© de Gruyter 2007
J. Inv. Ill-Posed Problems **15** (2007), 1–21 DOI 10.1515 / JIP.2007.xxx

Parameter estimation: A new approach to weighting a priori information

J.L. Mead

Communicated by

Abstract. We propose a new approach to weighting initial parameter misfits in a least squares optimization problem for linear parameter estimation. Parameter misfit weights are found by solving an optimization problem which ensures the penalty function has the properties of a χ^2 random variable with n degrees of freedom, where n is the number of data. This approach differs from others in that weights found by the proposed algorithm vary along a diagonal matrix rather than remain constant. In addition, it is assumed that data and parameters are random, but not necessarily normally distributed.

The proposed algorithm successfully solved three benchmark problems, one with discontinuous solutions. Solutions from a more idealized discontinuous problem show that the algorithm can successfully weight initial parameter misfits even though the two-norm typically smoothes solutions. For all test problems sample solutions show that results from the proposed algorithm can be better than those found using the L-curve and generalized cross-validation. In the cases where the parameter estimates are not as accurate, their corresponding standard deviations or error bounds correctly identify their uncertainty.

Key words. Parameter estimation, Tikhonov regularization, Maximum likelihood estimate, Non-necessarily Gaussian noise .

AMS classification. 65F22, 93E24, 62M40.

1. Introduction

Parameter estimation is an element of inverse modeling in which measurements or data are used to infer parameters in a mathematical model. Parameter estimation is necessary in many applications such as biology, astronomy, engineering, Earth science, finance, and medical and geophysical imaging. Inversion techniques for parameter estimation are often classified in two groups: deterministic or stochastic.

Both deterministic and stochastic approaches must incorporate the fact that there are uncertainties or errors associated with parameter estimation [3], [13]. For example, in a deterministic approach such Tikhonov regularization [17] or stochastic approaches which use frequentist or Bayesian probability theory, it is assumed that data contain noise. The difference between deterministic and stochastic approaches is that in the former it is assumed there exists “true” parameter values for a given set of data while in the latter the data, parameter values or both are random variables. [14].

Parameter estimation can be viewed as an optimization problem in which an ob-

This work was supported by the NSF-Idaho EPSCoR Program and the National Science Foundation under award number EPS-0447689.

jective function representing data misfit is minimized in a given norm, [11]. From a deterministic point of view the two-norm, i.e. a quadratic objective function, is the most attractive mathematically because the minimum can be explicitly written in closed form. From the stochastic point of view this choice of two-norm is statistically the most likely solution if data are normally distributed. However, this estimate is typically less accurate if the data are not normally distributed or there are outliers in the data [16].

A more complete optimization problem will include a statement about the parameter misfit, in addition to the data misfit. This statement could be a deterministic bound such as a positivity constraint on the parameters, or a regularization term which ensures that the first or second derivative of the parameters is smooth.

When parameter misfits are included in stochastic approaches their corresponding a priori probability distributions must be specified. The advantage and disadvantage of the stochastic viewpoint is that prior information about the probability distribution of data or parameters must be specified. A priori information for the distribution of data is tractable because data can (theoretically) be collected repeatedly in order to obtain a sample from which one can infer its probability distribution. A priori inference of the parameter probability distribution is less reliable than that for the data because it must rely on information from the uncertain data [13].

Which ever way one views the problem; positivity constraints, regularization, or probability distributions, typically a weighted objective function is minimized. A significant difference between methods and their solutions lies in how the weights are chosen.

An experimental study in [15] compares deterministic and stochastic approaches to seismic inversion for characterization of a thin-bed reservoir. Their conclusion is that deterministic approaches are computationally cheaper but results are only good enough for identifying general trends and large features. They state that stochastic inversion is more advantageous because results have superior resolution and offer uncertainty estimates.

The experimental results in [15] which suggest that the stochastic approach is more accurate than the deterministic approach may occur because the stochastic approach better weights the data and parameter misfits. Most deterministic approaches such as positivity constraints or regularization only use constant or simple weights on the parameter misfits. On the other hand, stochastic approaches which specify prior normal or exponential probability distributions weight the parameter misfit with an inverse covariance matrix. Weighting with accurate non-constant, dense matrices is desirable but it implies that there is good a priori information. How do we obtain this information, i.e how do we find accurate weights on the data and parameter misfits?

In this work we use the following piece of a priori information to better weight the parameter misfit: The minimum value of a quadratic cost function representing the data and parameter misfit is a χ^2 random variable with n degrees of freedom, where n is the number of data. For large n, this is true regardless of the prior distributions of the data or parameters.

For the linear problem, an explicit expression for the minimum value of the cost function is given as a function of the weight on the parameter misfit. Since the cost function follows a χ^2 distribution, this minimum value has known mean and variance.

To calculate a weight on the parameter misfit a new optimization problem is solved which ensures the minimum of the cost function lies within the expected range.

In Section 2 we describe current approaches to solving linear discrete ill-posed problems. In Section 3 we describe the new approach and the corresponding algorithm, in Section 4 we show some numerical results, and in Section 5 we give conclusions and future work.

2. Linear Discrete Ill-posed Problems

In this work discrete ill-posed inverse problems of the form

$$\mathbf{d} = \mathbf{G}\mathbf{m} \tag{2.1}$$

are considered. Here $\mathbf{d}$ is a n dimensional vector containing measured data, $\mathbf{G}$ is a forward modeling operator written as an $n \times m$ matrix and $\mathbf{m}$ is a m dimensional vector of unknown parameter values.

2.1. Deterministic approaches

Frequently it is the case that there is no value of $\mathbf{m}$ that satisfies (2.1) exactly. Simple and useful approximations may be found by optimizing

$$\min_{\mathbf{m}} ||\mathbf{d} - \mathbf{G}\mathbf{m}||_p^p. \tag{2.2}$$

The most common choices for p are $p = 1, 2$. If $p = 2$, this is least squares optimization which is the simplest approach to analyze and statistically results in the most likely solution if the data are normally distributed. However, the least squares solution is typically not accurate if one datum is far from the trend.

If $p = 1$ accurate solutions can still be found if there are a few data far from the trend. In addition, it is statistically the most likely solution if the data are exponentially distributed. As p increases from 2 the largest element of $\mathbf{d} - \mathbf{G}\mathbf{m}$ is given successively larger weight [11].

Least squares solutions are the simplest to analyze mathematically because the value at which the minimum occurs can be stated explicitly. In other words,

$$\min_{\mathbf{m}} ||\mathbf{d} - \mathbf{G}\mathbf{m}||_2^2 = \min_{\mathbf{m}} (\mathbf{d} - \mathbf{G}\mathbf{m})^T(\mathbf{d} - \mathbf{G}\mathbf{m}) \tag{2.3}$$

has a unique minimum occurring at

$$\hat{\mathbf{m}}_{ls} = (\mathbf{G}^T\mathbf{G})^{-1}\mathbf{G}^T\mathbf{d}. \tag{2.4}$$

However, the inverse solution is not that simple because typically $\mathbf{G}^T\mathbf{G}$ is not invertible and the problem must be constrained or regularized. One common way to do this is Tikhonov regularization in the two-norm where $\mathbf{m}$ is found by solving

$$\min_{\mathbf{m}} \left\{ ||\mathbf{d} - \mathbf{G}\mathbf{m}||_2^2 + \lambda ||\mathbf{L}(\mathbf{m} - \mathbf{m}_0)||_2^2 \right\} \tag{2.5}$$

with $\mathbf{m}_0$ an initial parameter estimate (often taken to be 0), λ a yet to be determined regularization parameter, and $\mathbf{L}$ a smoothing operator possibly chosen to represent the first or second derivative.

