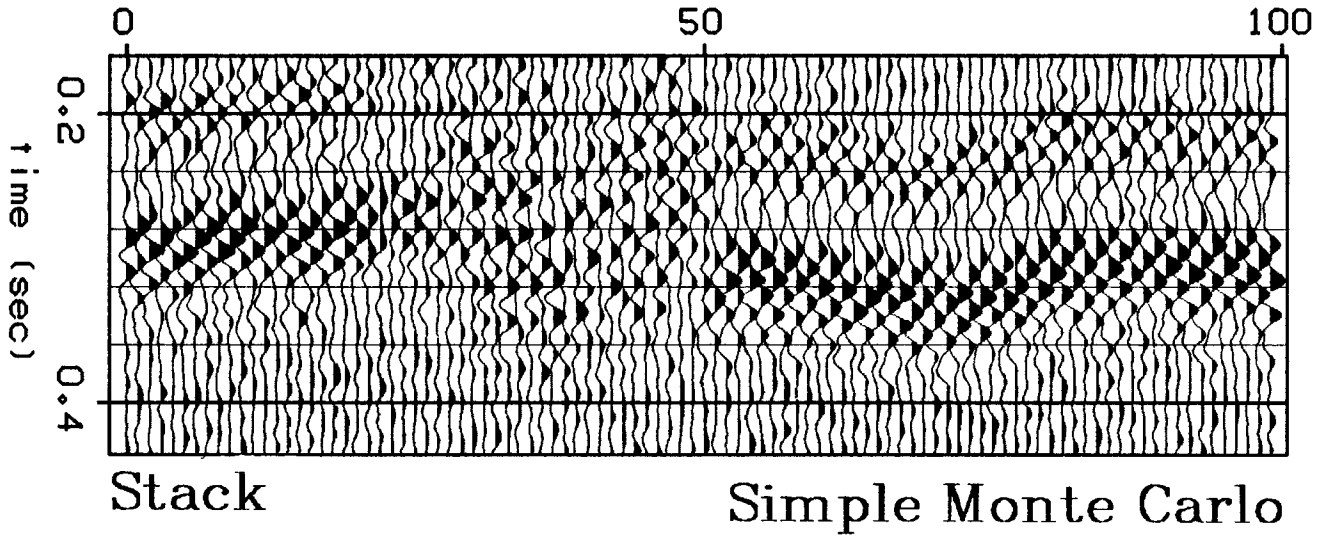

FIG. 6a. Common midpoint stack after applying the relaxation algorithm with constant $T = 0$ to the data of Figures 3a,b. Only perturbations which increased power were accepted. This simple Monte Carlo procedure produced not only a severe "cycle-skipping" problem, but substantial localized misrepresentation of structure. Full convergence (shown above) to this local extremum was achieved after only 28 iterations.

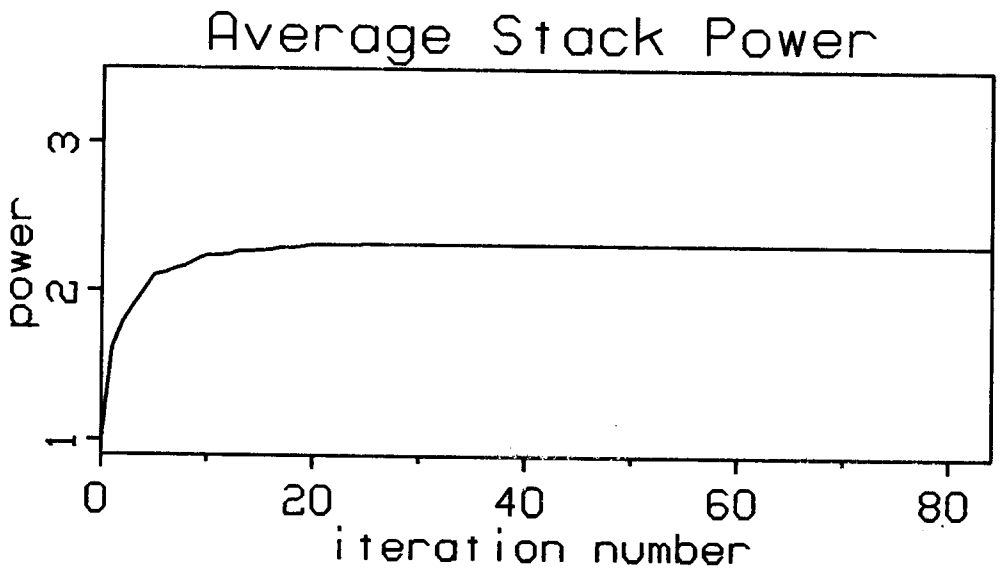

FIG. 6b. Graph of power versus iteration for the test leading to the result in Figure 6a. Convergence was achieved almost immediately after about 10 iterations. No perturbations were accepted after the 28th iteration, conclusively demonstrating that Figure 6a represents a local extremum of power.

