Discontinuous parameter estimates with least squares estimators

J.L. Mead

Boise State University, Department of Mathematicss, Boise, ID 83725-1555. Tel: 208426-2432, Fax: 208-426-1354
Email jmead@boisestate.edu.

Abstract

We discuss weighted least squares estimates of ill-conditioned linear inverse problems where weights are chosen to be inverse error covariance matrices. Least squares estimators are the maximum likelihood estimate for normally distributed data and parameters, but here we do not assume particular probability distributions. Weights for the estimator are found by ensuring its minimum follows a χ^2 distribution. Previous work with this approach has shown that the it is competitive with regularization methods such as the L-curve and Generalized Cross Validation (GCV) [23]. In this work we extend the method to find diagonal weighting matrices, rather than a scalar regularization parameter. Diagonal weighting matrices are advantageous because they give *piecewise* smooth least squares estimates and hence are a mechanism through which least squares can be used to estimate discontinuous parameters. This is explained by viewing least squares estimation as a constrained optimization problem. Results with diagonal weighting matrices are given for a benchmark discontinuous inverse problem from [15]. In addition, the method is used to estimate soil moisture from data collected in the Dry Creek Watershed near Boise, Idaho. Parameter estimates are found that combine two different types of measurements, and weighting matrices are found that incorporate uncertainty due to spatial variation so that the parameters can be used over larger scales than those that were measured.

Keywords: Least squares, covariance, regularization, hydrology

1. Introduction

Inverse problems arise when information about a system or inputs into a system are inferred from measurements. For example, parameters in a geophysical model, atmospheric event, biological or chemical processes, material or electrical properties or in constructing an image. These problems are typically iill-conditioned when the underlying infinite dimensional problem is ill-posed, or when the system cannot be resolved by the data.

Regularization is an approach to ill-posed or ill-conditioned inverse problems where a related problem is constructed with a solution estimate that is close to the solution of the original problem. Tikhonov regularization [39] is the most common approach while more recently total variation, first introduced in [32], has been widely used and analyzed. Quantitative estimates of the distance between the regularized problem and the original problem are well known for quadratic regularizers while some for total variation are given in [6]. Alternatively, regularized solutions can be viewed as solutions of the original problem if we regularize with constraints on possible states of the problem [3]. In other words, we can attempt to resolve the ill-posedness in the problem, and regularize it, by adding more information to the problem. This is the view of regularization we take in this work.

We describe our approach on discrete linear inverse problems of the form

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

where the goal is to recover parameters $\mathbf{x} \in \mathbb{R}^n$ from measurements $\mathbf{b} \in \mathbb{R}^m$ based on model $\mathbf{A} \in \mathbb{R}^{mxn}$, which is an ill-conditioned matrix. The matrix $\mathbf{A}$ may be from a linear problem, a linear

approximation to a nonlinear problem, or this linear system may be solved as part of a nonlinear algorithm. The problem may be viewed stochastically if we consider that measurements have noise $\boldsymbol{\epsilon}$ and the linear system becomes:

$$\mathbf{A}\mathbf{x} = \mathbf{b} + \boldsymbol{\epsilon}. \tag{1}$$

Solution of the linear system requires initial parameter estimates $\mathbf{x}_0$ which also have uncertainty $\mathbf{f}$, thus

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{f}. \tag{2}$$

Practitioners may supply $\mathbf{x}_0$ and statistical properties of $\mathbf{f}$, but more often $\mathbf{x}_0$ is taken to be $\mathbf{0}$ with unknown error statistics $\mathbf{f}$. This is an example of how including more information in the problem can be viewed as regularizing the problem. In this case, we find the parameter estimate $\hat{\mathbf{x}}$ that minimizes of the sum of the weighted errors $\boldsymbol{\epsilon}$ and $\mathbf{f}$:

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} ||\mathbf{W}_\epsilon(\mathbf{b} - \mathbf{A}\mathbf{x})||_{p_1} + ||\mathbf{W}_f(\mathbf{x} - \mathbf{x}_0)||_{p_2}. \tag{3}$$

If the problem is regularized with a first or second order derivative, the matrix $\mathbf{W}_f$ can weight the error in the initial derivative estimate rather than initial parameter estimate.

The choice of weights $\mathbf{W}_\epsilon$ and $\mathbf{W}_f$ significantly effects the estimate $\hat{\mathbf{x}}$, as does the choice of norm p_1 and p_2. Absolute deviation and total variation are the case when $p_2 = 1$ [2, 32], and this choice gives less weight to outliers than when $p_2 > 1$. Statistically, $p_2 = 1$ gives the most likely solution if the parameters are exponentially distributed [24]. Methods with this norm, for example LASSO, are often used for discontinuous problems in order to preserve edges of non-smooth solutions, but they can be computational complex because the functional is not differentiable at points for which $\mathbf{x}_0 = \mathbf{x}$.

This simplest choice of norm is $p_1 = p_2 = 2$ because the solution estimate $\hat{\mathbf{x}}$ can be given explicitly. This is generalized Tikhonov regularization [39], or ridge regression [19] and is the most likely solution when the data and parameters are normally distributed. However, when $\mathbf{W}_\epsilon$ or $\mathbf{W}_f$ is a constant times the identity matrix, large weight is placed on the error outliers and the estimate $\hat{\mathbf{x}}$ can be unecessarily smooth. Smoothness may be desirable if the data are too noisy, however, some important information in the noise can be lost. Optimally, an algorithm should reduce the oscillatory noise while preserving edges. If the solution contains sharp edges or fronts one approach is to take p_2 as close to 1 as possible, or find some approximation to it [14].

In this work we explore taking the simplest choice $p_1 = p_2 = 2$ and propose that since weighting matrices $\mathbf{W}_\epsilon$ or $\mathbf{W}_f$ can be chosen so that the estimate $\hat{\mathbf{x}}$ is piecewise smooth, $\hat{\mathbf{x}}$ can be discontinuous. Matrix weights allow accurate weighting of outliers in $\boldsymbol{\epsilon}$ or $\mathbf{f}$ so that their squared value does not inappropriately influence the solution estimate. The goal of this work is to develop an approach to estimate good weighting matrices.

In Section 2 we show how discontinuous parameters can be found with the least squares estimator in two dimensions by viewing regularization as constrained optimization. In Section 3 we describe a method to find weights which give piecewise smooth least squares estimates. This method is an extension of the χ^2 method [21, 22], and we illustrate its effectiveness in finding discontinuous parameters by solving a benchmark ill-conditioned problem in [15]. In Section 4 we use the method to find matrix weights that quantify uncertainty when estimating water content in soils. Results are shown for soil moisture estimation with the van Genuchten equation [40] in the Dry Creek Watershed near Boise, Idaho. The resulting parameters and their uncertainty can be used in large scale processes.

2. Regularization and Constrained Optimization

In order to see that appropriate weighting matrices can give non-smooth least squares estimates we write (3) with $p_1 = p_2 = 2$ as

$$\min \quad \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} A_{ij} x_j \right)^2 \tag{4}$$

$$\text{s.t} \qquad \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_0)_i^2 \leq t. \tag{5}$$

The solution is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} A_{ij} x_j \right)^2 + \lambda \sum_{i=1}^{n} (\mathbf{x} - \mathbf{x}_0)_i^2 \right\}$$

where λ represents the Lagrange multiplier for the constraint. This solution is equivalent to the solution from Tikhonov regularization with $\mathbf{W}_\epsilon = \mathbf{W}_f = \mathbf{I}$. Alternatively, the choice $p_1 = 2$, $p_2 = 1$, least squares with absolute deviation, has the same objective function (4) but the constraint is now

$$\text{s.t.} \quad \sum_{i=1}^{n} |\mathbf{x} - \mathbf{x}_0|_i \leq t \tag{6}$$

and the solution is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} A_{ij} x_j \right)^2 + \lambda \sum_{i=1}^{n} |\mathbf{x} - \mathbf{x}_0|_i \right\}.$$

Figure 1 contains contour plots of the objective function (4) and both constraints in two dimensions. Figure 1(a) shows the constraint (5) while Figure 1(b) shows (6). The contours of the objective function form ellipses, while the constraints form a circle in (a) and a diamond in (b). The solution of the constrained optimization problem is the lowest point $(\hat{x}, \hat{y})$ where the boundary of the constraint region intersects the lowest level contour of the objective function. In (a) where $p_2 = 2$ the geometry of the intersection of the circle and ellipse is smooth, and this is why least squares approaches produce smooth solutions. Alternatively in (b), when $p_2 = 1$, there are points in the constraint region that are perpendicular to the objective function, thus absolute deviation can produce non-smooth solutions [28].