The optimization problem (2.5) can be written equivalently as a constrained minimization problem:

$$\min_{\mathbf{m}} ||\mathbf{d} - \mathbf{G}\mathbf{m}||_2 \quad \text{subject to} \quad ||\mathbf{L}(\mathbf{m} - \mathbf{m}_0)||_2 \leq \delta. \tag{2.6}$$

In either formulation, (2.5) or (2.6), the optimization problem can be written as

$$\min_{\mathbf{m}} \left\{ (\mathbf{d} - \mathbf{G}\mathbf{m})^T(\mathbf{d} - \mathbf{G}\mathbf{m}) + (\mathbf{m} - \mathbf{m}_0)^T \lambda \mathbf{L}^T\mathbf{L}(\mathbf{m} - \mathbf{m}_0) \right\}. \tag{2.7}$$

When the optimization problem is written this way we see that the objective function is the sum of a data and parameter misfit. The function is normalized so that the data misfit has weight equal to one while the parameter misfit has weight $\lambda\mathbf{L}^T\mathbf{L}$. Thus $\lambda\mathbf{L}^T\mathbf{L}$ represents an a priori ratio of weights on the data and parameter misfits. Typically, $\mathbf{L}$ is taken to be the identity, first or second derivative operator. There are numerous approaches for choosing λ including the L-curve [8], Morozov's discrepancy principle [12] and generalized cross-validation [9].

The minimum of (2.7) occurs at

$$\hat{\mathbf{m}}_{rls} = \mathbf{m}_0 + (\mathbf{G}^T\mathbf{G} + \lambda\mathbf{L}^T\mathbf{L})^{-1}\mathbf{G}^T(\mathbf{d} - \mathbf{G}\mathbf{m}_0). \tag{2.8}$$

This deterministic parameter estimate, (2.8), from Tikhonov regularization does not use a priori knowledge, other than specification of the form of $\mathbf{L}$.

2.2. Stochastic approaches

Some stochastic formulations can lead to an optimization problem similar to (2.5). The difference between these stochastic approaches and the corresponding deterministic ones is the way in which the weights on the data and parameter misfits are chosen. For example, assume the data $\mathbf{d}$ are random, independent and identically distributed, following a normal distribution with probability density function

$$\rho(\mathbf{d}) = \text{const} \times \exp\left\{ -\frac{1}{2}(\mathbf{d} - \mathbf{G}\mathbf{m})^T\mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{G}\mathbf{m}) \right\}, \tag{2.9}$$

with $\mathbf{G}\mathbf{m}$ the expected value of $\mathbf{d}$ and $\mathbf{C}_d$ the corresponding covariance matrix. In order to maximize the probability that the data were in fact observed we find $\mathbf{m}$ where the probability density is maximum. This is the maximum likelihood estimate and it is the minimum of the argument in (2.9), i.e. the optimal parameters $\mathbf{m}$ are found by solving

$$\min_{\mathbf{m}} (\mathbf{d} - \mathbf{G}\mathbf{m})^T\mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{G}\mathbf{m}). \tag{2.10}$$

This is the weighted least squares problem and the minimum occurs at

$$\hat{\mathbf{m}}_{wls} = (\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G})^{-1}\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{d}. \tag{2.11}$$

Similar to (2.4), $\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G}$ is typically not invertible. In this case the stochastic problem can be constrained or regularized by adding more a priori information. For example, assume the parameter values $\mathbf{m}$ are also random following a normal distribution with probability density function

$$\rho(\mathbf{m}) = \text{const} \times \exp\left\{-\frac{1}{2}(\mathbf{m}-\mathbf{m}_0)^T\mathbf{C}_m^{-1}(\mathbf{m}-\mathbf{m}_0)\right\}, \tag{2.12}$$

with $\mathbf{m}_0$ the expected value of $\mathbf{m}$ and $\mathbf{C}_m$ the corresponding covariance matrix. If the data and parameters are independent then their joint distribution is

$$\rho(\mathbf{d},\mathbf{m}) = \rho(\mathbf{d})\rho(\mathbf{m}).$$

The maximum likelihood estimate of the parameters occurs when the joint probability density function is maximum, i.e. optimal parameter values are found by solving

$$\min_{\mathbf{m}}\left\{(\mathbf{d}-\mathbf{G}\mathbf{m})^T\mathbf{C}_d^{-1}(\mathbf{d}-\mathbf{G}\mathbf{m}) + (\mathbf{m}-\mathbf{m}_0)^T\mathbf{C}_m^{-1}(\mathbf{m}-\mathbf{m}_0)\right\}. \tag{2.13}$$

The minimum occurs at

$$\hat{\mathbf{m}} = \mathbf{m}_0 + (\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1})^{-1}\mathbf{G}^T\mathbf{C}_d^{-1}(\mathbf{d}-\mathbf{G}\mathbf{m}_0). \tag{2.14}$$

The stochastic parameter estimate (2.14) has been found under the assumption that the data and parameters follow a normal distribution and are independent and identically distributed.

2.3. Comparison between Deterministic and Stochastic approaches

Now we are in a situation to point out similarities between Tikhonov regularization in the two-norm and a stochastic approach for normally distributed data and parameters. The two equations (2.8) and (2.14) are equivalent if $\mathbf{C}_d^{-1} = \mathbf{I}$ and $\mathbf{C}_m^{-1} = \lambda\mathbf{L}^T\mathbf{L}$. Even though the two-norm smoothes parameter estimates and assumes independent and identically distributed parameters, following normal probability distributions, we can see under these simplifying assumptions how a stochastic approach would give better results. In the stochastic approach dense a priori covariance matrices better weight the the data and parameter misfits than when the weights are $\lambda\mathbf{L}^T\mathbf{L}$ with Tikhonov regularization. This is the justification for the success of the proposed method.

As further explanation of the advantage of a stochastic approach over a deterministic one consider the deterministic constraint

$$||\mathbf{m}-\mathbf{m}_0|| < \lambda.$$

When this constraint is applied, each element of $\mathbf{m}-\mathbf{m}_0$ is equally weighted which implies that the error in the initial guess $\mathbf{m}_0$ is the same for each element. Weighting in this manner will not be the best approach if a large temperature change or other such anomaly is sought. On the other hand, non-constant weights such as prior covariances $\mathbf{C}_m$ may vary along a diagonal matrix and hence give different weight to each element of $\mathbf{m}-\mathbf{m}_0$. The weights can be further improved if the prior is non-diagonal because then correlation between initial estimate errors can be identified.

Regardless of the norm in which the objective function is minimized or how the problem is formulated, the over-riding question is: How should weights on the terms in the objective function be chosen? In Section 3 we will show that if a quadratic cost function is used there is one more piece of a priori information that can be used to find weights on parameter misfits. In Section 4 we will show that when using weights chosen in this manner the parameter estimates are not smoothed and it need not be assumed that the data or parameters are normally distributed.

3. New approach

Rather than choosing a deterministic approach which uses no a priori information or a stochastic approach which may use incorrect a priori information we focus on finding the best way to weight data and parameter misfits in the two-norm using available a priori information.

Consider parameter estimation reformulated in the following manner. Given data $\mathbf{d}$, accurate mapping $\mathbf{G}$ and initial estimate $\mathbf{m}_0$, find $\mathbf{m}$ such that

$$\mathbf{d} = \mathbf{G}\mathbf{m}+\boldsymbol{\epsilon} \tag{3.1}$$

$$\mathbf{m} = \mathbf{m}_0+\mathbf{f} \tag{3.2}$$

where $\boldsymbol{\epsilon}$ and $\mathbf{f}$ are unknown errors in the data and initial parameter estimates, respectively.