The final normalized stack power for the solution in Figure 4c is 3.354. The known, desired solution has an average stack power of 3.399, so the computed solution is in error by approximately 1.3%. The difference between the estimated statics and the true statics is graphed in Figure 5. Note that, for both shots and receivers, the basic error occurs as a slight kink about two-thirds of the way along the line. The degree of noise contamination for this test was such that this long wavelength residual could not be resolved; other tests (not shown) with higher signal-to-noise ratios more successfully resolved the long wavelengths.

Figures 6a,b illustrate an example of "quenching" the solution - i.e., running the same algorithm, but always with $T = 0$. This is the simplest Monte Carlo procedure - only perturbations which increase power are accepted. The result is a trap in a poor local extremum that shows virtually no resemblance to the desired solution, demonstrating the necessity of accepting both increases and decreases in power. Figure 6a depicts the resulting stack, and Figure 6b shows the corresponding power versus iteration plot. Note that convergence was attained almost immediately after about 10 iterations. In this example, the algorithm was forced to continue for 85 iterations, though in actuality no perturbations were accepted after the 28th iteration. This conclusively demonstrates that Figure 6a represents a local extremum.

Future applications

The most immediate objective is the application of the residual statics algorithm to field data. Preliminary tests on data from Williston Basin, North Dakota were successful, but the results are not shown because the statics problem for these data can be solved by standard techniques, and does not merit the extra effort required for implementation of the relaxation algorithm. Efforts toward obtaining more appropriate test data are underway.

The residual statics algorithm may easily be adapted to the problem of frequency-dependent statics estimation. The simple statics model discussed in the previous section involves a phase shift that is linear with respect to frequency; a frequency-dependent model, however, allows more general variation in phase shifts. Frequency-dependent statics estimation is in many senses a surface-consistent *deconvolution.* Ideally, the source waveform and near-surface resonances would be deconvolved in a source- and receiver-consistent manner.

The frequency-dependent statics problem is notoriously difficult with models similar to (16) because phases must be "unwrapped" - the usual attempts at spectral decomposition have difficulties with phase shifts greater than $|\pi|$ (Sword, 1983). The optimization technique described here, however, does not encounter this difficulty. Adaptation of the algorithm is straightforward and requires just two elementary theorems from Fourier transform theory. The Rayleigh-Parseval theorem states that power in the time domain equals power in the frequency domain. So for a function $f(t)$ and its Fourier transform $F(\omega)$,

$$\sum_t |f(t)|^2 = \sum_\omega |F(\omega)|^2 .$$

In addition, the shift theorem states that time shifts in the time domain are equivalent to multiplication by a complex exponential in the frequency domain:

$$f(t-\tau) \supset e^{i\omega\tau} F(\omega) .$$

Then by letting the Fourier transform of $d_{yh}(t)$ be denoted by $D_{yh}(\omega)$, we may include frequency dependence in (18a) by writing

$$V_{s_i}[\mathbf{s}(\omega), \mathbf{r}(\omega)] = \sum_{y \in Y_{s_i}} \sum_\omega \left| \sum_h e^{i\omega[s_{i(y,h)}(\omega) + r_{j(y,h)}(\omega)]} D_{yh}(\omega) \right|^2 . \qquad (20)$$

Similar changes apply to equation (18b). Note that the s_i and r_j are now functions of ω.

