Data and model uncertainty estimation for linear inversion

Kasper van Wijk*, John A. Scales*, William Navidi† and Luis Tenorio†

* *Department of Geophysics and Center for Wave Phenomena, Colorado School of Mines.*
† *Department of Mathematics, Colorado School of Mines.*

ABSTRACT

Inverse theory concerns the problem of making inferences about physical systems from indirect noisy measurements. Information about the errors in the observations is essential to solve any inverse problem, otherwise it is impossible to say when a feature "fits the data." In practice, however, one seldom has a direct estimate of the data errors. We exploit the trade-off between data prediction and model or data structure to determine both model-independent and model-based estimates of the noise characteristics from a single realization of the data. Noise estimates are then used to characterize the set of reasonable models that fit the data, for example, by intersecting prior model parameter constraints with the set of data fitting models. This prior information can also be used to set bounds on the bias. We illustrate our methods with synthetic examples of vertical seismic profiling and cross-well tomography.

1 INTRODUCTION

The goal of geophysical inversion is to make quantitative inferences about the Earth from a finite number of indirect and noisy observations. Information about the data uncertainties is required to determine what Earth models are consistent with the observations. In practice the issue of data uncertainty is often ignored, with the problem of data fitting being replaced by the optimization of a data misfit function. Of course, this begs the question of when to terminate the optimization. Beyond this, it is common to simply assert a noise variance *a priori*.

With multiple realizations of the experiment, the distribution of random fluctuations could be directly quantified, but in geophysics this is seldom the case. If the data are sufficiently redundant they can be binned so as to approximate the situation of multiple realizations (Van Wijk et al., 1998). A clever variation of this theme exploits ray path redundancy in refraction tomography (Docherty, 1992). If the (binned) data are not truly redundant it is necessary to model their systematic variations. These variations can be first modeled without solving the inverse problem. In principle the estimation of the data uncertainties begins with an analysis of the error budget of the experiment, along with background noise recordings; for example see (Gouveia & Scales, 1998). The error budget for any geophysical field experiment is quite complicated and also involves many unknowns. Our goal here is to develop techniques which, if not as objective as an error budget analysis, will be easier to apply and nevertheless provide reasonable results.

We propose the following practical algorithm for estimating data uncertainties and checking model fits in geophysical inverse problems. First, we use a nonparametric regression method based on Tikhonov regularization to obtain a model-independent estimate of the data uncertainties. These error estimates may be used to determine sets of data fitting models that also satisfy other prior constraints. Alternatively, a second Tikhonov regularization can be used to estimate an Earth model. But without information of the data uncertainties, we do not know if the model provides an adequate fit to the data, and thus it cannot be regarded as a solution of the inverse problem. However, the residuals of this model also provide estimates of the data uncertainties that can be compared to the model-independent estimates to check goodness of fit. Once we have a reasonable Earth model we use prior information to construct bias-corrected confidence intervals for the true model.

2 THE INVERSE PROBLEM

In inverse theory a *model* is a mathematical parameterization of those properties of a physical system that are required to predict the data. In the Earth sciences,

models are usually functions of space and are therefore elements of an infinite dimensional space, known as the *model space*, a generic element of which we denote by x. In contrast, an experiment always results in a finite number of observations $\mathbf{d}$. *Data prediction* involves mapping elements from the model space into the data space. In addition, the data are contaminated by random errors $\mathbf{e}$ and systematic errors $\mathbf{s}(\mathbf{x})$:

$$\mathbf{d} = \mathbf{g}(x_{\mathrm{T}}) + \mathbf{e} + \mathbf{s}(x_{\mathrm{T}}), \tag{1}$$

where $\mathbf{g}$ is the forward modeling operator and x_{T} is the true model. Among the many contributors to the uncertainty in the data, the systematic errors consist primarily of un-modeled physics and effects of model discretization. The random errors are, by definition, those variations in the data that are not deterministically reproducible. A practical definition of noise is: that part of the data which we choose not to fit (Scales & Snieder, 1998).

In practice, models are often approximated by discretizing to a finite dimensional vector of parameters. Choosing too coarse a discretization may result in discretization artifacts and a limited ability to fit the data. Another, more subtle, consequence of discretization is that uncertainty estimates may be artificially small (Stark, 1992a). In Section 5 we adopt a practical approach by comparing the fits of different discretizations with model-independent fits to the data.

If the forward operator is linear, (1) reduces to

$$\mathbf{d} = \mathbf{A}\mathbf{x}_{\mathrm{T}} + \mathbf{e} + \mathbf{s}(\mathbf{x}_{\mathrm{T}}), \tag{2}$$

where $\mathbf{d} \in R^n$, the discrete version of the true model is $\mathbf{x}_{\mathrm{T}} \in R^m$, and the discretized forward operator is $\mathbf{A} \in R^{n \times m}$. From this point on, we ignore the systematic error term $\mathbf{s}(\mathbf{x}_{\mathrm{T}})$, assuming the forward modeling operator is accurate. The goal is then to find models that fit the systematic variations in the data. In practice the separation between systematic and random variations is not easy to make. Certainly, if the errors in the data are correlated and overlap the bandwidth of the data, it is impossible to separate the noise from the signal.