We can view $\mathbf{m}$ and $\mathbf{d}$ as random variables or alternatively, $\mathbf{m}$ as the "true" parameter estimate and $\mathbf{d}$ as data with error. In either case, parameter estimates are found by minimizing the errors in the data ($\boldsymbol{\epsilon}$) and initial estimates ($\mathbf{f}$) in a weighted least squares sense, i.e. solve the following optimization problem

$$\min_{\mathbf{m}} \left\{ (\mathbf{d}-\mathbf{G}\mathbf{m})^T\mathbf{W}_d(\mathbf{d}-\mathbf{G}\mathbf{m}) + (\mathbf{m}-\mathbf{m}_0)^T\mathbf{W}_m(\mathbf{m}-\mathbf{m}_0)\right\} \tag{3.3}$$

with $\mathbf{W}_d$ and $\mathbf{W}_m$ weights (yet to be determined) on the error in the data $\mathbf{d}$ and initial parameter estimates $\mathbf{m}_0$, respectively.

3.1. Choice of weights

The weights on the data misfit will be taken to be the inverse of the covariance of the data, i.e. $\mathbf{W}_d=\mathbf{C}_d^{-1}$. If the statistical properties of the data are not known this weight can be estimated by collecting the same data repeatedly and calculating the sample standard deviation. We do assume however, that the data is good and that $\mathbf{G}\mathbf{m}$ is the mean of $\mathbf{d}$.

To find the weights on the parameter misfit, $\mathbf{W}_m$, we use the following Theorem.

Theorem 3.1. *Define $\mathcal{J}(\mathbf{m})$ by*

$$\mathcal{J}_{\text{mls}} = (\mathbf{d}-\mathbf{G}\mathbf{m})^T\mathbf{C}_d^{-1}(\mathbf{d}-\mathbf{G}\mathbf{m}) + (\mathbf{m}-\mathbf{m}_0)^T\mathbf{C}_m^{-1}(\mathbf{m}-\mathbf{m}_0) \tag{3.4}$$

with $\mathbf{d}$ and $\mathbf{m}$ stochastic. In addition, assume the errors in the data $\mathbf{d}$ and initial guess $\mathbf{m}_0$ are not necessarily normally distributed but have mean zero and covariances $\mathbf{C}_d$ and $\mathbf{C}_m$, respectively. Then as the number of data n approaches infinity, the minimum value of (3.4) is a random variable and its limiting distribution is the χ^2 distribution with n degrees of freedom.

Proof. Case 1: Elements of $\mathbf{d}$ and $\mathbf{m}$ are independent and identically normally distributed. It is well known that under normality assumptions the first and second terms in the right hand side of (3.4) are χ^2 random variables with $n-m$ and m degrees of freedom, respectively.

Case 2: Elements of $\mathbf{d}$ and $\mathbf{m}$ are independent and identically distributed but not normally distributed. The minimum value of $\mathcal{J}(\mathbf{m})$ occurs at

$$\hat{\mathbf{m}} = \mathbf{m}_0 + (\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1})^{-1}\mathbf{G}^T\mathbf{C}_d^{-1}(\mathbf{d} - \mathbf{G}\mathbf{m}_0). \tag{3.5}$$

Re-write the matrix in (3.5) by noting that

$$\begin{aligned}\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G}\mathbf{C}_m\mathbf{G} + \mathbf{G}^T &= \mathbf{G}^T\mathbf{C}_d^{-1}\left(\mathbf{G}\mathbf{C}_m\mathbf{G}^T + \mathbf{C}_d\right)\\ &= \left(\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1}\right)\mathbf{C}_m\mathbf{G}^T,\end{aligned}$$

thus

$$\left(\mathbf{G}^T\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{C}_m^{-1}\right)^{-1}\mathbf{G}^T\mathbf{C}_d^{-1} = \mathbf{C}_m\mathbf{G}^T\left(\mathbf{G}\mathbf{C}_m\mathbf{G}^T + \mathbf{C}_d\right)^{-1}. \tag{3.6}$$

Let $\mathbf{h} = \mathbf{d} - \mathbf{G}\mathbf{m}_0$ and $\mathbf{P} = \mathbf{G}\mathbf{C}_m\mathbf{G}^T + \mathbf{C}_d$, then

$$\hat{\mathbf{m}} = \mathbf{m}_0 + \mathbf{C}_m\mathbf{G}^T\mathbf{P}^{-1}\mathbf{h}. \tag{3.7}$$

The minimum value of $\mathcal{J}(\mathbf{m})$ is

$$\begin{aligned}\mathcal{J}(\hat{\mathbf{m}}) &= \left(\mathbf{h} - \mathbf{G}\mathbf{C}_m\mathbf{G}^T\mathbf{P}^{-1}\right)^T\mathbf{C}_d^{-1}\left(\mathbf{h} - \mathbf{G}\mathbf{C}_m\mathbf{G}^T\mathbf{P}^{-1}\right)\\ &\quad + \left(\mathbf{C}_m\mathbf{G}^T\mathbf{P}^{-1}\mathbf{h}\right)^T\mathbf{C}_m^{-1}\left(\mathbf{C}_m\mathbf{G}^T\mathbf{P}^{-1}\mathbf{h}\right).\end{aligned} \tag{3.8}$$

Since $\mathbf{C}_d$ and $\mathbf{C}_m$ are covariance matrices, they are symmetric positive definite, and we can simplify (3.8) to:

$$\mathcal{J}(\hat{\mathbf{m}}) = \mathbf{h}^T\mathbf{P}^{-1}\mathbf{h}. \tag{3.9}$$

In addition, since $\mathbf{G}$ is full rank, $\mathbf{P}^{-1}$ and hence $\mathbf{P}$ are symmetric positive definite and we can define

$$\mathbf{h} = \mathbf{P}^{\frac{1}{2}}\mathbf{k}, \tag{3.10}$$

where

$$k_j = \sum_{i=1}^{n}(P^{-\frac{1}{2}})_{ji}h_i. \tag{3.11}$$

If the errors in the data $\mathbf{d}$ and initial guess $\mathbf{m}_0$ are normally distributed then h_j are normal and hence k_j are normal by linearity. On the other hand, if the errors in the data $\mathbf{d}$ and initial guess $\mathbf{m}_0$ are not normally distributed, then the central limit states that as n approaches infinity, k_j defined by (3.11) is a normally distributed random variable with zero mean and unit variance.

Now writing (3.9) in terms of $\mathbf{k}$ we have

$$\begin{aligned} \mathcal{J}(\hat{\mathbf{m}}) &= \mathbf{k}^T\mathbf{P}^{\frac{1}{2}}\mathbf{P}^{-1}\mathbf{P}^{\frac{1}{2}}\mathbf{k} && (3.12)\\ &= \mathbf{k}^T\mathbf{k} && (3.13)\\ &= k_1^2 + \ldots k_n^2. && (3.14) \end{aligned}$$

For large n the k_j are normally distributed random variables ir-regardless of the distribution of the errors in $\mathbf{d}$ and $\mathbf{m}_0$. Thus as n approaches infinity, $\mathcal{J}(\hat{\mathbf{m}})$ is a χ^2 random variable with n degrees of freedom. This is described more generally in [1]. □

We have shown that the objective function in (3.3) is a χ^2 random variable with n degrees of freedom regardless of the prior distributions of the data and parameter misfits. Thus the weights on the parameter misfits will be found via an optimization problem which ensures that the objective function in (3.3) lies within a critical region of the χ^2 distribution with n degrees of freedom.