There is a problem with this formulation that is related to the non-uniqueness of any residual statics solution. For the solution presented in the previous section, the longest spatial wavelength (the d.c. component) is fully undetermined by the data. In the

frequency-dependent case, the longest wavelength of each ω-component is also undetermined, but additional complexities arise because each component is independent of the others. Physically, however, the frequency-dependent phase shifts are expected to be locally correlated with each other to some degree. This *prior* knowledge may be incorporated in the potential function (19) by smoothing $\mathbf{s}(\omega)$ and $\mathbf{r}(\omega)$ over ω. Representing these smoothed functions by $\bar{\mathbf{s}}(\omega)$ and $\bar{\mathbf{r}}(\omega)$, the energy function for frequency-dependent statics is

$$E[\bar{\mathbf{s}}(\omega),\bar{\mathbf{r}}(\omega)] \;=\; -\sum_i V_{s_i}\,[\bar{\mathbf{s}}(\omega),\bar{\mathbf{r}}(\omega)] - \sum_j V_{r_j}[\bar{\mathbf{s}}(\omega),\bar{\mathbf{r}}(\omega)] \;.$$

Another candidate for inversion by stochastic relaxation is the problem of missing data; specifically, spatial interpolation of spatially aliased data. If there is only one aliased dip, then simple sinc interpolation may be locally applied along the direction of dip (Larner et al, 1980). But when two or more dips locally conflict and are aliased with respect to each other, then no straightforward interpolation operator is obvious. We have now truly a problem of estimation, and the Bayesian approach outlined here may be applicable. Again, the crucial element is the choice of potential function. A power-related potential could probably be effective for this problem, too. A correlation coefficient would be computed locally as a function of dip direction. The local potential would be the sum of the maximas in the correlation function, and energy would be the sum of these potentials. Each maximum would be included in the potential function so that dips of more than one direction could be detected. The neighborhood over which these calculations are made would encompass just a few traces and a few time samples.

The general problem of velocity inversion also may be approached much like residual statics estimation. A two-dimensional grid would be parameterized by velocity, and we would seek the velocity distribution yielding the maximum stack power. The problem is much more computationally intensive, however, because each perturbation and power calculation would necessitate far more effort than the simple shifts and sums needed in residual statics.

Remaining questions

Several important theoretical questions remain to be answered. Perhaps most importantly, much still needs to be learned with respect to the rate at which temperature is decreased. The notion of a *critical temperature* (Kinderman and Snell, 1980) in the theory of Markov random fields may offer some important answers. Above the critical temperature non-neighboring parameters are relatively independent of each other, but below this temperature the influence of non-neighboring parameters is strong. The statics tests exhibited

similar, abrupt behavior ("phase transitions") at certain temperature levels. Estimation of these critical temperatures in practical cases could greatly increase the efficiency of the relaxation algorithm.

Ideal potential functions may not be as simple as the ones described here. In particular, it may be advantageous to include spatial dependence in the function, so that different measures are computed in different locations (to account for variations in signal-to-noise ratio, spectral color, etc.).

This paper's approach is Bayesian in nature because a prior (Markov-Gibbs) probability model is specified. This inclusion of prior information is essential to the technique's success, but probably only a fraction of the available prior information has been exploited. We may reasonably expect that convergence will be faster and more accurate as more problem-dependent constraints are employed. For example, when the statics algorithm fails, it usually converges to a massive, obvious blunder. One way of protecting against these failures would be to construct a penalty function that detects unwanted dips (i.e., evanescent energy) on the output stack. This and most other types of prior information may be incorporated in the energy function. The more this is done, the more truly "Bayesian" the solution will be.

Much research and experimentation remain to be performed with regard to the important notion of *ergodicity*. Hammersley and Handscomb (1964) cast the Metropolis algorithm (the relaxation algorithm run at constant temperature) in terms of Markov chain theory. Ergodic averages (expected values) may be computed from successive realizations of a Markov chain if the chain is *irreducible*. A Markov chain is irreducible if all possible states are reachable from all other possible states. Thus the model $\mathbf{m}$ is irreducible if $P(\mathbf{m}) > 0$ for all $\mathbf{m}$ [or if $E(\mathbf{m}) < \infty$], which is the assumption made in equation (3b). If this is valid, then for constant $T = T_1$ we may compute the ergodic averages

$$\langle f(\mathbf{m}) \rangle = \sum_{\mathbf{m}} f(\mathbf{m}) P(\mathbf{M} = \mathbf{m}) = \frac{\sum_{\mathbf{m}} f(\mathbf{m}) e^{\frac{-E(\mathbf{m})}{T_1}}}{\sum_{\mathbf{m}} e^{\frac{-E(\mathbf{m})}{T_1}}},$$

where $\langle \bullet \rangle$ signifies expected value. The error in $\langle f(\mathbf{m}) \rangle$ drops off as $O(n^{-1/2})$, where n is incremented for each attempted perturbation. Calculation of these ergodic averages using the Metropolis technique is just a simple, uniformly weighted average of the Monte Carlo simulations at constant temperature. Thus the means, variances, covariances, etc. of the posterior probability distribution are easily obtained. We may even obtain an estimate of the posterior probability function by constructing a histogram. These are all important in the

solution of a geophysical inverse problem, because not only may we obtain simple answers (the maximum a posteriori solution) but we may also obtain estimates of resolution and accuracy. These latter quantities come directly from the posterior probability function. The posterior probability function is arguably the most fundamental information that a solution to an inverse problem can provide (Tarantola and Valette, 1982).