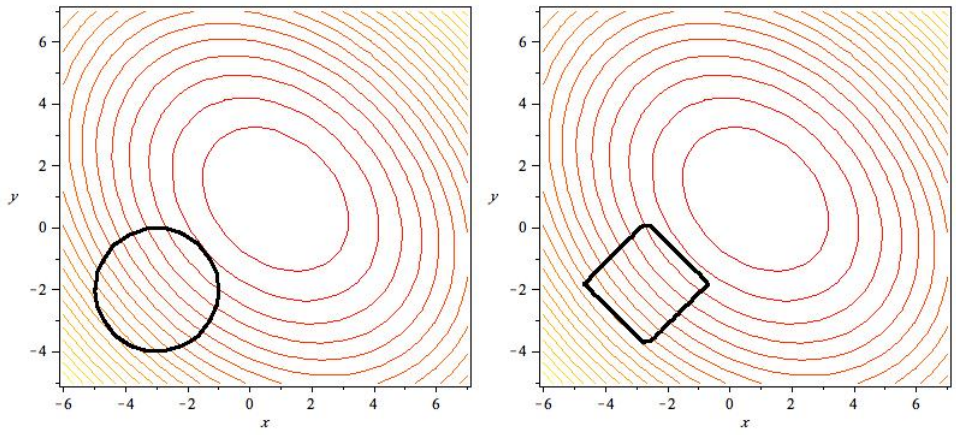
When $p_2 = 2$ we can turn the circle into a square, and hence introduce the ability to obtain non-smooth least squares solutions, by applying multiple constraints. The objective function remains (4) but we apply n constraints

$$\text{s.t.} \quad (\mathbf{x} - \mathbf{x}_0)_i^2 \leq t\sigma_i^2 \qquad i = 1, \ldots, n. \tag{7}$$

The solution is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\{ \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} A_{ij} x_j \right)^2 + \sum_{i=1}^{n} \lambda_i (\mathbf{x} - \mathbf{x}_0)_i^2 \right\} \tag{8}$$

where now there are n Lagrange multipliers $\lambda_1, \ldots, \lambda_n$. Multiple constraints in two dimensions are plotted in Figure 2. The solution must lie in the square formed by the two constraints, thus the constraint region can also be perpendicular to contours of the objective function and non-smooth estimates can be obtained with multiple quadratic constraints. The potential for non-smooth solutions in higher dimensions is significant, because computationally the solution when $p_2 = 2$ is significantly simpler than when $p_2 = 1$. However, this approach relies on calculating n Lagrange multipliers.



(a) (b)

Figure 1: 2-D Illustration of constrained optimization of (4), (a) quadratic constraint (5) (b) absolute constraint (6).

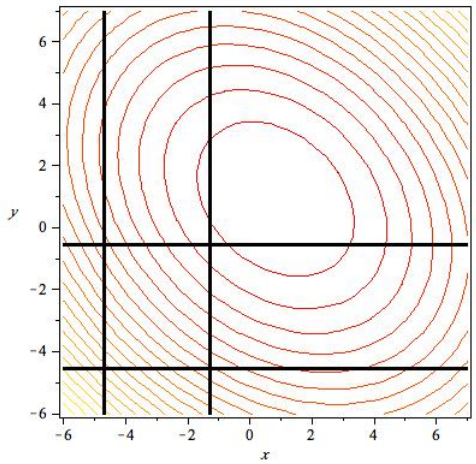


Figure 2: 2-D Illustration of constrained optimization of (4) with multiple quadratic constraints (7).

The approach we take to calculate these multipliers uses the fact that the solution (8) may be interpreted as the Maximum *a Posteriori* or MAP estimate [9]. If weights $\mathbf{W}_\epsilon$ and $\mathbf{W}_f$ are chosen to be error covariance matrices $\mathbf{C}_\epsilon = cov(\boldsymbol{\epsilon})$ and $\mathbf{C}_f = cov(\mathbf{f})$, the least squares estimate is

$$\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \left\{ (\mathbf{b} - \mathbf{A}\mathbf{x})^T \mathbf{C}_\epsilon^{-1} (\mathbf{b} - \mathbf{A}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_0)^T \mathbf{C}_f^{-1} (\mathbf{x} - \mathbf{x}_0) \right\}.$$

This is equivalent to the solution of the constrained optimization problem (4) and (8) with multiple constraints if we rescale the problem to

$$\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}} + \tilde{\boldsymbol{\epsilon}} \tag{9}$$

where the tilde ˜ represents pre-multiplication of the original variable by $\mathbf{C}_\epsilon^{-1/2}$, e.g. $\tilde{\mathbf{A}} = \mathbf{C}_\epsilon^{-1/2}\mathbf{A}$. In addition, we assume there is no correlation in the errors, i.e.

$$\begin{aligned} \mathbf{C}_\epsilon &= \text{diag}((\sigma_1^\epsilon)^2, \ldots, (\sigma_m^\epsilon)^2) \\ \mathbf{C}_f &= \text{diag}((\sigma_1^f)^2, \ldots, (\sigma_n^f)^2) \end{aligned}$$

with $(\sigma_i^\epsilon)^2 = var(\epsilon_i)$ and $(\sigma_i^f)^2 = var(f_i)$, and $\sigma_i^f = 1/\lambda_i$. This statistical interpretation of the weights in (3) and multipliers in (8) gives us a method for calculating them, and we term it the χ^2 method for parameter estimation and uncertainty quantification [21, 22, 23, 31].

3. χ^2 Method

The χ^2 method can be used to estimate weights for the initial parameter or data misfits in (3) when $p_1 = p_2 = 2$. It is based on the assumption that the matrix weights $\mathbf{W}_\epsilon$ and $\mathbf{W}_f$ are inverse covariances matrices $\mathbf{C}_\epsilon^{-1}$ and $\mathbf{C}_f^{-1}$, respectively, however, the specific probability distributions of the errors $\boldsymbol{\epsilon}$ and $\mathbf{f}$ are not needed or calculated. This makes the method more efficient than statistical or Bayesian approaches to inverse problems. The interpretation of the weighting matrices as inverse covariances is intuitive even in a non-stocastic setting because the errors in initial estimates are penalized according to their inverse variance, while scalar weights give all errors the same penalty. This means that the parameter estimate $\hat{\mathbf{x}}$ in the χ^2 method is the one that minimizes the errors in a weighted least squares sense, where the weights are chosen to be the inverse covariance matrices for the errors, i.e $\hat{\mathbf{x}}$ minimizes the functional

$$\mathcal{J}(\mathbf{x}) = (\mathbf{b} - \mathbf{A}\mathbf{x})^T \mathbf{C}_\epsilon^{-1} (\mathbf{b} - \mathbf{A}\mathbf{x}) + (\mathbf{x} - \mathbf{x}_0)^T \mathbf{C}_f^{-1} (\mathbf{x} - \mathbf{x}_0). \tag{10}$$

Even though we use least squares, we do not assume the data or parameter errors are normally distributed, thus this is an M-estimator [11].

3.1. Relationship to Discrepancy Principle

If we have an estimate of the measurement noise σ_ϵ, Morozov's discrepancy principle [26] can be used to choose the regularization parameter $\alpha > 0$ in

$$\hat{\mathbf{x}}_\alpha = \arg\min_{\mathbf{x}} ||\mathbf{b} - \mathbf{A}\mathbf{x}||_2^2 + \alpha ||\mathbf{x} - \mathbf{x}_0||_2^2.$$

It is based on the residual principle in the following theorem.

Theorem 1. *(Discrepancy Principle [29]) Assume $||\mathbf{b}||_2 > \delta$. If α is the root of*

$$||\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_\alpha||_2 = ||\mathbf{b} - \mathbf{A}(\mathbf{A}^T\mathbf{A} + \alpha\mathbf{I})^{-1}\mathbf{A}^T\mathbf{b}||_2 = C\delta$$

where $C > 1$ is a constant, then

$$\lim_{\delta \to 0} ||\hat{\mathbf{x}}_\alpha - \mathbf{x}|| = 0$$

This theorem states that the regularized estimate $\hat{\mathbf{x}}_\alpha$ will tend towards the true solution $\mathbf{x}$ as the size of the noise goes to zero. In real applications the noise does not typically go to zero, and the discrepancy principle is implemented in the following manner: choose α so that

$$||\mathbf{b} - \mathbf{A}\hat{\mathbf{x}}_\alpha||_2^2 = ||\boldsymbol{\epsilon}||_2^2 \approx m\sigma_\epsilon^2.$$

If the problem is re-scaled as in (9) so that $||\tilde{\boldsymbol{\epsilon}}||_2 = 1$ then the discrepancy principle is implemented as

$$||\tilde{\mathbf{b}} - \tilde{\mathbf{A}}\hat{\mathbf{x}}_\alpha||_2^2 \approx m. \tag{11}$$

In the following Theorem and Corollary we give some well known properties about least squares functionals as random variables, see for example [1]. They will be used to show that even though the residual principle holds as the noise goes to zero, (11) may not be a good criteria from which to choose the regularization parameter.