4. Algorithm

We can determine, within specified confidence intervals, values of $\mathbf{W}_m = \mathbf{C}_m^{-1}$ in (3.3) that ensure $\mathcal{J}(\hat{\mathbf{m}})$ given by (3.8) is a χ^2 random variable with n degrees of freedom when there is a large amount of data. The larger the confidence interval, the bigger the set of possible values of $\mathbf{W}_m$. These values of $\mathbf{W}_m$ will be for a specified covariance matrix for the errors in the data $\mathbf{C}_d$, and for a specified initial guess $\mathbf{m}_0$. Thus for each matrix $\mathbf{W}_m$ there is an optimal parameter value $\hat{\mathbf{m}}$ uniquely defined by (3.7).

If $\mathbf{W}_m = \lambda^{-1}\mathbf{I}$ then this algorithm is similar to approaches such as the L-curve for finding the regularization parameter λ in Tikhonov regularization. However, the advantage of this new approach is when $\mathbf{W}_m$ is not a constant matrix and hence the weights on the parameter misfits vary. Moreover, when $\mathbf{W}_m$ has off diagonal elements, correlation in initial parameter estimate errors can be modeled.

One advantage to viewing optimization stochastically is that once the optimal parameter estimate is found, the corresponding uncertainty or covariance matrix for $\hat{\mathbf{m}}$ is given by [16]

$$\text{cov}(\hat{\mathbf{m}}) = \mathbf{W}_m^{-1} - \mathbf{W}_m^{-1}\mathbf{G}^T\mathbf{P}^{-1}\mathbf{G}\mathbf{W}_m^{-1}, \tag{4.1}$$

with $\mathbf{P} = \mathbf{G}\mathbf{W}_m^{-1}\mathbf{G}^T + \mathbf{C}_d$. In Section 5 the numerical estimates of $\hat{\mathbf{m}}$ are plotted with error bars which represent standard deviations of these estimates. These standard deviations are the square root of the diagonal elements of (4.1).

4.1. Confidence Intervals

For large n $\mathcal{J}(\hat{\mathbf{m}})$ has mean n and variance $\sqrt{2n}$. The $(1-\alpha)$% confidence interval for the mean is

$$\mathrm{P}\left(-z_{\alpha/2} < \left(\mathcal{J}(\hat{\mathbf{m}}) - n\right)/\sqrt{2} < z_{\alpha/2}\right) = 1-\alpha, \tag{4.2}$$

or

$$\mathrm{P}\left(n-\sqrt{2}z_{\alpha/2} < \mathcal{J}(\hat{\mathbf{m}}) < n+\sqrt{2}z_{\alpha/2}\right) = 1-\alpha, \tag{4.3}$$

where $z_{\alpha/2}$ is the z-value on the normal curve above which we find an area of $\alpha/2$. Thus for a given $(1-\alpha)$ confidence interval we find values of $\mathbf{W}_m^{-1}$ that ensure

$$n-\sqrt{2}z_{\alpha/2} < \mathcal{J}(\hat{\mathbf{m}}) < n+\sqrt{2}z_{\alpha/2}$$

or

$$n-\sqrt{2}z_{\alpha/2} < \mathbf{h}^T\left(\mathbf{G}\mathbf{W}_m^{-1}\mathbf{G}^T + \mathbf{C}_d\right)^{-1}\mathbf{h} < n+\sqrt{2}z_{\alpha/2}. \tag{4.4}$$

By choosing a value of $\alpha = 0.05$, for example, we are stating that we are 95% confident that the mean of $\mathcal{J}(\hat{\mathbf{m}})$ is n. In this case the interval in which the cost function lies is $[n-2.77, n+2.77]$, while for $\alpha = 0.10$ it is $[n-3.64, n+3.64]$. For large n this is a small interval, thus in our experiments the optimization problem is to find a $\mathbf{W}_m$ such that

$$\mathbf{h}^T\left(\mathbf{G}\mathbf{W}_m^{-1}\mathbf{G}^T + \mathbf{C}_d\right)^{-1}\mathbf{h} = n. \tag{4.5}$$

4.2. Optimization

There is actually a set of feasible solutions $\mathbf{W}_m$ that ensure (4.5) holds. This boundary surface is well behaved as long as $\mathbf{P}$ is well-conditioned. The solution we seek is the one in which $||\mathbf{W}_m^{-1}||$ is minimized because this will most likely result in the strongest "regularization" or well-conditioned matrix to invert in (2.14). The norm we choose is the Frobenius norm, i.e.

$$||A||_F^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2,$$

because it is continuously differentiable and it is equivalent to the 2-norm via

$$\frac{1}{\sqrt{n}}||A||_F \leq ||A||_2 \leq ||A||_F.$$

If $\mathbf{W}_m$ is to represent the inverse covariance matrix $\mathbf{C}_m$ it must be symmetric positive definite, thus we define

$$\mathbf{C}_m = \mathbf{L}\mathbf{L}^T$$

with $\mathbf{L}$ lower triangular. We also assume that $\mathbf{G}$ is full rank and n is large. However, these assumptions may be dropped if $\mathbf{P}$ is symmetric positive definite and the data are normally distributed. The corresponding algorithm is given in Table 1 and results from it are given in Section 5.

Table 1. Algorithm for weights

Minimize	$\|\|\mathbf{LL}^T\|\|_F^2$
subject to	$n - \sqrt{2}z_{\alpha/2} < \mathbf{h}^T \left(\mathbf{GLL}^T\mathbf{G}^T + \mathbf{C}_d\right)^{-1} \mathbf{h} < n + \sqrt{2}z_{\alpha/2}$
	$\mathbf{GLL}^T\mathbf{G}^T + \mathbf{C}_d$ well conditioned

5. Numerical Tests

Matlab was used in all numerical tests, and the optimization problem in Table 1 was solved using the function *fmincon* in the Optimization Toolbox. Future work involves finding more efficient approaches to the optimization problem so that the proposed methodology can be used for more realistic problems. For each test problem the inputs vary, and they are: the data $\mathbf{d}$, mapping $\mathbf{G}$, initial parameter estimate $\mathbf{m}_0$ and data misfit weight $\mathbf{W}_d = \mathbf{C}_d^{-1}$. In all cases, this weight is a diagonal matrix which is the inverse of the data variances, i.e. $diag(\sigma_i^d)^2$. More accurate representations of the error covariance from sample data takes considerably more work, see for example [7]. The outputs from the proposed method include the weights on the parameter misfit $\mathbf{W}_m = \mathbf{LL}^T$ calculated by the algorithm given in Table (1), parameter estimates (2.14), and their corresponding uncertainties (4.1). In the first test, Section 5.1, the emphasis is on showing how the calculated weights can properly identify a discontinuity, and hence only the value of the weights $\mathbf{W}_m$ found by the algorithm in Table 1 are shown. In the remaining tests, Section 5.2, the emphasis is on parameter estimation and the corresponding uncertainties, thus the weights are not shown.

5.1. Discontinuous parameters in an idealized model

In the first test we sought to determine if a diagonal $\mathbf{W}_m$ could be found which accurately weights the error in the initial parameter misfit when the parameters are discontinuous. The parameters values are $(0, \ldots, 0, 1, \ldots, 1, 0, \ldots, 0)$ with $m = n = 70$. The matrix $\mathbf{G}$ is taken to be the identity so that initial parameter errors are known and the accuracy of the weights are easy to identify. A more realistic test with a discontinuous solution is given in Section 5.2.2.