If the errors $\mathbf{e}$ are uncorrelated and have known statistical characteristics we could define sets of data-fitting models in a variety of ways. For example, a criterion that is useful for independent, normally-distributed errors is the normalized χ^2:

$$\chi^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\sum_j A_{ij} x_j - d_i}{\sigma_i} \right)^2, \tag{3}$$

where σ_i is the standard deviation of the ith datum. For example, any model $\mathbf{x}$ such that $\chi^2(\mathbf{x}) < 1$ fits the data on average to one standard deviation. But in practice how do we determine σ_i? In many published examples of inverse calculations involving real data the issue is sidestepped by either making a particular *a priori* choice of the data errors [e.g., Scales (1987), Oldenburgh et al.(1997)] or avoiding the problem altogether by simply optimizing the difference between the observed and predicted data. Obviously the latter strategy begs the question of when to terminate the optimization. Once again various *rules of thumb* are used (e.g., 95% variance reduction). These approaches are unsatisfactory, since subtle, apparently inconsequential choices can lead to misinterpretations of the data (Scales & Tenorio, 2001). Equation (3) can be readily generalized to l_p norm measures of data misfit. For $p \approx 1$, the l_p norm is robust in the presence of long-tailed noise (Scales et al., 1988).

In recent years methods such as Tikhonov regularization, with the regularization parameter chosen by the L-curve or generalized cross validation (GCV), have become popular methods (Li & Oldenburg, 1999) to find stable solutions of ill-posed inverse problems. Estimates of the noise variance are based on the residual sum of squares of the fitted model. In this paper we concentrate on this last point: understanding the random fluctuations in the data. We use Tikhonov regularization methods but other approaches such as truncated SVD or conjugate gradient could also be used to regularize the optimization process.

3 ESTIMATING DATA UNCERTAINTIES

Estimating the variance σ_i of the ith datum d_i means estimating the variability of d_i about its mean

$$\mu_i = (\mathbf{A}\mathbf{x}_{\mathrm{T}})_i.$$

Let us assume that the variance is constant $\sigma_i^2 = \sigma^2$. We consider the following two complementary ways of estimating σ^2:

(i) Tikhonov regularization to obtain an estimate $\widehat{\boldsymbol{\mu}}_\lambda$ of $\boldsymbol{\mu}$ by minimizing

$$\min_{\boldsymbol{\mu}} \left(\|\boldsymbol{\mu} - \mathbf{d}\|^2 + \lambda \|\mathbf{R}\boldsymbol{\mu}\|^2 \right), \tag{4}$$

for some operator $\mathbf{R}$. The variance of the residual vector $\widehat{\boldsymbol{\mu}}_\lambda - \mathbf{d}$ is used as an estimate of σ^2. This is a model-independent estimate because it does not require the solution of the inverse problem. Also, since it is independent of the data mapping, (4) can be applied to problems with a nonlinear forward operator. If the variance is not constant we can use a sliding window along the residual vector (Van Wijk et al., 1998).

(ii) Tikhonov regularization to obtain a model estimate $\widehat{\mathbf{x}}_\lambda$ by minimizing

$$\min_{\mathbf{x}} \left(\|\mathbf{A}\mathbf{x} - \mathbf{d}\|^2 + \lambda \|\mathbf{R}\mathbf{x}\|^2 \right). \tag{5}$$

We set $\widehat{\boldsymbol{\mu}}_\lambda = \mathbf{A}\widehat{\mathbf{x}}_\lambda$ and estimate σ^2 as in (i).

Both methods require the selection of a regularization operator $\mathbf{R}$ and a regularization parameter λ. In either (4) or (5), $\mathbf{R}$ is used to damp (if $\mathbf{R}$ is the identity matrix) or penalize roughness (if $\mathbf{R}$ is a derivative operator) of $\boldsymbol{\mu}$ or $\mathbf{x}$. For the latter, $\mathbf{R}$ also has to regularize $\mathbf{A}$. This

means that the the null spaces of $\mathbf{R}$ and $\mathbf{A}$ cannot overlap.

The regularization parameter λ determines the trade-off between the data misfit and a model or data structure penalty term [e.g. Tikhonov & Arsenin (1977), Green & Silverman (1994)]. The L-curve method optimizes that trade-off as a function of the regularization parameter λ (Lawson & Hanson, 1974). On a double logarithmic plot this curve often takes the shape of an "L". The knee in the curve is defined as the optimal λ. GCV could also be used to determine λ but, for the problems we have studied, GCV tends to significantly overestimates the value of λ. The minimum of the GCV auxiliary function of λ tends to be very broad, leading to numerical instability of the GCV estimate.

Good estimates of $\mathbf{x}_\mathrm{T}$ and $\boldsymbol{\mu}$ should lead, provided there are no systematic errors, to residuals that approximate the true errors $\mathbf{e}$. Therefore there should not be a substantial difference between the model-independent and the model-based variance estimates. But without prior information on the errors how can we tell if an estimate is "good"? This makes the argument somewhat circular but, in the absence of *a priori* information on the noise, *some* assumptions must be made. These assumptions may be based on the physics and the properties of the forward operator. For example, if $\mathbf{A}$ is the discretization of an integral operator that smooths the model, we expect $\boldsymbol{\mu}$ to be the discretization of a function that is at least as smooth as the true model x_T. In this case a smoothness constraint on $\boldsymbol{\mu}$ is easier to justify. This smoothness also results in estimates of σ^2 that are better than the model-based estimates. The reason is the following. Methods like GCV and the L-curve determine a λ that balances the bias and variance of the regularization estimates. This results in biases that are larger precisely where the model has the most interesting structure (Cummins et al., 2001). The less structure in the model, the smaller the bias in the variance estimates.

4 SOLVING THE INVERSE PROBLEM

Until now we have focused on estimating the variance of random data errors. Once these uncertainties are known, the inference problem (characterizing the range of models that fit the data) can be tackled. Our purpose here is not to treat the inference problem exhaustively, but simply to illustrate several approaches that could be taken. A key ingredient in any inverse problem is prior model information. Since the data mapping operator is not invertible, it is usually impossible to achieve finite uncertainty on the model parameters without prior information. In practice, prior information on the true model may come, for example, from geologic information on the correlation length of layering in the region or from well-log measurements (Gouveia & Scales, 1998).