Theorem 2. *Assume the elements of* $\mathbf{b}$ *and* $\mathbf{x}$ *are independent and identically distributed random variables with covariance matrices* $\mathbf{C}_\epsilon$ *and* $\mathbf{C}_f$*, respectively.*

Part A: Let $\mathcal{J}_{obs}(\mathbf{x}) = ||\tilde{\mathbf{b}} - \tilde{\mathbf{A}}\mathbf{x}||_2^2$*. Then* $\mathcal{J}_{obs}(\mathbf{x})$ *is a* χ^2 *random variable with* m *degrees of freedom if the elements of* $\mathbf{b}$ *are normally distributed.*

Part B: Let $\mathcal{J}(\mathbf{x})$ *be as defined by (10). Then* $\mathcal{J}(\mathbf{x})$ *is a* χ^2 *random variable with* $m+n$ *degrees of freedom if the elements of* $\mathbf{b}$ *and* $\mathbf{x}$ *are normally distributed.*

Since the expected value of a χ^2 random variable is equal to its degrees of freedom, the discrepancy principle could be viewed as using Part A of Theorem 2 to find the regularization parameter. However, if the functionals in Parts A and B are evaluated at parameter estimates found with the data, the degrees of freedom in the resulting χ^2 variable are reduced by the number of parameters as stated in the following Corollary.

Corollary 3. *If the same assumptions hold as in Theorem 2 then*

Part A: Let $\hat{\mathbf{x}}_{obs} = \arg\min_{\mathbf{x}} \mathcal{J}_{obs}(\mathbf{x})$*, then* $\mathcal{J}_{obs}(\hat{\mathbf{x}}_{obs})$ *is a* χ^2 *random variable with* $m-n$ *degrees of freedom if the elements of* $\mathbf{b}$ *are normally distributed.*

Part B: Let $\hat{\mathbf{x}} = \arg\min_{\mathbf{x}} \mathcal{J}(\mathbf{x})$ *as defined by (10). Then* $\mathcal{J}(\hat{\mathbf{x}})$ *is a* χ^2 *random variable with* m *degrees of freedom if the elements of* $\mathbf{b}$ *and* $\mathbf{x}$ *are normally distributed.*

Corollary 3 follows from the fact that $\mathcal{J}_{obs}(\hat{\mathbf{x}}_{obs})$ and $\mathcal{J}(\hat{\mathbf{x}})$ can be written in quadratic form and hence follow a generalized χ^2 distribution [7]. Lastly, it was shown that the results in Theorem 2 and Corollary 3 still hold for non-normally distributed $\mathbf{x}$ and $\mathbf{b}$ in [21], as stated in the following Corollary.

Corollary 4. *If the number of data* m *is large and the same assumptions hold as in Theorem 2, then the asymptotic distributions of* $\mathcal{J}_{obs}(\mathbf{x})$*,* $\mathcal{J}_{obs}(\hat{\mathbf{x}}_{obs})$*,* $\mathcal{J}(\mathbf{x})$ *and* $\mathcal{J}(\hat{\mathbf{x}})$ *are still* χ^2 *with* m*,* $m-n$*,* $m+n$ *and* m *degrees of freedom, respectively, regardless of the distribution of* $\mathbf{b}$ *and* $\mathbf{x}$*.*

Part A of Corollary 3 states that if the number of data is equal to the number of unknowns, we expect that $\mathcal{J}_{obs}(\hat{\mathbf{x}}_{obs}) \approx m - n = 0$, while the discrepancy principle has $\mathcal{J}_{obs}(\hat{\mathbf{x}}_\alpha) \approx m$. Even though $\hat{\mathbf{x}}_{obs}$ and $\hat{\mathbf{x}}_\alpha$ are not equivalent, the degrees of freedom in $\mathcal{J}_{obs}(\hat{\mathbf{x}}_\alpha)$ still must be reduced by the number of parameters since the data were used to find $\hat{\mathbf{x}}_\alpha$. The degrees of freedom must be reduced when the parameter estimates are found with the data because the terms in the χ^2 sum are no longer independent [10].

Similar to the discrepancy principle, the χ^2 method uses the fact that the expected value of a least squares functional is the number of degrees of freedom in the corresponding χ^2 random variable, but now condition (11) is replaced with the regularized residual i.e.

$$||\tilde{\mathbf{b}} - \tilde{\mathbf{A}}\hat{\mathbf{x}}||_2^2 + ||\mathbf{C}_f^{-1/2}(\hat{\mathbf{x}} - \mathbf{x}_0)||_2^2 \approx m. \tag{12}$$

This condition is consistent with Part B of Corollary 3. Algorithms for the χ^2 method when $\mathbf{C}_f = \sigma_f^2\mathbf{I}$ and $\mathbf{C}_\epsilon = \sigma_\epsilon^2\mathbf{I}$ are given in [22, 31], and we refer to this case as the scalar χ^2 method. This is Tikohonv regularization [39] and it was shown that the scalar χ^2 method is an attractive alternative to the L-curve [13], Generalized Cross Validation (GCV) [43] and Unbiased Predictive Risk Estimation (UPRE).

3.2. χ^2 Method for diagonal matrix weights

The scalar χ^2 method is based on solving the single equation (12) for the regularization parameter $\alpha = \sigma_f^{-2}$ ($\mathbf{C}_f = \sigma_f^2\mathbf{I}$) or to estimate the standard deviation in the data error σ_ϵ ($\mathbf{C}_\epsilon = \sigma_\epsilon^2\mathbf{I}$). Since the methodology to compute $\mathbf{C}_\epsilon$ or $\mathbf{C}_f$ is the same, we will use the more general notation $\mathbf{C}$ and σ to represent $\mathbf{C}_f$ or $\mathbf{C}_\epsilon$, and σ_f or σ_ϵ, respectively. To estimate more dense matrices $\mathbf{C}$, we now specify a system of N equations from which we can solve for $\mathbf{C} = \mathrm{diag}(\sigma_1, \ldots, \sigma_N)$. The system represents multiple χ^2 tests and based on the following theorem.

Theorem 5. *Let $\mathbf{P} = \mathbf{A}\mathbf{C}_f\mathbf{A}^T + \mathbf{C}_\epsilon$, $\mathbf{r} = \mathbf{A}\mathbf{x}_0 - \mathbf{b}$, and $\mathbf{k} = \mathbf{P}^{-1/2}\mathbf{r}$ so that $\mathcal{J}(\hat{\mathbf{x}}) = \mathbf{r}^T\mathbf{P}^{-1}\mathbf{r} = k_1^2 + \ldots k_m^2$. If $\boldsymbol{\epsilon}$ and $\mathbf{f}$ are independent and identically normally distributed with covariance matrices $\mathbf{C}_\epsilon$ and $\mathbf{C}_f$, respectively, then $k_{p_{i-1}+1}^2 + \ldots + k_{p_i+p_{i-1}}^2$, with $p_0 = 0$ and $\sum_{i=1}^N p_i = m$, $N \leq m$, follows a χ^2 distribution with p_i degrees of freedom for $i = 1, \ldots, N$.*

Proof. Since $\mathbf{C}_f$ and $\mathbf{C}_\epsilon$ are covariance matrices, $\mathbf{P}$ is symmetric positive definite and we can define $\mathbf{k} = \mathbf{P}^{-1/2}\mathbf{r}$. It was shown in [21] that $\mathcal{J}(\hat{\mathbf{x}}) = \mathbf{r}^T\mathbf{P}^{-1}\mathbf{r} = k_1^2 + \ldots k_m^2$. When data errors are normally distributed, each k_i is also normally distributed since it is the linear combination of m normal variables. Hence each k_i^2 follows a χ^2 distribution with 1 degree of freedom. Since there are m elements in $\mathbf{k}$, we can form $N \leq m$ different sums, and each sum will have p_i degrees of freedom, where p_i is the number of terms in the sum.

We write the series of sums of squares in Theorem 5 as an N dimensional linear system

$$\begin{aligned} k_1^2 + \ldots + k_{p_1}^2 &= p_1, && (13)\\ k_{p_1+1}^2 + \ldots + k_{p_2+p_1}^2 &= p_2, && (14)\\ &\vdots && (15)\\ k_{p_{N-1}+1}^2 + \ldots + k_{p_N+p_{N-1}}^2 &= p_N. && (16) \end{aligned}$$

This gives us N equations from which we can solve for N unknowns in the weighting matrix $\mathbf{C}^{-1}$ (i.e. either $\mathbf{C}_f^{-1}$ or $\mathbf{C}_\epsilon^{-1}$ in (10)). This system (13)-(16) is coupled because each k_i is a function of $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_N)$, and we will denote it by $\mathbf{F}(\boldsymbol{\sigma}) = \mathbf{0}$. We note that N does not necessarily equal the number of parameters n or data m. When it does, each data or initial parameter misfit has different weight.