Results from the algorithm are plotted in Figures 1 and 2 when the data and parameters are taken from normal and exponential distributions, respectively, with a standard deviation of 5×10^{-2}. The parameter misfit $\mathbf{m} - \mathbf{m}_0$ is plotted along with the diagonal entries of $\mathbf{W}_m^{-1}$ found by the proposed algorithm. Each of the four plots in Figures 1 and 2 represent differential initial estimates.

In three of the four plots, i.e. for $(m_0)_i = 1, 0, 5$ $i = 1, \ldots, 70$, the diagonal elements of $\mathbf{W}_m^{-1}$ found by the proposed algorithm do indeed jump between smaller and larger values accurately reflecting the error in the initial parameter estimate $(\mathbf{m} - \mathbf{m}_0)_i$. In the fourth plot the calculated $(\mathbf{W}_m^{-1})_i$ still accurately weights the parameter misfit, but in this case the misfit is constant at 0.5. Recall that a large diagonal element $(\mathbf{W}_m^{-1})_{ii}$ gives small weight to $(\mathbf{m} - \mathbf{m}_0)_i$, which is desired when $(\mathbf{m} - \mathbf{m}_0)_i$ is large. That is, if

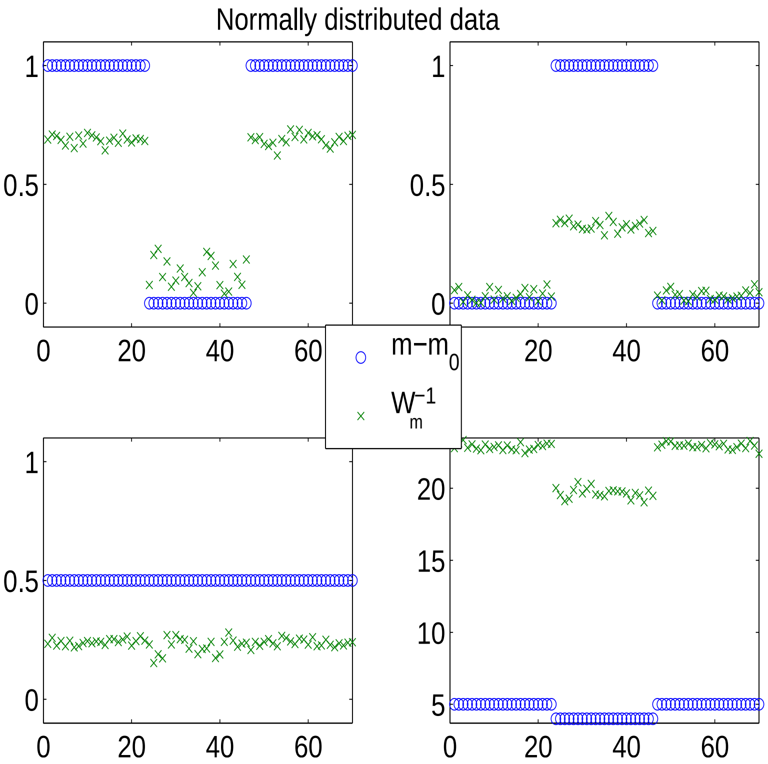

Figure 1: Parameter misfits and their corresponding weights found with the proposed algorithm in Table 1. Clockwise from top left: $\mathbf{m}_0 = (1, \dots, 1), (0, \dots, 0), (0.5, \dots, 0.5), (5, \dots, 5)$.

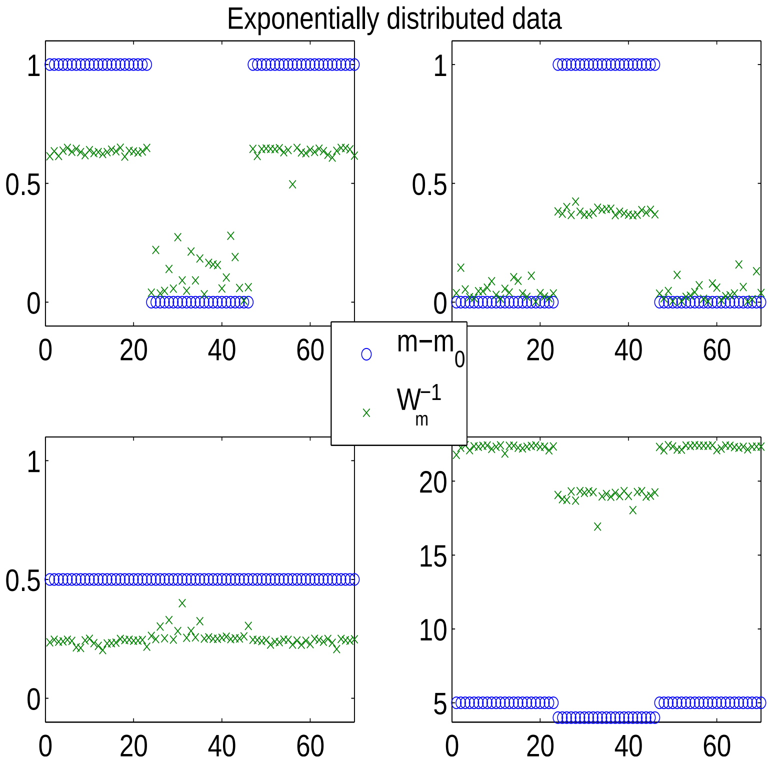

Figure 2: Parameter misfits and their corresponding weights found with the proposed algorithm in Table 1. Clockwise from top left: $\mathbf{m}_0 = (1, \ldots, 1), (0, \ldots, 0), (0.5, \ldots, 0.5), (5, \ldots, 5)$.

we have a bad initial guess $(\mathbf{m}_0)_i$, we don't want to find a $(\mathbf{m})_i$ near it but instead give small weight to minimizing $(\mathbf{m}-\mathbf{m}_0)_i$. The difficulty in weighting parameters misfits in (3.3) is that typically we do not know the accuracy of $\mathbf{m}_0$ a priori. However, in this simple example, the weights found by the proposed algorithm do appropriately weight the parameter misfits without a priori knowledge.

5.2. Benchmark Problems from [5]

Both analysis routines and test problems from [5] were used to compare and test the algorithm in Table 1. Results from the proposed algorithm were compared to those found from Tikhonov regularization with the L-curve, and and from generalized cross-validation. The L-curve approach plots the parameter misfit (with $\mathbf{m}_0 = \mathbf{0}$ and weighted by $\lambda \mathbf{L}^T\mathbf{L}$) versus the data misfit (with weight $\mathbf{I}$) to display the compromise between minimizing these two quantities. When plotted in a log-log scale the curve is typically in the shape of an L, and the solution is the parameter values at the corner. These parameter values are optimal in the sense that the error in the weighted parameter misfit and data misfit are balanced. Generalized cross-validation is based on the theory that if some data are left out, the resulting choice of parameters should result in accurate prediction of these data.

There are a total of 12 test problems in [5] and here we solve three of them: *Phillips*, *Wing* and *Shaw*. They are all derived from approximating a Fredholm integral of the first kind:

$$\int_a^b K(s,t)f(t)dt = g(s). \tag{5.1}$$

The *Phillips* and *Wing* problems use Galerkin methods with particular basis functions to approximate the Fredholm integral while the *Shaw* problem uses a weighted sum quadrature method. All approaches or problems lead to a system of linear algebraic equations $\mathbf{Gm} = \mathbf{d}$ however, in the Phillips and Wing test problem $\mathbf{Gm}$ is different from $\mathbf{d}$.