4.1 Surgery

With sufficient prior information it may be possible to construct empirical or theoretical Bayesian probability distributions characterizing the range of feasible models [e.g., Scales & Tenorio (2001), Gouveia & Scales (1998), Moraes & Scales (2000)]. However, in this paper we focus on frequentist methods of inference. We use deterministic prior information to reduce the range of data fitting models. For example, following Stark (1992b) [see also Evans & Stark (2001)], let:

$$\mathbf{d} = \mathbf{A}\mathbf{x}_\mathrm{T} + \mathbf{e}.$$

Assuming that the noise has been characterized, as in Section 3, we can find a $1-\alpha$ confidence set $\Xi \in R^n$ for the data errors; i.e.,

$$P\left[\,\mathbf{e} \in \Xi\,\right] \geq 1-\alpha.$$

Now let D be the set of models $\mathbf{x}$ such that $\mathbf{d} - \mathbf{A}\mathbf{x} \in \Xi$. Thus D is the *preimage* of Ξ under the action of the forward operator. It then follows that D is a $1-\alpha$ confidence region for the model $\mathbf{x}_\mathrm{T}$,

$$P\left[\,\mathbf{x}_\mathrm{T} \in D\,\right] \geq 1-\alpha.$$

Now suppose we are certain that $\mathbf{x}_\mathrm{T}$ is in some set C. This is an example of deterministic prior information. Then $C \cap D$ is a $1-\alpha$ confidence set for $\mathbf{x}_\mathrm{T}$. Since any element of $C \cap D$ might be the truth, we can try to characterize all the models in this set—this is the *inference problem.*

Conceptually this approach is straightforward: we first find the set of data fitting models, then we perform surgery on this set by intersecting it with a prior constraint set. This gives us a (presumably) smaller set of models that fit the data and are *a priori* feasible.

4.2 Confidence intervals and bias corrections

The surgical approach above did not require a particular estimated model. We now use an estimator obtained by Tikhonov regularization to construct bias-corrected confidence intervals for model parameters.

To construct a confidence interval from the value of an estimator, we must determine the bias and variance of the estimator. An estimator can be stable to random perturbations of the data (small variance) but be far from the truth (biased). On the other hand, we can have an unbiased estimator that is very sensitive to data perturbations (large variance).

Let $\mathbf{A}_\lambda^\dagger$ be the regularized pseudo-inverse for (5) associated with a fixed λ, and $\widehat{\mathbf{x}}_\lambda$ the regularized solution

$$\widehat{\mathbf{x}}_\lambda = \mathbf{A}_\lambda^\dagger \mathbf{d} = \left(\mathbf{A}^T\mathbf{A} + \lambda \mathbf{R}^T\mathbf{R}\right)^{-1}\mathbf{A}^T\mathbf{d}, \tag{6}$$

where $\mathbf{R}$ is a regularizing operator that damps or smooths the solution. The covariance matrix of $\widehat{\mathbf{x}}_\lambda$ is

$$\mathrm{cov}(\widehat{\mathbf{x}}_\lambda) = \mathbf{A}_\lambda^\dagger\, \mathrm{cov}(\mathbf{d})\, \mathbf{A}_\lambda^{\dagger\,T}. \tag{7}$$

Assuming that the covariance of the data is $\sigma^2\mathbf{I}$, the variance of the ith model parameter is

$$\mathrm{Var}\left[\,(\widehat{\mathbf{x}}_\lambda)_i\,\right] = \sigma^2 \left(\mathbf{A}_\lambda^\dagger \mathbf{A}_\lambda^{\dagger\,T}\right)_{ii}. \tag{8}$$

To construct $1-\alpha$ confidence intervals for model parameters we use the Gaussian approximation

$$(\widehat{\mathbf{x}}_\lambda)_i \pm z_{\alpha/2}\,\widehat{\sigma}\sqrt{\left(\mathbf{A}_\lambda^\dagger \mathbf{A}_\lambda^{\dagger\,T}\right)_{ii}}, \tag{9}$$

where $z_{\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Gaussian distribution, $\widehat{\sigma}^2$ is the variance estimate obtained in Section 3 and λ is determined by the L-curve. This interval could be shifted by a bias in the model estimate given by

$$\mathrm{Bias}(\widehat{\mathbf{x}}_\lambda) \equiv \mathrm{E}\,(\,\widehat{\mathbf{x}}_\lambda - \mathbf{x}_\mathrm{T}\,) = \mathbf{C}\mathbf{R}\,\mathbf{x}_\mathrm{T} = \mathbf{B}\,\mathbf{x}_\mathrm{T}, \tag{10}$$

where

$$\mathbf{C} = -\lambda\left(\,\mathbf{A}^T\mathbf{A} + \lambda\mathbf{R}^T\mathbf{R}\,\right)^{-1}\mathbf{R}^T \quad \text{and} \quad \mathbf{B} = \mathbf{C}\mathbf{R}.$$

$\mathbf{B}$ can also be written in terms of the resolution matrix $\mathbf{A}_\lambda^\dagger\mathbf{A}$: $\mathbf{B} = \mathbf{A}_\lambda^\dagger\mathbf{A} - \mathbf{I}$. A good resolution of the estimator implies a small bias, and vice versa.