Corollary 6. *As the number of data increases, i.e. $m \to \infty$, Theorem 5 still holds as a limiting χ^2 distribution.*

Proof. It is important to note that as we take fewer sum of squares of k_i, we do not violate the fact that the distribution of each k_i are approximately normal when m is large, regardless of the distribution of data errors. This can be seen by noting that if data errors are not normal, each k_i is the linear combination of m random errors r_i, i.e.

$$k_i = \sum_{j=1}^{m} P_{ij}^{-1/2}\mathbf{r}_j = P_{i1}^{-1/2}r_1 + \ldots + P_{im}^{-1/2}r_m, \quad i = 1, \ldots, m. \tag{17}$$

The Central Limit Theorem [10] states that each k_i, $i = 1, \ldots, m$, will have a distribution that tends to the normal distribution as m increases regardless of the probability distribution of r_i. Thus when the number of data m is large, we can assume that each k_i is normally distributed.

3.3. Benchmark discontinuous problem

Errors in data or initial parameter estimates can be grouped to give $\mathbf{C}$ block diagonal structure. As an example, we solved the discontinuous *wing* test problem from [15] with $m = n = 96$, and considered $\mathbf{C} = \mathbf{C}_f$ with three diagonal blocks. The *wing* test problem is a one-dimensional inverse problem with a discontinuous solution arising from the discretization of a Fredholm integral equation of the first kind [45]. The goal is to find $f(s)$ given the Kernel $K(t, s)$ and data $d(t)$:

$$d(t) = \int_a^b K(t, s) f(s) ds.$$

Small errors in data d are greatly amplified in the solution f, thus the problem is ill-posed and it's discretization results in an ill-conditioned linear system of equations. The accuracy of the one dimensional results from *wing* are viewed by creating a "true" or mean solution $\mathbf{x}$, artificially adding noise $\boldsymbol{\epsilon}$ to data created by it, and then comparing the computed estimate $\hat{\mathbf{x}}$ to $\mathbf{x}$.

In Figure 3 the noisy data from which the estimate was obtained is plotted on the left, and the estimated parameters along with the mean or true solution are plotted on the right. The noise in the data were taken from a normal distribution, while the initial parameter estimate was taken to be zero. The 'o' curve is the estimate when $\mathbf{C}_f = \sigma_f \mathbf{I}$ while the '+' curve is the estimate with $\mathbf{C}_f = \text{diag}(\sigma_1, \ldots, \sigma_1, \sigma_2, \ldots, \sigma_2, \sigma_3, \ldots, \sigma_3)$. Since there are two discontinuities we chose $N = 3$ with $p_i = 96/3$, $i = 1, 2, 3$, so that $\mathbf{C}_f$ has three diagonal blocks. We see that the estimate with the matrix weight captures the discontinuity in the true solution at 33 and 65, while the scalar weight smoothes the estimate at those points. In addition, since the χ^2 tests apply to non-normally distributed data, we illustrate the estimate when the data are from an exponential distribution in Figure 4. We see there that the discontinuity is again captured when the parameter misfit is weighted with a matrix with three diagonal blocks.

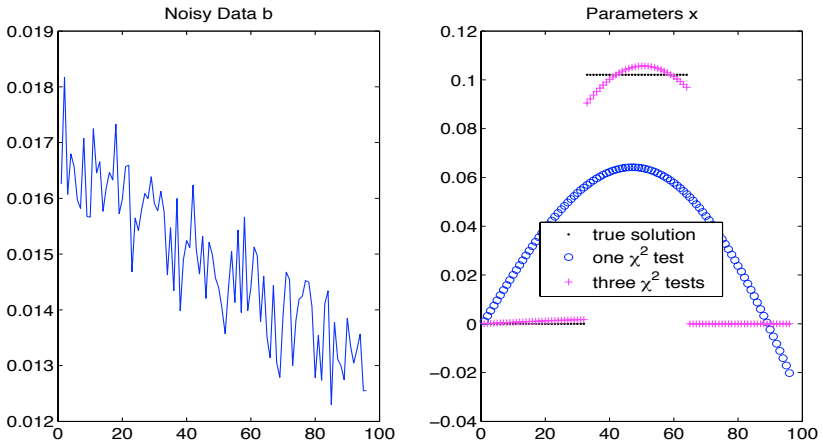


Figure 3: Benchmark problem *wing* [15] with discontinuous solution and normally distributed data.

The roots of $\mathbf{F}(\boldsymbol{\sigma})$ with $\mathbf{F}, \boldsymbol{\sigma} \in \mathbb{R}^N$ given by (13)-(16) were found in Matlab using *sqrtm* to calculate $\mathbf{P}^{1/2}$ and *fsolve* to solve the nonlinear system. Since $\mathbf{P}$ is symmetric positive definite there is a unique positive square root. The Matlab function *sqrtm(P)* calculates this from the Schur deomposition $\mathbf{P} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$ so that $\mathbf{P}^{1/2} = \mathbf{Q}\mathbf{D}^{1/2}\mathbf{Q}^T$ [17]. The Matlab function *fsolve* in the Optimization Toolbox uses nonlinear least squares methods based on reflective newton methods [8].

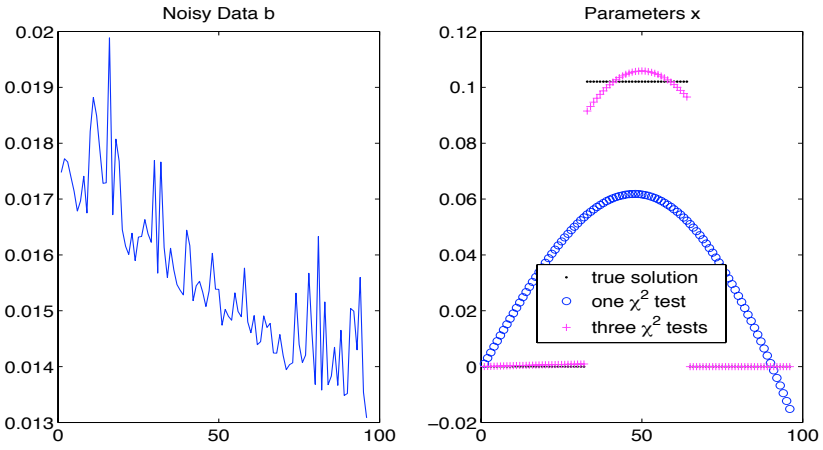


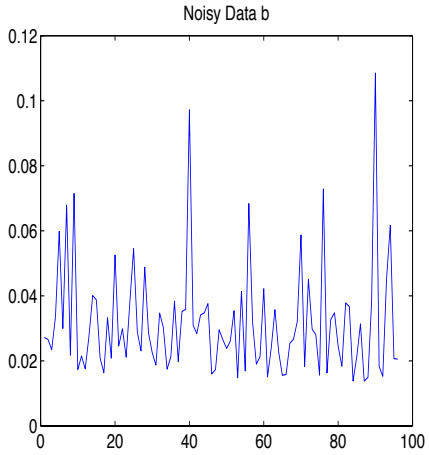
Figure 4: Benchmark problem *wing* [15] with discontinuous solution and exponentially distributed data.

3.4. Variable χ^2 method

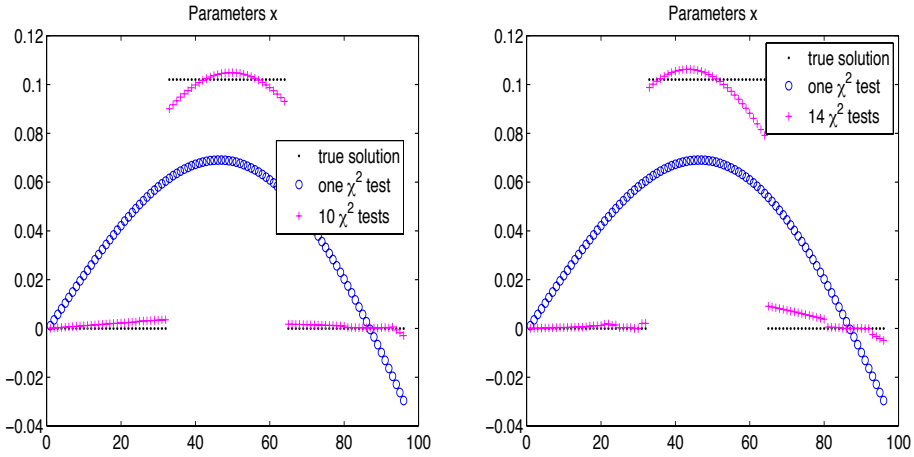
The algorithm using *fsolve* and *sqrtm* does not always converge. In addition, the results obtained in Figures 3-4 occured with knowledge of the location of the discontinuities in the solution. When the location of discontinuities or large variations in the solution are not known, we developed the variable χ^2 method to determine the number of χ^2 tests needed in a given region. The procedure begins with one χ^2 test (12) and if it is not satisfied within a given tolerance, the problem is split in two and $\mathbf{F}(\boldsymbol{\sigma}) = \mathbf{0}$ is solved with $N = 2$. This process continues and the number of χ^2 principles in regions where an estimate cannot be found is doubled while the estimate is held fixed in regions where the nonlinear system is satisfied. For example, in (13)-(16) when we cannot find $\boldsymbol{\sigma}$ such that $k^2_{p_1+1}+\ldots+k^2_{p_2+p_1} = p_2$, but we can satisfy all other equations, we change the N dimensional system to the following system of N+1 equations:

$$\begin{aligned}
k_1^2 + \ldots + k_{p_1}^2 &= p_1, && (18)\\
k_{p_1+1}^2 + \ldots + k_{p_2/2+p_1}^2 &= p_2/2, && (19)\\
k_{p_2/2+1}^2 + \ldots + k_{p_2+p_2/2}^2 &= p_2/2, && (20)\\
k_{p_2+1}^2 + \ldots + k_{p_3+p_2}^2 &= p_3, && (21)\\
&\vdots && (22)\\
k_{p_i+1}^2 + \ldots + k_{p_N+p_i}^2 &= p_N. && (23)
\end{aligned}$$