Noise is added to $\mathbf{d}$ from normal or exponential distributions. The standard deviation varies with each element in $\mathbf{d}$, but is of the order of 10^{-2}. The initial estimate $\mathbf{m}_0$ is found similarly, i.e. by adding noise to $\mathbf{m}$ from normal or exponential distributions with a varying standard deviation of the order of 10^{-2}.

The weights on the data misfit and the initial estimate $\mathbf{m}_0$ are the same for all three solution methods (L-curve, GCV and the algorithm in Table 1) and are $\mathbf{W}_d^{-1} = \mathbf{C}_d = diag(\sigma_i^d)^2$ as in the first numerical example. The algorithm in Table 1 is used to find a diagonal weight $\mathbf{W}_m$ for the parameter misfit. Future work involves finding weights $\mathbf{W}_m$ with more structure.

Since we assume in the proposed algorithm that the data and parameters are random, but from arbitrary distributions, we can assign posterior uncertainties via (4.1). These are represented as error bars in Figures 3-9.

5.2.1 Phillips test problem

This problem was presented by D.L. Phillips [6] and for (5.1) uses

$$\begin{aligned} K(s,t) &= \psi(s-t) \\ f(t) &= \psi(t) \\ g(s) &= (6-|s|)\left(1+\frac{1}{2}\cos\left(\frac{\pi s}{3}\right)\right)+\frac{9}{2\pi}\sin\left(\frac{\pi|s|}{3}\right) \end{aligned}$$

with

$$\psi(x) = \begin{cases} 1+\cos(\frac{\pi x}{3}, & |x| < 3 \\ 0, & |x| \geq 3 \end{cases} .$$

Sample solutions $\hat{\mathbf{m}}$ of $\mathbf{Gm} = \mathbf{d}$ derived from the approximation of (5.1) are plotted in Figures 3-5. The reference solution is the value of $\mathbf{m}$ given by the test problem.

In these samples almost every parameter estimates found by the new algorithm is better than those found by the other two methods. The error bars associated with the parameter estimates found by the new algorithm are all small in Figures 3 and 4 because the estimates are good. However, there are estimates for which the error bars do not reach the reference solution.

In Figure 5 another normally distributed sample solution set is plotted on the left. The randomly generated data in this sample have the same standard deviation as the samples in Figures 3 and 4 however, in this run the data were more noisy. The plot on the right is of the absolute error and standard deviation of each parameter estimate found by the new algorithm. The absolute error is the difference between the parameter estimate and the reference solution. This plot shows that standard deviation estimates from the new algorithm are of the same order of magnitude as the absolute error, or distance from the reference solution. By looking more closely at Figures 3-5 we see that in fact error bars on parameter estimates from the new algorithm often, but not always, reach the reference solution.

5.2.2 Wing test problem

The solution of this test problem contains discontinuous parameters, which is a good test of the proposed algorithm since it uses two-norm. In this problem

$$\begin{aligned} K(s,t) &= te^{-st^2} \\ g(s) &= \frac{e^{-s/9}-e^{-4s/9}}{2s} \end{aligned}$$

and

$$f(t) = \begin{cases} 1 & \frac{1}{3} < t < \frac{2}{3} \\ 0 & \text{otherwise} \end{cases} .$$

Sample parameter estimates for all three methods are given in Figures 6 and 7. Figure 6 shows one sample when the data are normally distributed while Figure 7 contains

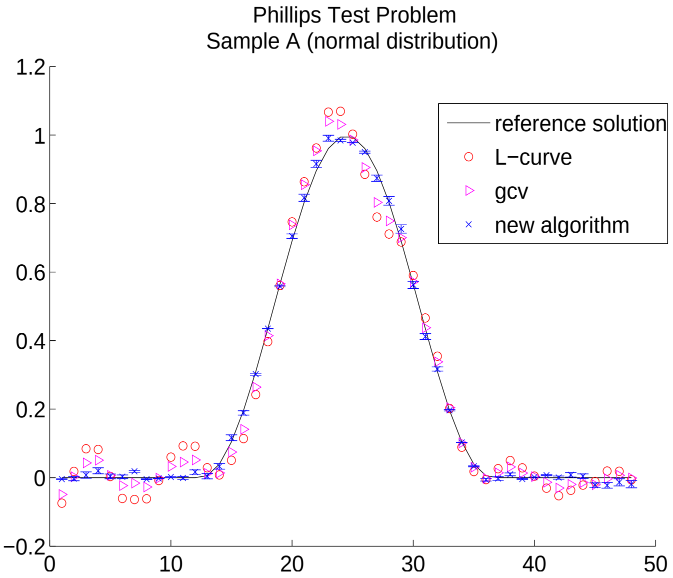

Figure 3: Sample parameter estimates for the Phillips test problem. Estimates are found by (i) the L-curve, (ii) generalized cross-validation and (iii) the proposed algorithm in Table 1 which also has error bars. The data noise are from a normal distributions.

two sample solutions when the data are exponentially distributed. Generalized cross validation did not perform well in all cases while the results from the L-curve and the proposed algorithm are good. All estimates from the new algorithm are as good as or better than those from the L-curve. Since the estimates are good, the corresponding error bars are all small. There are instances for which the error bars do not reach the reference solution however, the standard deviation still represents the absolute error well.

5.2.3 Shaw test problem

This test problem is a one dimensional image restoration model. In this problem

$$\begin{aligned} K(s,t) &= (\cos(s)+\cos(t))^2\left(\frac{\sin(u)}{u}\right)^2 \\ u &= \pi\left(\sin(s)+\sin(t)\right) \\ f(t) &= 2e^{-6(t-0.8)^2}+e^{-2(t+0.5)^2}. \end{aligned}$$

This is discretized by collocation to produce $\mathbf{Gm}$, while $\mathbf{d}$ is found by multiplying $\mathbf{G}$ and $\mathbf{m}$.

Two sample results from normal distributions are shown in Figure 8. The left plot shows a sample where all methods performed well while the parameter estimates found

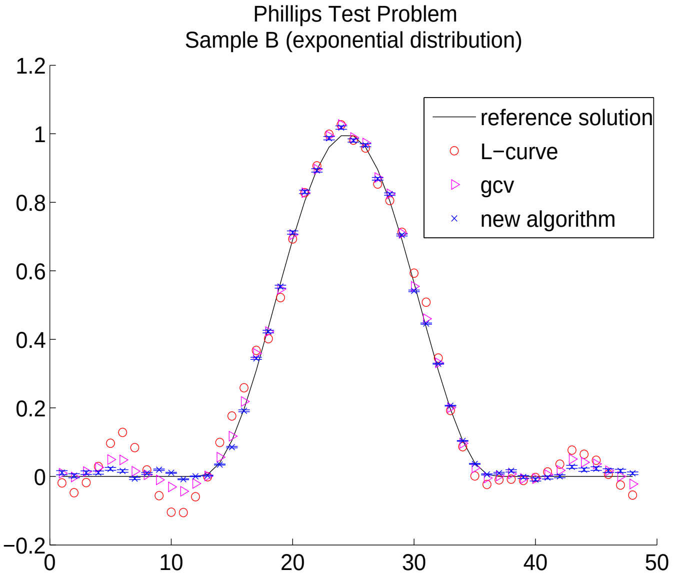

Figure 4: Sample parameter estimates for the Phillips test problem. Estimates are found by (i) the L-curve, (ii) generalized cross-validation and (iii) the proposed algorithm in Table 1 which also has error bars. The data noise are from an exponential distributions.