Whether or not the Markov chain is irreducible is unfortunately difficult prove. Irreducibility comes directly from assuming that $P(\mathbf{m}) > 0$ for all $\mathbf{m}$, but this may become questionable if the random perturbations are sampled from a distribution that is not uniform over the range of all possible values for M_{ij}. This question is important, because significant gains in computational efficiency may be possible by progressively narrowing the distribution from which perturbations are obtained.

Conclusions

This paper has presented the basis for further investigations of nonlinear inverse problems in reflection seismology. The prior assumption of a Markov random field is a good, simplifying model for many seismic data processing problems, and the Markov-Gibbs equivalence leads to the application of a powerful optimization technique designed to attack the problem of local extrema.

The preliminary application of these ideas to the problem of residual statics estimation has been successful. Tests were performed on synthetic data, but the data presented difficulties usually unseen in field datasets. Application to field data is the next goal in the continuation of these efforts.

Future applications to frequency-dependent statics estimation, missing data restoration, and other inverse problems are envisioned. The necessary assumption is that parameters are locally interactive and define a Markov random field. Viewed from this perspective, the difficult, nonlinear inverse problems in reflection seismology may gain new and valuable solutions.

ACKNOWLEDGMENTS

I thank Fabio Rocca for encouraging me to pursue a probabilistic approach to residual statics estimation. Jon Claerbout suggested the power criterion. I had many helpful discussions with Francis Muir, John Toldi, Stew Levin, Shuki Ronen, and most other members of the SEP. Shuki also generously provided some helpful technical assistance. Western

Geophysical Company, via Helmut Jakubowicz and others, provided some data used for early tests; unfortunately the data proved unsuitable for illustration. Finally, I thank Stuart Geman for providing a preprint of his unpublished manuscript.

REFERENCES

Aki, K. and Richards, P., Quantitative Seismology: San Francisco, W. H. Freeman and Co., 932 p.

Donoho, D.L., 1979, Estimation of time delay at poor S/N: Paper presented at the 1979 EAEG, Hamburg.

Geman, S. and Geman, D., 1983, Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images: preprint.

Hammersley, J.M. and Handscomb, D.C., 1964, Monte Carlo methods: London, Chapman and Hall, 178 p.

Kinderman, R. and Snell, J.L., 1980, Markov random fields and their applications: Providence, American Mathematical Society, 142 p.

Kirkpatrick, S., Gelatt, C.D., Jr., and Vecchi, M.P., 1983, Optimization by simulated annealing: Science, v. 220, p. 671-680.

Larner, K.L., Gibson, B., and Rothman, D., 1980, Trace interpolation and the design of seismic surveys: Paper presented at the 1980 SEG, Houston.

Metropolis, N., Rosenbluth, A., Rosenbluth, M., Teller, A., and Teller, E., 1953, Equation of state calculations by fast computing machines: Journal of Chemical Physics, v. 21, p. 1087-1092.

Parker, R.L., 1977, Understanding inverse theory: Ann. Rev. Earth Planet. Sci., v. 5, p. 35-64.

Reif, F., 1965, Fundamentals of statistical and thermal physics: San Francisco, McGraw-Hill Book Company, 651 p.

Rothman, D., 1980, Probabilistic residual statics: SEP-37, p. 152-156.

Shore, J.E. and Johnson, R.W., Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy: IEEE Trans. Inform. Theory, v. IT-26, p. 26-37.

Sword, C., 1983, The generalized frequency-dependent surface-consistent problem: SEP-35, p. 19-42..

Taner, M.T., Koehler, F., and Alhilali, K.A., 1974, Estimation and correction of near-surface time anomalies: Geophysics, v. 39, p. 441-463.

Tarantola, A. and Valette, B., 1982, Inverse problems = Quest for information: Journal of Geophysics, v. 50, p. 159-170.

Wiggins, R., Larner, K., and Wisecup, D., Residual statics as a general linear inverse problem: Geophysics, v. 41, p. 922-938.