Results from the variable χ^2 method with *wing* test problem are shown in Figure 5. The data in this example were from a normal distribution, and are more noisy than those in Figure 3. The tolerance was set at $\chi^2 < 1$, but it was not met on all intervals. The iteration procedure was stopped in two places, after 10 and 14 χ^2 tests. The resulting vector $\mathbf{p}$ formed by the right hand side of the system (18)-(23) was $p = [16\ 8\ 8\ 32\ 16\ 8\ 1\ 1\ 2\ 4]$ and $p = [16\ 4\ 2\ 2\ 4\ 2\ 2\ 32\ 16\ 4\ 4\ 2\ 2\ 4]$, respectively. We can see that as the number of χ^2 tests increased, the first discontinuity in the solution is better resolved, but the estimate at the second discontinuity is worse. It is not clear from these results that adding more χ^2 tests would resolve the discontinuity better.



(a)



(b) (c)

Figure 5: Benchmark problem *wing* [15] with normally distributed data (a). Estimate with 10 χ^2 tests $p = [16\ 8\ 8\ 32\ 16\ 8\ 1\ 1\ 2\ 4]$ (b), and 14 χ^2 tests $p = [16\ 4\ 2\ 2\ 4\ 2\ 2\ 32\ 16\ 4\ 4\ 2\ 2\ 4]$.

Even though the estimate $\hat{\mathbf{x}}$ is not calculated during the iteration process, the solution of the nonlinear system which defines the χ^2 tests is very expensive and does not always converge. Future work involves finding more efficient, reliable methods for solutions of the system, and more robust methods for determining the appropriate number of terms in each χ^2 test. We now turn to an application, and we will see that, as is the case with many applications, there is some *a priori* knowledge of where to break up the χ^2 tests.

4. Soil hydraulic properties

In 1980 van Genuchten [40] introduced the widely used empirical equation for a continuous representation of soil moisture θ as a function of pressure head ψ:

$$\theta(\psi) \quad = \quad \theta_r + \frac{(\theta_s - \theta_r)}{(1 + |\alpha\psi|^n)^m} \qquad h < 0. \tag{24}$$

This equation, coupled with Mualem's representation [27] of hydraulic condicuivity, and Richards' equation [30] for unsaturated flow gives soil moisture estimates over time, and larger spatial scales than are given by point-scale measurements. However, since the parameters in van Genuchten's equation are determined at the point scale, it is not clear over what scales they can be used in Richards' equation [25]. The goal here is to find relevant parameters that can be used in a multiscale approach to Richards equation, such as [16], so that we can capture large-scale behavior without resolving the small-scale details.

The van Genuchten equation (24) contains five independent parameters: θ_r and θ_s are the residual and saturated water contents ($\text{cm}^3\text{cm}^{-3}$), respectively, while α (cm^{-1}), n and m (-), with m commonly set to $1 - 1/n$, are empirical fitting parameters. The parameters were estimated by inverting measurements of soil moisture θ as a function of pressure head ψ from both field and laboratory measurements of soil in the Dry Creek catchment near Boise, Idaho [20]. Direct field measurements of soil moisture content and pressure head are made at soil pits, and in addition, soil cores were taken to the laboratory. In the laboratory, the cores were analyzed by measuring percentages of sand, silt , and clay and bulk density of the soils. These measurements were used as inputs into *Rosetta* [36], a public domain code from the USDA Agricultural Research Service. Figure 6 contains plots of the field measurements and the laboratory curve generated by Rosetta at four different pits: SD5_15, NU10_15, SU10_20, and SU10_40, representing North or South facing slopes, up or down stream from a weir 5 or 10 *cm*, and at depths of 15, 20 or 40 *cm*.

In Figure 6 we see that the field measurements and Rosetta estimates of soil moisture as a function of pressure head are not in agreement. This is because Rosetta is based on neural network analyses of a wide range of soil types consisting of 2134 soil samples, of which the gravel content is not documented. However, the soil samples from the Dry Creek Watershed had appreciable gravel contents. In [12] they normalized the curve, but here we used the curve from Rosetta within the uncertainty ranges given by the software, and indicated by horizontal bars in the plots. Figure 6 also contains results from the χ^2 method. The χ^2 curve is the fit of the field and laboratory measurements, within the uncertainty ranges given by Rosetta, and that pass the single χ^2 test. This means that unlike regularization, here $\mathbf{C}_f$ is given by Rosetta, and we use the χ^2 method to find $\mathbf{C}_\epsilon$, the uncertainty in the data error.

4.1. Spatial variability and uncertainty quantification

It is difficult to define parameter sets over multiple scales because parameters vary due to spatial variability or heterogeneity of soil properties. Accounting for these variations in soil, and upscaling the parameters, has been addressed in a variety of ways, see for example [44] and [37]. However, if one parameter set is to be used over a large spatial scale, "effective parameters" are calculated which simulate field-scale behavior with a single soil hydraulic characteristic, see for example [37]

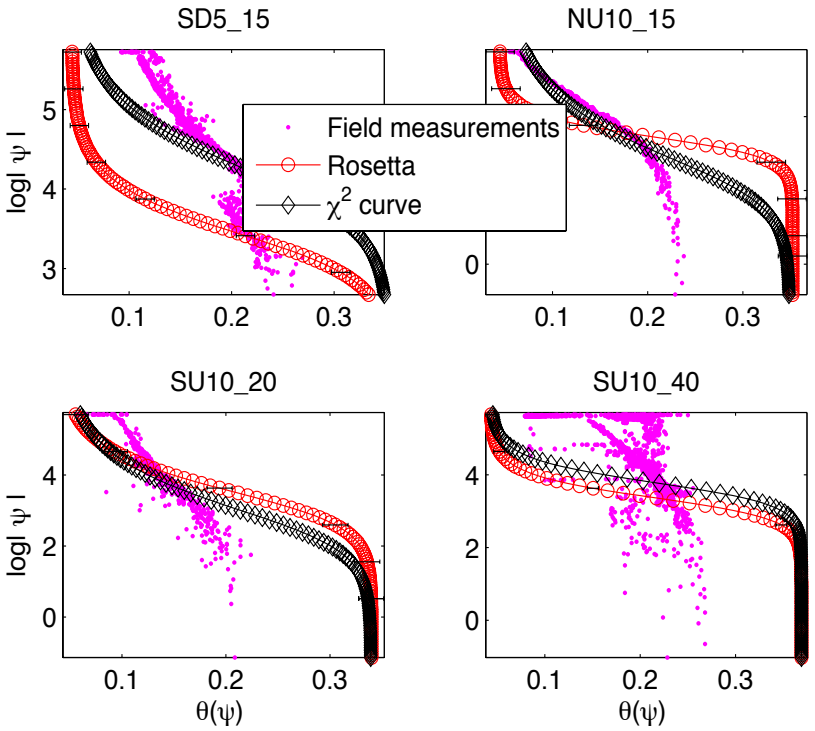


Figure 6: Soil water retention curves, $\theta(\psi)$, observed in the field, in the laboratory (Rosetta) and that found with χ^2 method.

NU10_15	SU10_20	SU10_40	SD5_15
0.005 (2.6%)	0.011 (6.4%)	0.044 (23.1%)	0.011 (6.4%)

Table 1: Single pit uncertainties calculated with the χ^2 method, and their percentage across all measurements.