In general the bias cannot be computed exactly because it depends on the true model. We now discuss some ideas on how prior information can be used to bound the bias of model estimates [see Xu (1998) for related work]. Such information may include bounds on the norm and/or the smoothness of the model. To bound the bias just note that

$$\begin{aligned} ||\mathrm{Bias}(\widehat{\mathbf{x}}_\lambda)|| &= ||\mathbf{C}\mathbf{R}\mathbf{x}_\mathrm{T}|| \le ||\mathbf{C}||\,||\mathbf{R}\mathbf{x}_\mathrm{T}|| \\ &\le ||\mathbf{C}||\,||\mathbf{R}||\,||\mathbf{x}_\mathrm{T}||\,, \end{aligned} \tag{11}$$

where $\mathbf{C}$ and $\mathbf{R}$ are known matrices. We can bound the bias with prior information on $\|\,\mathbf{x}_\mathrm{T}\,\|$ or $\|\,\mathbf{R}\mathbf{x}_\mathrm{T}\,\|$. See O'Sullivan (1986) and Tenorio (2001) for further discussions on bias assessment in more general regularization methods.

Note that prior information on $||\mathbf{R}\mathbf{x}_\mathrm{T}||$ may provide significantly tighter bounds on the bias than prior information on $||\mathbf{x}_\mathrm{T}||$ since we can choose $\mathbf{R}$ according to our knowledge of $\mathbf{x}$. For example, if $\mathbf{x}$ is known *a priori* to be smooth, we may want to penalize roughness using a second derivative operator $\mathbf{R}$ to make $||\mathbf{R}\mathbf{x}||$ small. The bias is zero if $\mathbf{x}$ is in the null space of $\mathbf{R}$.

4.2.1 Convex constraint set

In the previous section, we considered prior information on model norms. Now suppose that prior information on individual model parameters is available. Each component of the bias is a linear functional of the true model:

$$\left[\,\mathrm{Bias}(\widehat{\mathbf{x}}_\lambda)\,\right]_i = \mathbf{b}_i^T\mathbf{x}_\mathrm{T}, \tag{12}$$

where $\mathbf{b}_i$ is the ith row of $\mathbf{B}$. If we assume *a priori* that each component of the true model belongs to an interval C of the form

$$C : \mathbf{l}_i \le (\mathbf{x}_\mathrm{T})_i \le \mathbf{u}_i, \tag{13}$$

then we can find the maximum and minimum bias by solving linear programming problems for each component:

$$\max_{\mathbf{x}_\mathrm{T}\in C} \mathbf{b}_i^T\mathbf{x}_\mathrm{T}, \quad \min_{\mathbf{x}_\mathrm{T}\in C} \mathbf{b}_i^T\mathbf{x}_\mathrm{T} \qquad i = 1, 2, ..., m. \tag{14}$$

Once we have found the vectors that solve the optimization for each bias component, the values are used to bound the bias of the estimated model parameter. Admittedly, this will be a pessimistic estimate, but our goal is to be conservative in our interpretation of the data.

Additional prior information on the smoothness of the model parameters can be incorporated as a constraint in (14), for example,

$$\max_{\mathbf{x}_\mathrm{T}\in C} \mathbf{b}_i^T\mathbf{x}_\mathrm{T} \;\text{ with }\; \sum_j L_{ij}\mathbf{x}_{\mathrm{T}j} < v_i \quad i = 1, 2, ..., m, \tag{15}$$

for some vector of derivative bounds $\mathbf{v} = (v_i)$, where $\mathbf{L}$ is a second order difference operator.

In our examples we have assumed a linear data mapping. The confidence interval analysis can be applied approximately to mildly nonlinear nonlinear problems (i.e., those amenable to iterative linearization). For example, see Bates & Watts (1998).

5 1-D TRAVEL TIME TOMOGRAPHY

We investigate the discretized one-dimensional tomography problem of vertical radar or seismic profiling (VRP/VSP). The geometry is shown in Figure 1. A single source of elastic or electromagnetic energy is on the surface directly above a number of receivers that are lowered into a borehole. The source emits a pulse of energy and the arrival times to n receivers form the data vector $\mathbf{d}$. The model vector $\mathbf{x}$ consists of m layers of equal thickness. The slowness is constant in each layer. The travel time t along a single ray is the integral of the local *slowness* s (i.e., the reciprocal of velocity v, along the ray):

$$t = \int_{\mathrm{ray}} \frac{1}{v}\,dl = \int_{\mathrm{ray}} s\,dl. \tag{16}$$

The discrete forward operator is a $n\times m$ matrix $\mathbf{A}$ whose element A_{ij} is the length of the ith ray in the jth layer. For non-zero offsets, expression 16 is nonlinear in s due to refraction of the rays. In practice this nonlinearity is mild and can be overcome by iterative linearization.

The true model has decreasing slowness down to a depth of 40m with an isolated high slowness zone centered at 20m depth (see Figure 2). To simulate a truly (piece-wise) continuous model, the synthetic data are computed with a 1000 layer approximation of this model. The right side of Figure 2 shows the exact data contaminated by uncorrelated Gaussian pseudo-random