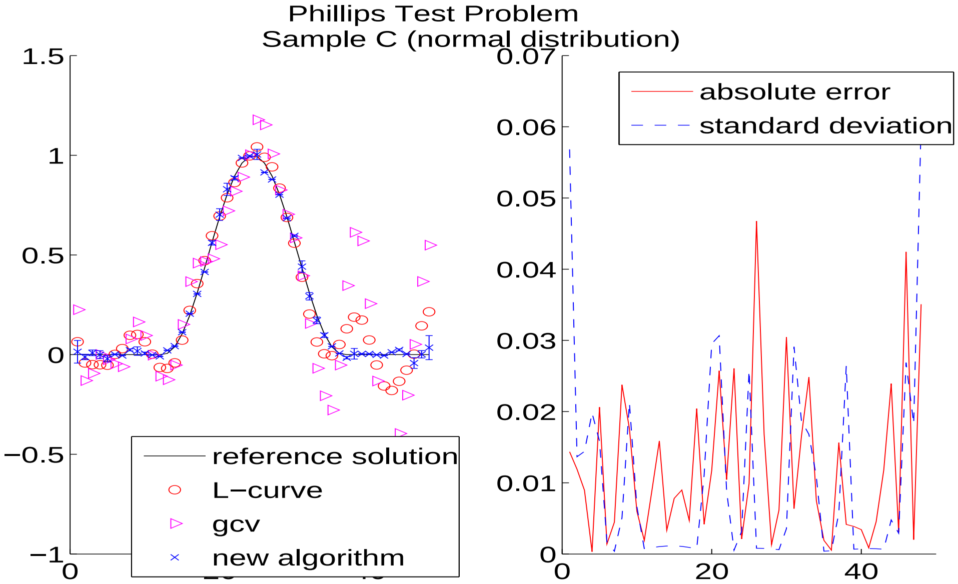

Figure 5: Sample parameter estimates for the Phillips test problem. Estimates are found by (i) the L-curve, (ii) generalized cross-validation and (iii) the proposed algorithm in Table 1 which also has error bars. The data noise are from an exponential distributions.

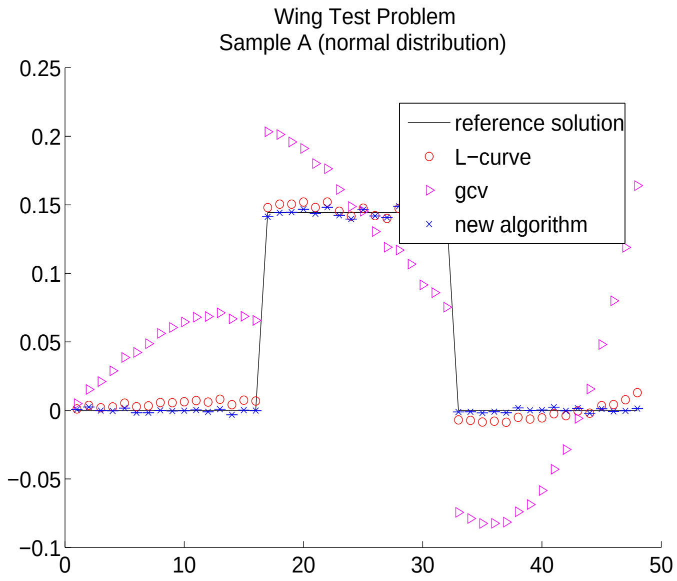

Figure 6: Sample parameter estimates for the Wing test problem. Estimates are found by (i) the L-curve, (ii) generalized cross-validation and (iii) the proposed algorithm in Table 1 which also has error bars. The data noise are from a normal distributions.

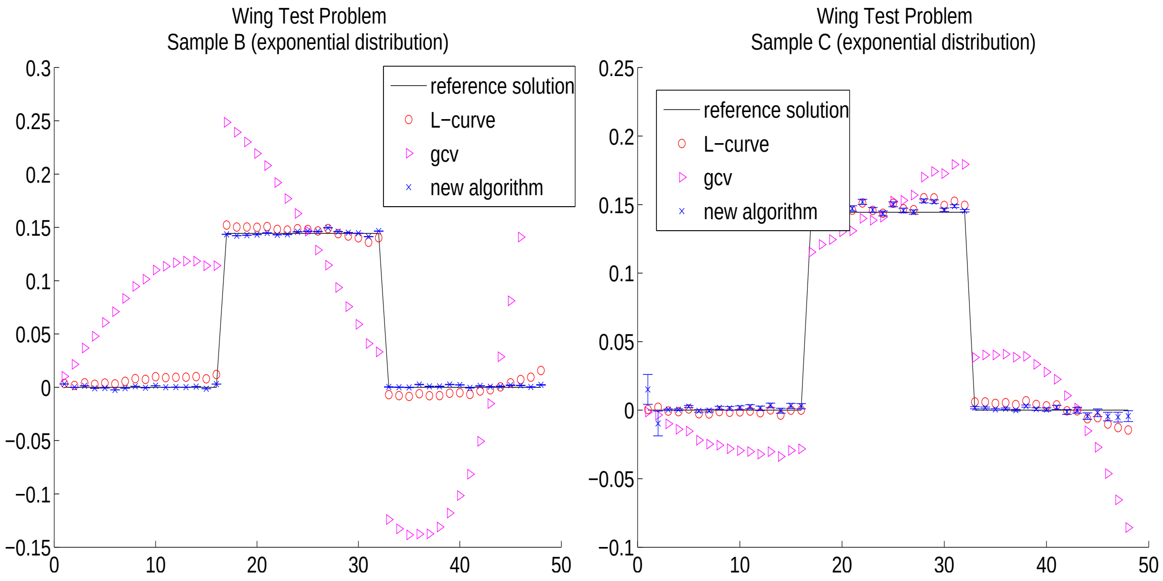

Figure 7: Sample parameter estimates for the Wing test problem. Estimates are found by (i) the L-curve, (ii) generalized cross-validation and (iii) the proposed algorithm in Table 1 which also has error bars. The data noise are from exponential distributions.