	Depth SU10_20 & SU10_40	North/South slopes SU10_20 & NU10_15	Up/downstream SU10_20 & SD 5_15
χ^2 method	0.048 (22.5%)	0.014 (7.7%)	0.031 (14.5%)
Std. Dev.	0.017 (8.0%)	0.010 (5.2%)	0.011 (5.1%)

Table 2: Multiple pit uncertainties calculated with the χ^2 method, the standard deviation of all measurements, and the corresponding percentage each uncertainty is across all measurements.

and [46]. In this work we calculate effective parameters using the χ^2 method. The spatial variability in the parameter estimates is treated as uncertainty in data measurements and quantified with $\mathbf{C}_\epsilon$. This differs from previous approaches where the interest is in selecting specific parameters that reduce uncertainty in predictions [38]. In our approach, we vary the scale, calculate the data error uncertainty with the χ^2 method, and use that uncertainty estimate σ_ϵ to determine over what scales one parameter set should be used. These uncertainty estimates can also be used to identify sets of measurements that contain the most information for parameters identification, as in the PIMLI methodology [42].

4.2. Soil moisture results

4.2.1. Spatial variability

Data error uncertainty, $\mathbf{C}_\epsilon$, is typically attributed to measurement error but here we assume that it also includes uncertainty due to heterogeneties or spatial variability, and approximate it with the χ^2 method. The spatial variability of measurements, and curves resulting from parameter estimates, are illustrated in Figure 7. Starting clockwise from top left we combine measurements at two pits on North and South facing slopes, two pits up and downstream from the weir, two pits at different depths, and lastly we combine measurements from all four pits. The Rosetta curve is averaged over multiple pits while the χ^2 curve combines averaged data with averaged parameters and uncertainties from Rosetta. These curves do not look that different from each other, so the values of σ_ϵ are given in Tables 1-2.

Table 1 gives the uncertainty σ_ϵ found by the scalar χ^2 method, so that the combination of field measurements and Rosetta estimates pass the single χ^2 test. The corresponding number in parentheses represents the percentage this value is across all measurements, i.e $\sigma_\epsilon/(\max\theta - \min\theta)$. Table 2 gives similar σ_ϵ calculated by the χ^2 method, but for the measurements averaged across pits, as illustrated in Figure 7. Similar to Table 1, in parentheses we give the standard deviation as the average across all field measurements, but now across multiple pits. The values in these tables most likely underestimate the true uncertainties [34], however, their change in magnitude as we combine pits can be used to understand the effect of using point measurements across scales.

In Table 1 we see that the largest uncertainty comes from the pit at $40cm$ depth and it is 23.1% of the soil moisture measurements made across all pressure heads measurements at this pit. This is very high, but is typical of measurements in the sub-surface. The values in Table 1 are used as a baseline for determining the scales over which one parameter set can be used. The uncertainties typically increase as we combine data sets from multiple pits, as seen in Table 2. When the uncertainty

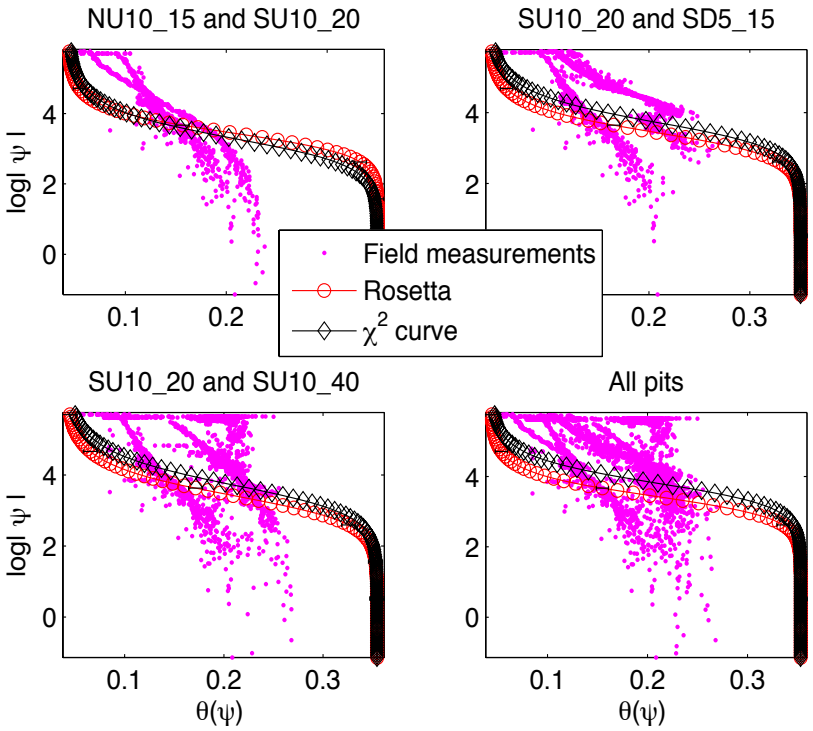


Figure 7: Soil water retention curves for multiple pits, $\theta(\psi)$, observed in the field, in the laboratory (Rosetta) and found with χ^2 method.

calculated with the χ^2 method is significantly higher with multiple pits than from a single pit, this indicates that a single parameter set should not be used for that particular pit combination.

In the first column of Table 2 we see that the χ^2 uncertainty after combining pits at depth is roughly the same percentage as that from the single pit at 40 *cm* depth in Table 1. This shows that the high uncertainty at depth 40 *cm* dominates the uncertainty at 20 *cm* depth, and hence different parameter sets should be used at these depths. In addition, we see that this uncertainty is significantly higher than the standard deviation of the measurements of both pits, which is 8%. The conclusion that we need to use different parameter sets at different depths is consistent with the flow modeling in [12], where they also found that they needed to use two different parameter sets.

Alternatively, when we combined measurements from the north and south facing slopes in column 2 of Table 2, we see that the χ^2 uncertainty does not change that much from the single pit uncertainty percentage in columns 1 and 2 of Table 1. This implies that we can use the same parameter set for north and south facing slopes. This is a surprising result since measurements on north and south facing slopes are often different in this region due to the high elevation and significant sunshine. However, we note that these parameters represent soil type, not moisture content, so χ^2 results indicate that the soils are of the same type on different the slopes.

The last column in Table 2 contains results from combining measurements from pits up and downstream from a weir. The uncertainty did increase from those calculated at single pits in columns 3 and 4 of Table 1. This implies that we may want to use different parameters upstream than down. This is another reasonable conclusion because it indicates that the soil type changes along the stream. The hypotheses that the same parameter set can be used on North and South facing slopes, but should not be used at different depths nor at locations 15 *cm* downstream is logical, and we have quantified their validity with the χ^2 method.

Lastly, we point out the difference between the uncertainty found by the χ^2 method and the standard deviation of measurements across multiple pits. When measurements from the North and South slopes are combined the uncertainties are of the same magnitude. However, when measurements up and downstream from the weir are combined the uncertainty calculated by the χ^2 method is significantly higher than the standard deviation of the measurements. This implies that just using the standard deviation of measurements when combining pits may underestimate the true uncertainty, and not be a good indication of when to use different parameter sets.

4.2.2. Uncertainty across pressure head measurements

We also show results from the χ^2 method with diagonal, rather than scalar weighting. This illustrates how *a priori* knowledge can be used to determine where to break up the χ^2 tests in (13)-(16). Scalar weighting $\mathbf{C}_\epsilon = \sigma_\epsilon^2 \mathbf{I}$ in Section 4.2.1 implies that uncertainty in soil moisture measurements is constant across all pressure head measurements. This is a poor physical assumption since the pressure head measurements are made independent of soil moisture measurements. There are thousands of measurements at these pits, thus the number of χ^2 tests will be significantly less than the number of data, and $\mathbf{C}_\epsilon$ will be block diagonal. The dimension of each block represents the data grouping, and we determine the grouping by plotting data from the single pit SU10_40 on a regular rather than log scale in Figure 8. Figure 8(a) shows the grouping with two χ^2 tests, i.e. $N = 2$ in (13)-(16) and roughly breaks up the data evenly at 250 *cm*. In Figure 8(b) the data were split evenly in three at 80cm and 275cm. This process was continued in Figures 8(c)-(e) to evenly split the data in 4, 5, and 10 groups. In each of the plots we give a representative σ_{ϵ_i} to illustrate the region where σ_{ϵ_i} is assumed constant.

The σ_{ϵ_i}, $i = 1, \ldots, N$, were found with the χ^2 method and the corresponding curves with $N = 2, 3, 4, 5$, and 10 are plotted in Figure 9. This is for the single pit SU10_40 with the data groupings defined in Figures 8(a)-(e). There we see that the estimates do not vary that greatly as we change the number of χ^2 test. This indicates that the scalar χ^2 method is sufficient for producing parameter estimates. However, we continue to investigate the effectiveness of diagonal weighting in Table 3.