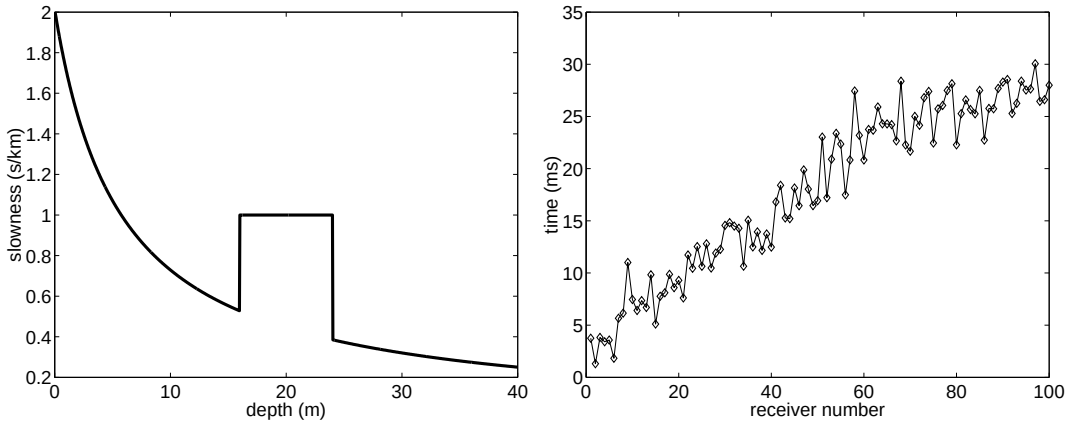

Figure 2. The true slowness model (left) and synthetic data associated with this model (right). The data are contaminated with uncorrelated noise drawn from a Gaussian distribution with $\mu = 0$ ms and $\sigma = 2$ ms.

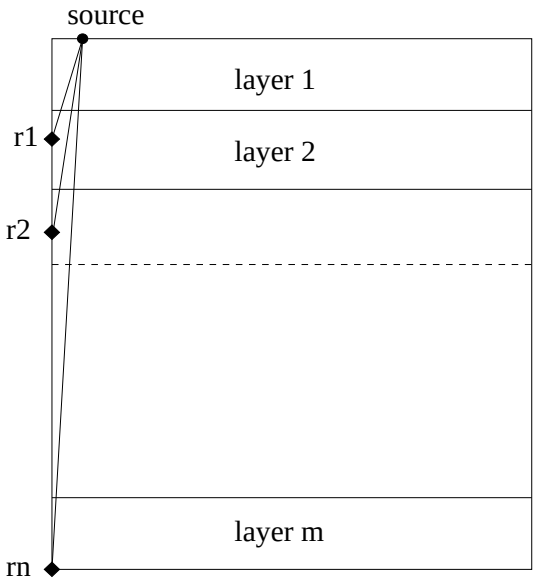

Figure 1. Geometry of the synthetic VSP experiment. The source is at zero offset but for visual purposes is drawn at a small offset.

noise of mean $\mu = 0$ ms and standard deviation $\sigma = 2$ ms.

5.1 Estimating the noise variance

From here on subscripts **I** and **L** stand, respectively, for Tikhonov estimates that use the L-curve with the identity or a discrete derivative operator as penalty matrix. In addition, we use a variation on the L-curve method that uses **L** but is based on a plot of the logarithmic residual as a function of $\log(1/\lambda)$. We denote these results with the subscript $1/\lambda$. This variation is to point out that there are different possible quantities to investigate model structure. The nonparametric estimate of the data uncertainty will be sub-scripted by μ.

Model-based estimates of σ were obtained using the three regularization schemes (R_I-curve, R_L-curve and $R_{1/\lambda}$-curve). Figure 3 shows a typical R_L-curve for one realization of the noisy data. Approximate 95% confidence intervals for σ are given in Table 1.

To get model independent estimates of the noise variance, we apply the regularization scheme described in equation (4) with a second difference regularization operator to 100 realizations of the noisy data to find $\boldsymbol{\mu}$ that predicts the bulk of the systematic variations in the data and leads to the following 95% confidence interval for σ: 2.02 ± 0.03 ms.

We see that model-based estimates have a larger bias than those from the model-independent methods, but they are not too far off from the true value. We use the model-independent estimate to approximate the variance of each model parameter estimate with (8), which is in turn used to construct confidence intervals for the model $\mathbf{x}_T$ in Figure 4. These intervals may be biased by nonzero terms in (10). The bias could be reduced by using a local cross-validation method that has been recently developed by Cummins et al. (2001). But we would still want to estimate the leftover bias to correct the confidence intervals.

Note that the model variances in Figure 4 are relatively large near the surface and at the maximum depth of the model. Near the surface $\widehat{\mathbf{x}}_\lambda$ is sensitive to the random errors in the data, because there are several layers in the model before the first receiver is reached; near the bottom this is caused by poor ray coverage of the deepest layers.

5.2 The bias

Figure 5 shows corrected confidence intervals obtained by subtracting the exact bias, as defined by (10), from

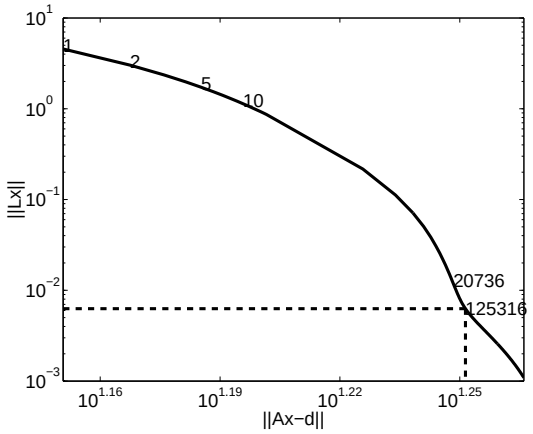

Figure 3. Example of an R_L-curve for the VSP data. The optimal value of the regularization parameter λ_L is chosen to be the point of maximum curvature.