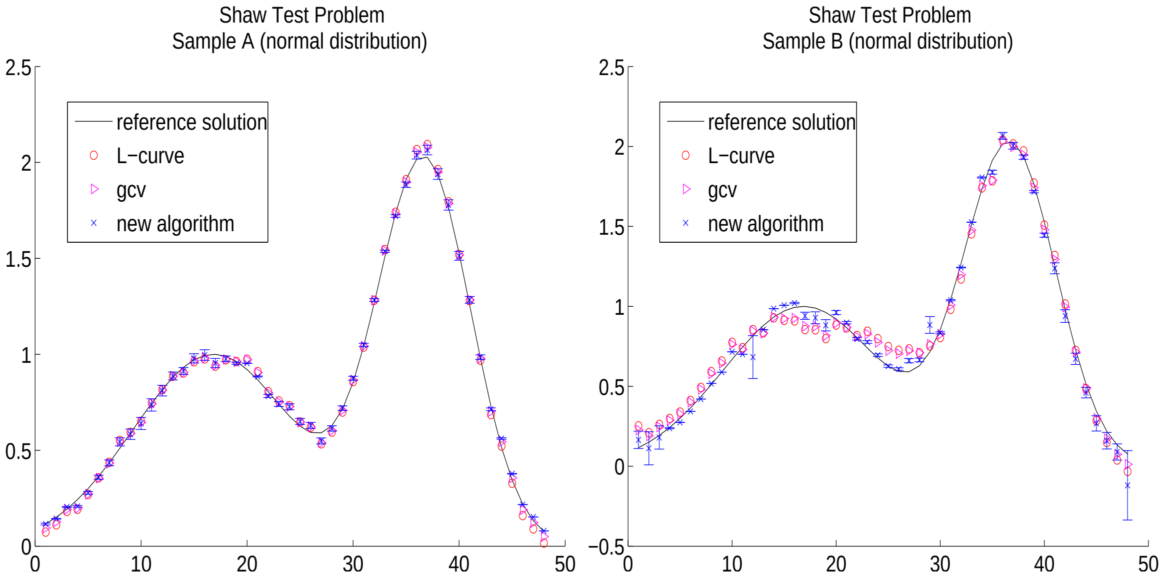

Figure 8: Sample parameter estimates for the Shaw test problem. Estimates are found by (i) the L-curve, (ii) generalized cross-validation and (iii) the proposed algorithm in Table 1 which also has error bars. The data noise are from a normal distributions.

by the new algorithm were slightly closer to the reference solution. Correspondingly, the error bars on the parameter estimates correctly identify small uncertainty. The right plot is a sample for which the majority of the parameter estimates found by the new algorithm are better than those found by the other two methods. However, there are a few estimates which are worse. The new algorithm is still useful in these instances because as is typically the case, the error bars reach or come near the reference solution in every estimate.

Figure 9 shows one sample result when the data are taken from an exponential distribution. Here we see that the L-curve and GCV estimates are much worse than those found with the new algorithm. The results from this sample are typical when data are taken from an exponential distribution. Since the data are not normally distributed, least squares estimation is statistically not the best approach. However, we see that by appropriately weighting the errors in the data and parameter misfits with the proposed algorithm, the two-norm is still a useful estimator.

6. Conclusions

We propose a new algorithm which combines ideas from deterministic and stochastic parameter estimation. From a deterministic point of view the new approach is an improvement because it effectively expands Tikhonov regularization in the two-norm in such a way that the regularization parameter can vary along a diagonal to accurately

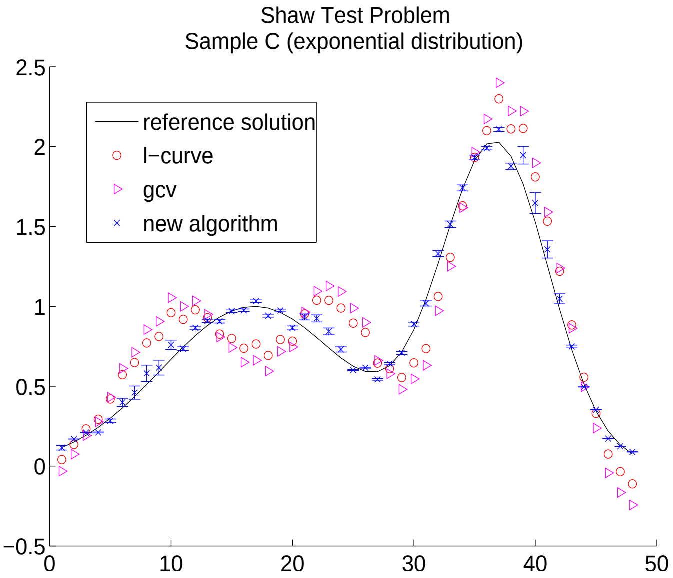

Figure 9: Sample parameter estimates for the Shaw test problem. Estimates are found by (i) the L-curve, (ii) generalized cross-validation and (iii) the proposed algorithm in Table 1 which also has error bars. The data noise are from exponential distributions.

weight initial parameter misfits. The benefits from a stochastic point of view are that with this approach, a priori information about the parameters is not needed nor is it necessary to assume normally distributed data or parameters. Rather than identify a priori distributions of the parameters, the parameter misfit weight is found by ensuring that the cost function, a sum of weighted data and parameter misfits, is a χ^2 random variable with n degrees of freedom.

Optimization is done in least squares sense however, we find that if the misfits are accurately weighted, the parameter estimates are not smoothed. This was shown by both solving benchmark problems in parameter estimation, and by investigating the calculated weights on the initial parameter estimates in a simpler, idealized problem.

In the benchmark problems the proposed algorithm typically gave better parameter estimates than those found from the L-curve and generalized cross validation. In the cases for which the proposed algorithm did not perform better, corresponding error bars or uncertainty estimates correctly identify the error.

The goal of the proposed algorithm is to accurately weight initial parameter misfits in a least squares minimization problem. Optimal weights will be dense weighting matrices which appropriately identify initial parameter misfit errors, and their correlations. Thus future work involves finding dense weighting matrices, rather than diagonal matrices, in addition to improving the optimization routine. Limitations of the algorithm include the need for good initial parameter estimates, and the computational time of the optimization problem in Table 1.

7. References

References

1. Bennett A 2005 *Inverse Modeling of the Ocean and Atmosphere* (Cambridge University Press) p 234
2. Casella G and Berger R 2001 *Statistical Inference* (California: Duxbury) 688p
3. Chandrasekaran S, Golub G, Gu M and Sayed A 1998 Parameter estimation in the presence of bounded data uncertainties *SIMAX* **19** 235-52
4. Golub G Hansen P and O'Leary D 1999 Tikhonov Regularization and Total Least Squares *SIAM J. Matrix. Anal. App.* **21** 185-94
5. Hansen P 1994 Regularization Tools: A Matlab Package for Analysis and Solution of Discrete Ill-posed Problems *Numerical Algorithms* **6** 1-35
6. Phillips D 1962 A technique for the numerical solution of certain integral equations of the first kind *J. ACM* , **9** 84-97
7. Huang J, Liu N, Pourahmadi M and Liu L 2006 Covariance matrix selection and estimation via penalised normal likelihood *Biometrika* **93** 85-98
8. Hansen P 2002 Analysis of discrete ill-posed problems by means of the L-curve *SIAM Review* **34** 561-80.
9. Hansen P 1998 *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion* (SIAM Monographs on Mathematical Modeling and Computation 4) p 247
10. Hansen P 1994 Regularization Tools A Matlab Package for Analysis and Solution of Discrete Ill-Posed Problems *Num. Alg* **6** 1-25
11. Menke W 1989 *Geophysical Data Analysis: Discrete Inverse Theory* (San Diego: Academic Press) p 289
12. Morozov 1984 *Methods for Solving Incorrectly Posed Problems* (New York: Springer Verlag)
13. Scales J and Tenorio L 2001 Prior Information and Uncertainty in Inverse Problems *Geophysics* **66** 389-97
14. Scales J Snieder R 1997 To Bayes or not to Bayes? *Geophysics* **63** 1045-46
15. Sancevero S Remacre A, and Portugal R 2005 Comparing deterministic and stochastic seismic inversion for thin-bed reservoir characterization in a turbidite synthetic reference model of Campos Basin, Brazil *The Leading Edge* 1168-72.
16. Tarantola A 2005 *Inverse Problems Theory and Methods for Model Parameter Estimation* (SIAM) p 342
17. Tikhonov A and Arsenin V 1977 *Solutions of Ill-Posed Problems* (New York: Wiley) p 272

Received

Author information

J.L. Mead , Department of Mathematics, Boise State University, Boise, ID 83725-1555, USA.
Email: jmead@boisestate.edu