	One χ^2 test	Two χ^2 tests	Three χ^2 tests	Four χ^2 tests	Five χ^2 tests	Ten χ^2 tests
σ_{ϵ_1}	0.0440 (23.1%)	0.0535 (28.1%)	0.0523 (27.4%)	0.0522 (27.4%)	0.0521(27.3%)	0.0542 (28.5%)
σ_{ϵ_2}		0.0287 (15.1%)	0.0382 (20.0%)	0.0500 (26.2%)	0.0506 (26.6%)	0.0501 (26.3%)
σ_{ϵ_3}			0.0212 (11.1%)	0.0261 (13.7%)	0.0462 (24.23%)	0.0521 (27.3%)
σ_{ϵ_4}				0.0248 (13.0%)	0.0125 (6.5%)	0.0482 (25.3%)
σ_{ϵ_5}					0.0290 (15.2%)	0.0531 (27.9%)
σ_{ϵ_6}						0.0400 (21.0%)
σ_{ϵ_7}						0.0162 (8.5%)
σ_{ϵ_8}						0.0067 (3.5%)
σ_{ϵ_9}						0.0244 (12.8%)
$\sigma_{\epsilon_{10}}$						0.0355 (18.6%)

Table 3: Uncertainty estimates for SU10_40 found by χ^2 method with diagonal weighting matrices.

In Table 3 the uncertainty estimates σ_{ϵ_i}, with $\mathbf{C}_\epsilon = \text{diag}(\sigma_{\epsilon_1}, \ldots, \sigma_{\epsilon_1}, \sigma_{\epsilon_2}, \ldots, \sigma_{\epsilon_2}, \sigma_{\epsilon_N}, \ldots, \sigma_{\epsilon_N})$, $N = 2, 3, 4, 5, 10$ are given. With two χ^2 tests we see that the largest estimates of uncertainty are in measurements at pressure heads with large magnitude, i.e above 250 *cm*. This is due to the fact that the same soil moisture content was observed at near similar pressure head measurements. With three χ^2 tests the largest uncertainty is at pressure head measurements above 80 cm, and with four χ^2 tests there is significant uncertainty between 60 *cm* and 250 *cm* but the majority is above 250 *cm*. As the number of χ^2 tests increases to 10 we see that the calculated uncertainty is largest above 250 *cm* and below 50 *cm*. Large uncertainty at low pressure head measurements is due to the fact that in this semi-arid climate, soils do not get fully saturated and there are not many soil moisture measurements at this level. These conclusions can be made by observing the data set, they are validated using χ^2 tests, and quantified with the χ^2 method. We conclude that the single uncertainty estimate across all pressure heads, 0.0440, is a reasonable estimate from which to infer parameters at soil pit SU10_40.

5. Summary and Conclusions

We have developed the χ^2 method to calculate diagonal weighting matrices for weighted least squares estimation. This is an extension of the scalar χ^2 method [21, 22, 23, 31] which can be viewed as a regularization method. The new method amounts to solving multiple χ^2 tests to give an equal number of equations as the number of unknowns in the diagonal weighting matrix for data or parameter misfits. The unknowns in the diagonal weighting matrcies are solved for using *fsolve* and *sqrtm* in *Matlab*. Future work involves efficient and reliable implementation of the solution of this nonlinear system of equations (13)-(16), and investigation of the appropriate number of χ^2 tests. We have shown here that these diagonal weighting matrices can give discontinuous parameter estimates with a least squares estimator on a benchmark inverse problem.

We applied both scalar and diagonal weighting matrices found by the χ^2 method to estimate the water content in soils in a watershed near Boise, Idaho. The χ^2 method was used to quantify uncertainty in point scale field measurements so that the resulting parameter estimates can be used on the watershed scale. The measurement uncertainties calculated by the χ^2 method incorporate spatial variability, define scales over which one parameter set can be used, and indicate areas where measurement uncertainty needs to be reduced.

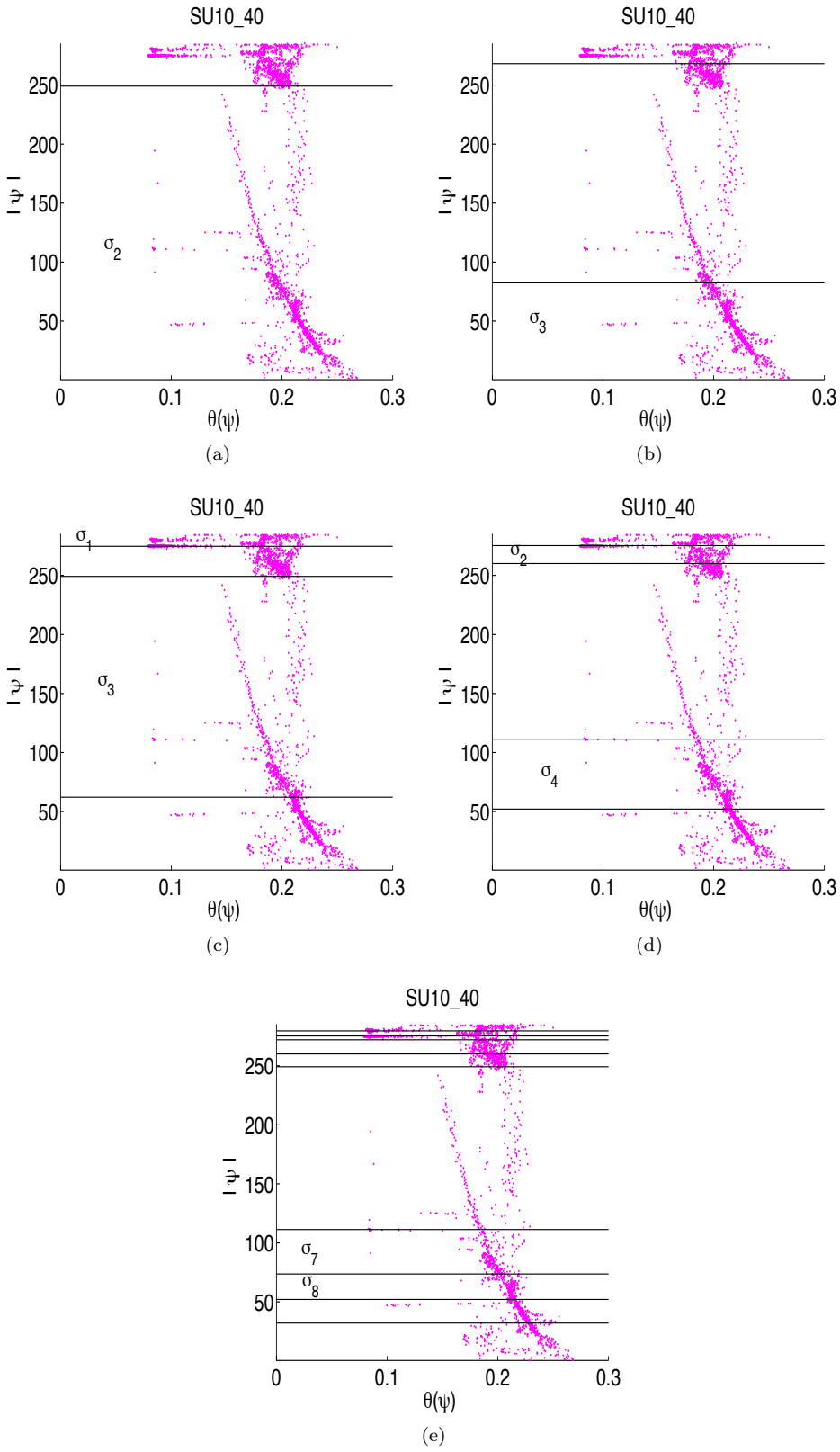


Figure 8: Intervals for multiple χ^2 tests (13)-(16) (a) $N = 2$ (b) $N = 3$ (c) $N = 4$ (d) $N = 5$ (e) $N = 10$.

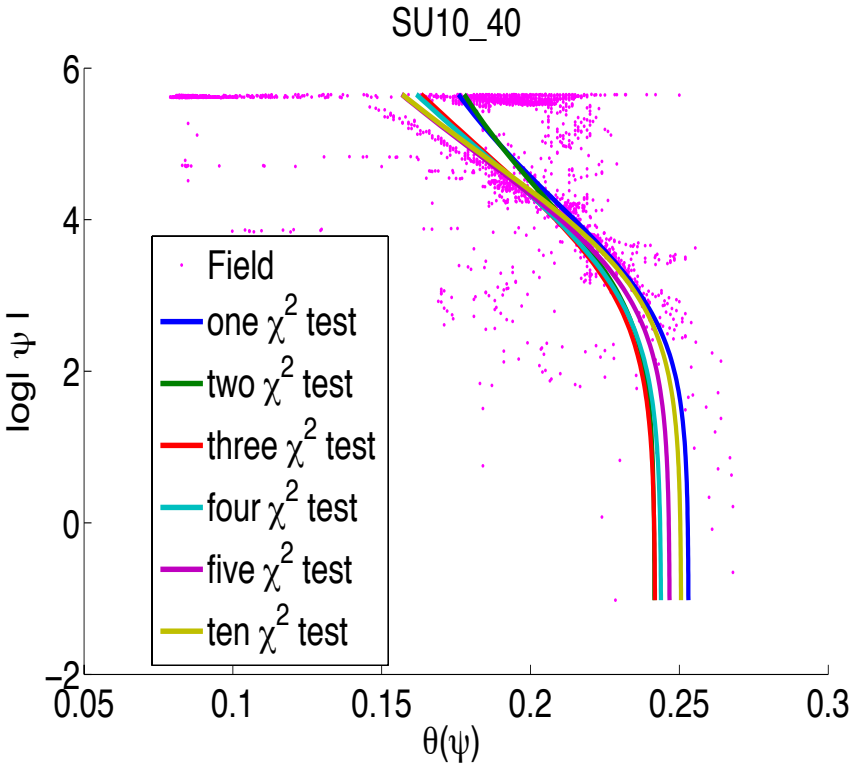


Figure 9: Soil water retention curve for SU10_40 found with multiple χ^2 tests.