$\widehat{\sigma}_\mu$	$\widehat{\sigma}_I$	$\widehat{\sigma}_L$	$\widehat{\sigma}_{1/\lambda}$
2.02 ± 0.03	1.90 ± 0.03	1.92 ± 0.03	1.93 ± 0.03

Table 1. Approximate 95% confidence intervals (in ms) for the true standard deviation $\sigma = 2.0$ ms of the VSP data. The first column corresponds to the model-independent estimate, the others are model-based estimates from the three different L-curves.

the model estimate $\widehat{\mathbf{x_L}}$. The results are now consistent with the true model. Of course, the exact bias is a function of the true unknown model, but we will show next how to use prior information on the true model to put bounds on the bias.

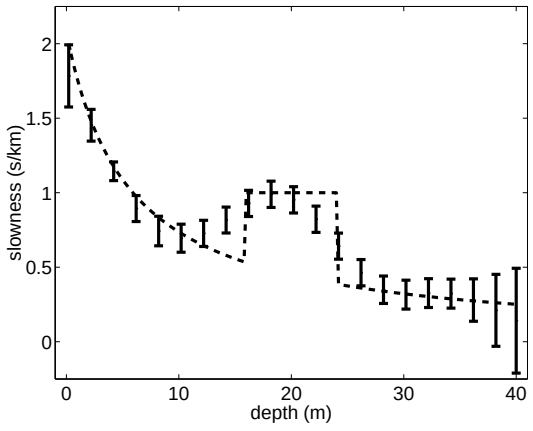

Figure 4. 1σ confidence intervals (9) for the VSP model. Every tenth model parameter interval is plotted. The intervals show a trend coming from nonzero terms in the bias (10).

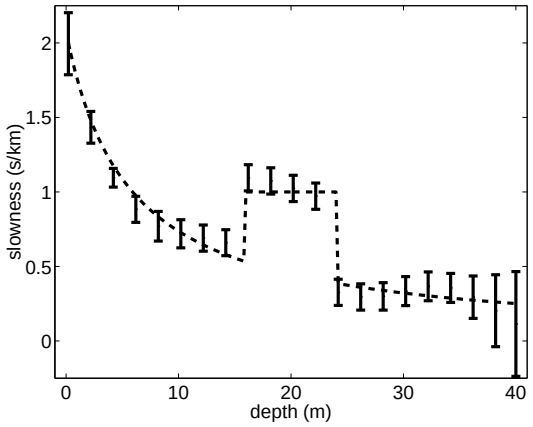

Figure 5. Same confidence intervals as in Figure 4 but corrected by subtracting the exact bias from the estimator $\widehat{\mathbf{x}_L}$.

5.3 Bias corrections

First, we assume prior information in the form of upper and lower bounds on the slowness model, only (in s/km):

$$2 > s_i > 0 \quad \forall \text{ i}. \tag{17}$$

These constraints on the model parameters are not enough to provide bounds for the bias or to narrow the prior bounds with the *surgery* approach as described in Section 4.1. A simple explanation of this is the following. Since there are more layers than observations in this 1-D experiment, it is possible to have highly fluctuating slownesses in the layers between two successive receivers and still have the average travel time fit the data. Therefore, without additional prior information on the smoothness of the model, uncertainty estimates based on bias upper bounds cannot be smaller than the prior bounds and are still given by (13). However, we can narrow the range of possible model parameter values that fit the data using information on the smoothness of the true model. For this example we used $(\mathbf{L}\mathbf{x}_\mathrm{T})_i < 0.001$ s/km^3 for all i. In fact, this is true everywhere, except at the discontinuity of the high slowness area.

The confidence sets (Figure 6) are the model parameter variance corrected by the minimum and maximum bias from expression (15). Finally, to be consistent with our prior assumptions, the confidence sets are bounded by the assumed upper and lower limits as defined in (17).

5.4 Discretization Effects

Geophysical inverse problems differ from parameter estimation problems for the simple reason that the models to be estimated are infinite dimensional. Ignoring this fact and simply choosing an *ad hoc* discretization can result in both discretization artifacts and optimistic error estimates. This issue is thoroughly explored by Stark

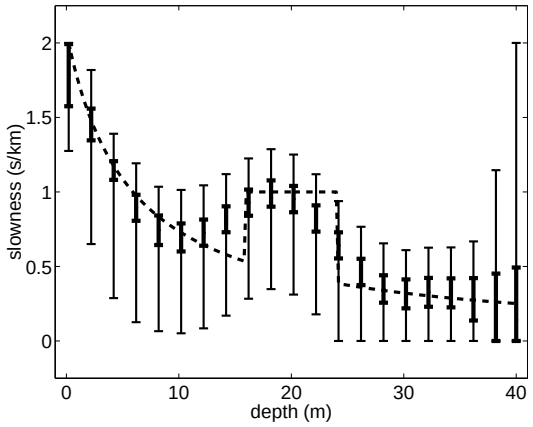

Figure 6. Bias corrections for the confidence intervals in Figure 4. The thick lines correspond to the original biased intervals. The upper and lower limits of the thin lines are computed by adding (subtracting) the upper (lower) bound of the bias (see equation 14), intersected with the hard bounds of (17). The bias corrections are based on prior knowledge on the model.

m	10	20	50	100	200	300
$\widehat{\sigma}_I$	1.92	1.87	1.87	1.89	1.90	1.89
$\widehat{\sigma}_L$	1.91	1.86	1.90	1.90	1.92	1.89
$\widehat{\sigma}_{1/\lambda}$	1.92	1.88	1.91	1.92	1.93	1.91

Table 2. Standard deviation estimates (ms) for the three different L-curves as a function of discretization for 100 realizations. The true standard deviation is $\sigma = 2$ ms, whereas the nonparametric estimate $\widehat{\sigma}_\mu = 2.02$ ms.