6. Acknowledgements

I am grateful to Rosemary Renaut for helpful discussions and support with the χ^2 method, and to Jim McNamara and Molly Gribb for supplying the data.

References

[1] Aster, R.C., Borchers, B. and Thurber, C., 2005, *Parameter Estimation and Inverse Problems*, Academic Press, p 301.

[2] Barrodale, I. and Roberts, F. D. K. 1973, *An improved algorithm for discrete L_1 linear approximation*, SIAM Journal on Numerical Analysis, **10**, 839848.

[3] Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Marzouk, Y., Tenorio,L., van Bloemen Waanders, B., Willcox, K. (editors), 2010, *Large-Scale Inverse Problems and Quantification of Uncertainty*, Wiley, pp. 372.

[4] Box, G.E.P., and G.C. Tiao, 1973, "Bayesian Inference in Statistical Analyses," Addison-Wesley-Longman, Reading, Mass..

[5] Casella G and Berger R, 2001 *Statistical Inference* (California: Duxbury) 688p.

[6] Burger, M. , Resmerita, E. and He, L., 2007, *Error estimation for Bregman iterations and inverse scale space methods in image restoration*, Computing **81(2-3)**, 109-135.

[7] Davies, R,B., 1980, *Algorithm AS155: The distribution of a linear combination of χ^2 random variables*, Applied Statistics **29**, 323333.

[8] Coleman, T.F. and Li, Y., 1994, *On the Convergence of Reflective Newton Methods for Large-Scale Nonlinear Minimization Subject to Bounds*, Mathematical Programming, Vol. 67, Number 2, 189-224.

[9] DeGroot, M..H., 1970, *Optimal Statistical Decisions*, New York: McGraw-Hill.

[10] N.R. Draper and H. Smith "Appied Regresion Analysis", *Wiley*, New York, 3rd edition, 1998

[11] Evans, S.N. and Stark, P.B., 2002, *Inverse Problems as Statistics*, Inverse Problems, **18**, R55-R97.

[12] Gribb, M.M , Forkutsaa, I. , . Hansena, A, Chandler, D.G and McNamara, J.P., 2009, *The Effect of Various Soil Hydraulic Property Estimates on Soil Moisture Simulations*, Vadose Zone Journal Vol. 8 No. 2, p. 321-331.

[13] Hansen, P.C. "Analysis of discrete ill-posed problems by means of the L-curve SIAM Rev. 34 56180, 1992.

[14] Hansen, P. C., Nagy, J. G., and O'Leary, D. P., 2006, *Deblurring Images, Matrices, Spectra and Filtering*, (SIAM Fundamentals of Algorithms).

[15] Hansen, P. C., 2007,*Regularization Tools Version 4.0 for Matlab 7.3*, Numerical Algorithms, **46**, 189-194.

[16] He, X, and Ren, L, A , *Multiscale finite element linearization scheme for the unsaturated flow problems in heterogeneous porous media*, Water. Resour. Res., Vol. 32, W08417, 2006.

[17] Higham, N.J., 2008, *Functions of Matrices Theory and Computation*, SIAM, p .425.

[18] HYDRUS, http://www.pc-progress.com/en/Default.aspx?hydrus-3d.

[19] Marquardt, D. W., 1970, *Generalized inverses, ridge regression, biased linear estimation, and nonlinear estimation*, Technometrics, **12** (3), 591-612.

[20] McNamara, J. P., Chandler, D. G., Seyfried, M., and Achet, S., 2005, *Soil moisture states, lateral flow, and streamflow generation in a semi-arid, snowmelt-driven catchment*, Hydrological Processes, **19**, 4023-4038.

[21] Mead, J., 2008, *Parameter estimation: A new approach to weighting a priori information*, J. Inv. Ill-posed Problems, **16**, 2, 175-194.

[22] Mead, J., and Renaut, R. A., *A Newton root-finding algorithm for estimating the regularization parameter for solving ill-conditioned least squares problems*, Inverse Problems, Vol. 25, No. 2, 025002.

[23] Mead, J., and Renaut, R. A., 2010 *Least squares problems with inequality constraints as quadratic constraints*, Linear Algebra and its Applications, **432**, 1936-1949.

[24] Menke W, 1989 *Geophysical Data Analysis: Discrete Inverse Theory* (San Diego: Academic Press) p 289.

[25] Mertens J., Madsen H., Kristensen M., Jacques D., Feyen J., 2004, *Sensitivity of soil parameters in unsaturated zone modelling, and the relation between effective, laboratory and in-situ measurements*, Hydrological Processes Volume 19 Issue 8, Pages 1611 - 1633.

[26] Morozov, V. A., "On the solution of functional equations by the method of regularization", *Sov. Math.Dokl.* 7 4147, 1966.

[27] Mualem, Y., 1976, *A new model for predicting the hydraulic conductivity of unsaturated porous media*, Water Resour. Res., **12**, 513-522.

[28] O'Leary, D.P. and Rust, B. W. , 1986, *Confidence intervals for inequality-constrained least squares problems, with applications to ill-posed problems*, SIAM J. on Sci. Comput. **7** 473-489.

[29] Ramm, A.G., 2005, *Inverse Problems*, Springer Science, New York.

[30] Richards, L. A., 1931, *Capillary conduction of liquids in porous media*, Physics, **1**, 318-333.

[31] Renaut, R.A,, Hnetynkova, I. and Mead, J. L., 2010, *Regularization Parameter Estimation for Large Scale Tikhonov Regularization Using A Priori Information,* Computational Statistics and Data Analysis, **54**, 3430-3445.

[32] Rudin L.I., Osher, S. and Fatemi, E., 1992, '*Nonlinear total variation based noise removal algorithms*, Physica D, **60**, 259-268.

[33] Tarantola, A., 2005, *Inverse Problem Theory and Methods for Model Parameter Estimation*, (SIAM).

[34] Scales J and Tenorio L, "Prior Information and Uncertainty in Inverse Problems" *Geophysics* 66 389-97, 2001.

[35] Schaap, M.G. and W. Bouten. 1996, *Modeling water retention curves of sandy soils using neural networks*, Water Resour. Res. 32:3033-3040.

[36] Schaap, M.G., F.J. Leij and M.Th. van Genuchten. 2001, *Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions*, J. Hydrol. 251:163176.

[37] Smith, R.E. and Diekkrüger, "Effective soil water characteristics and ensemble soil water profiles in heterogenous soils", *J. Geophy. Res.* vol. 32, No. 7, p 1993-2002, 1996.

[38] Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian, "Bayesian recursive parameter estimation for hydrological models," *Water Resour. Res* 37, 2521-2535, 2001.

[39] Tikhonov A, and Arsenin V, 1977 *Solutions of Ill-Posed Problems* (New York: Wiley) p 272.

[40] van Genuchten, M. Th., 1980, *A closed-form equation for predicting the hydraulic conductivity of unsaturated soils.*, Soil Sci. Soc. Am. J., **44**, 892-898.

[41] Vogel, C. R., 2002, *Computational Methods for Inverse Problems*, (SIAM Frontiers in Applied Mathematics).

[42] Vrugt, J.A. and W. Bouten, "Toward improved identificability of hydrologic model parameters: The information content of experimental data", *Water. Resour. Res.*, 38(12), 1312-1225, 2002.

[43] Wahba G "Practical approximate solutions to linear operator equations when the data are noisy", *SIAM J. Numer. Anal.* 14 65167,1977.

[44] Warrick, A.W. "Soil Water Dynamics", Oxford University Press, p 391, 2003.

[45] Wing, G.M. and Zahrt J.D., 1991, '*A primer on Integral Equations of the First Kind*, SIAM, Philadelphia, p. 109.

[46] Zhu, J. M.H., Young, and M.T. van Genuchten, "Upscaling Schemes and Relationships for the Gardner and van Genuchten Hydraulic Functions for Heterogeneous Soils", *Vadose Zone Journal*, 6:186-195, 2007.