(1992a) who shows how to use the theory of conjugate duality to put bounds on discretization errors of continuous inverse problems. We adopted a much simpler strategy by comparing the standard deviation estimates for up to 300 layers. Table 2 shows that the error estimation is reasonably robust to discretization. By comparing model-independent and model-based variance estimates we see that 10 layers are enough to explain the variations in the data. A selection of a higher number of layers has to be justified with prior information about the unknown Earth model.

5.5 Concluding the VSP example

All proposed error estimation methods performed within 5% of the truth, with the model-independent estimate being slightly superior in accuracy. The latter does not involve model discretization, which eliminates one more subjective choice and improves numerical efficiency. With the error estimate we compute model parameter variances. These variances, combined with bias corrections from prior information on the true model, lead to confidence intervals that are in agreement with the exact solution.

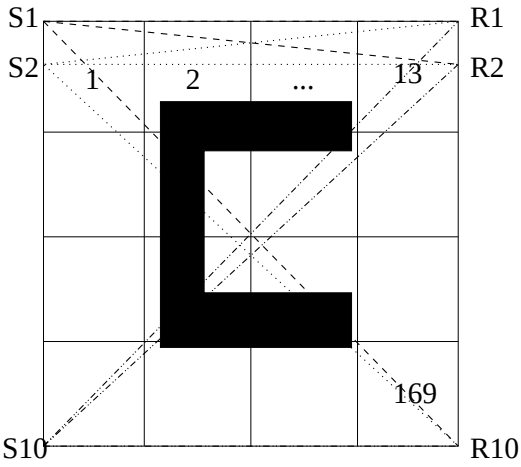

Figure 7. The true model and the shot/receiver geometry of the synthetic tomography problem. The "C" has a slowness of the 0.5 s/km, whereas the background slowness is 1 s/km.

6 CROSS-WELL TOMOGRAPHY

In the VSP example above all error estimation procedures provided reasonable estimates. However, as the next example shows, there are no guarantees that these methods work generally.

The true model, shown with the shot/receiver geometry in Figure 7, is a 169 cell model with a C-shaped structure centered between two boreholes. The background slowness is 1 s/km and the "C" has slowness 0.5 s/km. In both, the forward operation and the inverse calculation, we assume straight rays, which makes this problem linear. The exact travel times from 10 receivers and 10 shots are contaminated by Gaussian noise with mean zero and standard deviation $\sigma = 0.52$ ms.

We can get a model-independent noise estimate by smoothing the travel times in any parameterization we want. The best, specially when using a roughness penalty approach, is to use one which leads to a smooth structure that can be easily estimated. For example, we can apply a similar method used in the VSP example to each of the 10 sources, and average the 10 variance estimates. This gives the estimate $\widehat{\sigma}_\mu = 0.57$ ms. If, on the other hand, we use travel time ordered by path length, we get $\widehat{\sigma}_\mu = 0.88$ ms. These two representations of the data are depicted in Figure 8. In either case we have penalized roughness with a second derivative but it is not clear that this is a reasonable constraint on the $\boldsymbol{\mu}_i$ for either of the two orders. In the VSP case, there was a natural way of presenting the data from the shallow to the deep receivers. However, in this 2-D example there is not such a an obviously natural representation of the data. In cases like this, it may be better to obtain

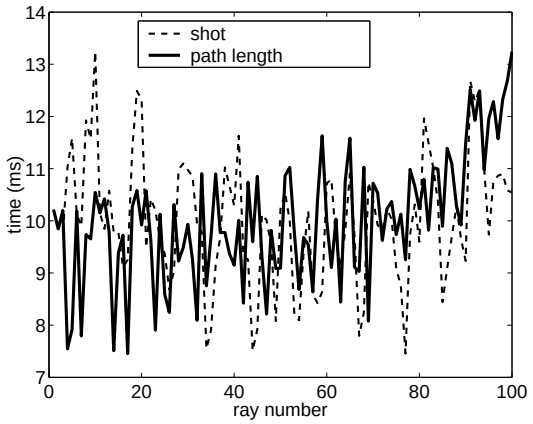

Figure 8. The travel times ordered by shot and ray path length. Note that the representation by path length is smoother than the data ordered shot by shot. This has consequences for the way we use the L-curve on the nonparametric error estimate.

σ	$\widehat{\sigma}_I$	$\widehat{\sigma}_L$	$\widehat{\sigma}_{1/\lambda}$	$\widehat{\sigma}_\mu$
0.52	–	0.50	0.62	0.88

Table 3. Estimates of the noise standard deviation (in ms) from the three algorithms for a cross-well tomography problem.

model-independent estimates without using a roughness penalty approach, for example, by thresholding coefficients of the data in a wavelet representation (Donoho & Johnstone, 1995).

Table 3 contains the error estimates for the different algorithms. Especially the R_L-curve, can separate the signal from the noise. It seems that penalizing the roughness of the model is a physically natural approach. The R_I-curve method fails since there is no "L," and thus does not offer an error estimate.

6.1 Concluding the tomography example

In this 2-D example, a representation of the data that corresponds to a smooth representation of $\boldsymbol{\mu}$ is not as apparent as in the VSP case. Finding such representation requires more physical information. Other nonparametric regression methods can be used to obtain noise variance estimates that do not implicitly depend on roughness constraints [e.g., Green & Silverman (1994)]. We have not done this here. However, penalizing the roughness of the model as in the R_L-curve seems natural and provides an accurate estimate of the noise.

In this example we merely want to tackle the error estimation procedure. An analog analysis to the VSP problem on the confidence intervals on the model parameters could be performed as well.

7 CONCLUSIONS AND DISCUSSION

In geophysics we are often faced with the challenge of estimating Earth model parameters given a single set of noisy observations, without information on the noise. At some point we must make subjective choices, in order to proceed quantitatively. These choices can be made on the model $\mathbf{x}_\mathrm{T}$ or on the mapped model $\boldsymbol{\mu} = \mathbf{A}\mathbf{x}_\mathrm{T}$, but, depending on the type of forward operator, smoothness assumptions may be easier to justify, or may be less compromising, for $\boldsymbol{\mu}$.

We have used the L-curve to obtain regularized estimates of $\boldsymbol{\mu}$ and $\mathbf{x}_\mathrm{T}$. It stands to reason that 'good' estimates of each should lead to residuals with similar characteristics. A comparison of these residuals can therefore be used as a goodness of fit check.

The L-curve can often tell us when significantly more structure is needed to improve the data prediction. but there are no guarantees that the L-curve method, or any other method, will always work with a given regularization operator. We suggested some variations of the L-curve methods with similar trade-offs between model structure and data prediction.

The methods that we have proposed, although only demonstrated with uncorrelated noise, should be applicable to correlated noise as well, provided the bandwidth of the noise does not overlap the bandwidth of the data.

Once the noise variance has been estimated one can determine the stability of the estimator via the calculation of the model covariance matrix. However, to fully assess the model uncertainties we must also estimate the bias. Since the bias depends on the true model, in practice we can only put bounds on the bias using *a priori* bounds on the true model. We have shown examples on how this can be done in practice.

8 ACKNOWLEDGMENTS

This work was begun while one author (JS) was on sabbatical at the École Supérieure de Physique et de Chimie Industrielles de la Ville de Paris as a Professor of the French Academy of Sciences. He wishes to express his gratitude to the Academy and TotalFinaElf for their sponsorship of this professorship. He would also like to thank his colleagues in the Laboratoire Ondes et Acoustique for their hospitality. This work was also partially supported by the sponsors of the Consortium Project on Seismic Inverse Methods for Complex Structures at the Center for Wave Phenomena and the Army Research Office under Project DAAG55-98-1-0277. We also acknowledge stimulating discussions with Mike Knoll and Bill Clement of Boise State University. Finally, we would like to acknowledge Dr. Per Christian Hansen for the use of his free MATLAB optimization toolbox.

REFERENCES

Bates, D. M. & Watts, D. G., 1998. *Nonlinear Regression Analysis and Its Applications*, Wiley.

Cummins, D. J., Filloon, T. G., & Nychka, D., 2001. Confidence intervals for nonparametric curve estimates: toward more unifom pointwise coverage, *Journal of the American Statistical Association*, **96**, 233–246.

Docherty, P., 1992. Solving for the thickness and velocity of the weathering layer using 2-D refraction tomography, *Geophysics*, **57**(10), 1307–1318.

Donoho, D. L. & Johnstone, I. M., 1995. Adapting to unknown smoothness via wavelet shrinkage, *Journal of the American Statistical Association*, **90**, 1200–1224.

Evans, S. & Stark, P., 2001. Inverse problems as statistics, *Inverse Problems*, **To appear**.

Gouveia, W. & Scales, J. A., 1998. Bayesian seismic waveform inversion: parameter estimation and uncertainty analysis, *JGR*, **103**, 2759–2779.

Green, P. J. & Silverman, B. W., 1994. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman & Hall, London.

Lawson, C. L. & Hanson, R. J., 1974. *Solving Least Squares Problems*, Prentice Hall.

Li, Y. & Oldenburg, D. W., 1999. 3-D inversion of DC resistivity data using an L-curve criterion, in *Annual Meeting Abstracts*, pp. 251–254, Society of Exploration Geophysicists.

Moraes, F. & Scales, J. A., 2000. Local Bayesian inversion: theoretical developments, *Geophys. J. Int.*, **141**, 713–723.

Oldenburg, D., Li, Y., & Ellis, R., 1997. Inversion of geophysical data over a copper gold porphyry deposit: a case history for Mt. Milligan, *Geophysics*, **62**, 1419–1431.

O'Sullivan, F., 1986. A statistical perspective on ill-posed inverse problems, *Statistical Science*, **1**(4), 502–527.

Scales, J., 1987. Tomographic inversion via the conjugate gradient method, *Geophysics*, **52**, 179–185.

Scales, J. A. & Snieder, R. K., 1998. What is noise?, *Geophysics*, **63**(4), 1122–1124.

Scales, J. A. & Tenorio, L., 2001. Prior information and uncertainty in inverse problems, *Geophysics*, **66**(2), 389–397.

Scales, J. A., Gersztenkorn, A., & Treitel, S., 1988. Fast ℓ_p solution of large, sparse linear systems: application to seismic travel time tomography, *Journal of Computational Physics*, **75**, 314–333.

Stark, P., 1992. Minimax confidence intervals in geomagnetism, *Geophys. J. Int.*, **108**, 329–338.

Stark, P., 1992. Inference in infinite dimensional inverse problems: discretization and duality, *JGR*, **97**, 14,055–14,082.

Tenorio, L., 2001. Statistical regularization of inverse problems, *SIAM Review*, **43**, 347–366.

Tikhonov, A. N. & Arsenin, V. Y., 1977. *Solutions of Ill-posed Problems*, Winston, Washington, D.C.

Van Wijk, K., Scales, J. A., Navidi, W., & Roy-Chowdhury, K., 1998. Estimating data uncertainties for least squares optimization, in *Annual Project Review*, vol. CWP 283, Center for Wave Phenomena, Colorado School of Mines.

Xu, P., 1998. Svd methods for linear ill-posed problems, *Geophys. J. Int.*, **135**(2), 505–514.