Fundamentals of inverse problems

Odd Kolbjørnsen

Department of Mathematical Sciences
Norwegian University of Science and Technology
Norway
2002

1 Introduction

Inverse problems can be defined as problems that consist in finding the cause of an observed effect. An inverse problem is always paired with a direct problem that provide the effect of a given cause. This definition requires the formulation of any specific problem to be based on physical laws and that physics must specify what is a cause and what is an effect as well as provide the equations relating the effects to the causes (Bertero 1989). Inverse problems arise naturally if one is interested in determining the internal structure of a system based on the system's observed behavior or in determining the unknown input that give rise to an observed output (Hansen 1998).

In a mathematical language an inverse problem relate to an operator equation,

$$y = K(z)\,, \tag{1}$$

with $K : \mathcal{Z} \to \mathcal{Y}$ being a possibly nonlinear operator. The direct problem is to determine the effect, y, of a given cause, z, whereas the inverse problem is to determine the cause, z, of an observed effect y. The function space $\mathcal{Z}$ is commonly denoted the model space or parameter space, while $\mathcal{Y}$ is denoted the data space.

Expression (1) is unlikely to hold when y is a measured quantity, since measurements have finite precision. In addition Expression (1) may be inaccurate in the sense that the operator K does not model all aspects of the physical processes that produce the observations. The problem is hence more realistically stated as,

$$y = K(z) + \varepsilon\,, \tag{2}$$

with ε being an error term.

To give a comprehensive account for all aspects of inverse problems, is impossible in a short introduction since the field have so many branches spanning physical, mathematical, computational and statistical aspects. The current presentation include basic mathematical and statistical definitions that are relevant

for inverse problems, and discuss some of the philosophies that underlies the different solution methods. The presentation will concentrate on the case where $K : \mathcal{Z} \to \mathcal{Y}$ is a compact linear operator since this theory is by far the best developed. The inversion methods and the underlying philosophies are frequently generalized to solve nonlinear inverse problems. This is briefly discussed.

The presentation is organized as follows. Mathematical aspects of inverse problems are presented in in Section 2. Inversion by regularization is presented in Section 3. The statistical theory of point estimation is presented in 4. Statistical minimax inversion and Bayesian inversion are presented in Section 5 and 6 respectively. In Section 7 the three inversion methodologies are compared with respect to similarities and differences. Section 8 contains some concluding remarks and the authors personal preferences. In Section 9 the content of the thesis is discussed in light of the current introduction.

2 Mathematics of inverse problems

The presentation in this section is based on Engl, Hanke and Neubauer (1996), Kirsch (1996) and Hansen (1998). It contain basic mathematical definitions and discusses approximate solutions to inverse problems.

2.1 Problem classification

According to the informal definition above a problem is classified as direct or inverse by the physics defining the problem. From a mathematical point of view problems are more naturally labeled as being well-posed or ill-posed. A problem is well-posed if there exists a unique, stable solution. The notion of a well-posed problem is attributed to Hadamard (1902, 1923). Although there is no formal connection between the two sets of labels, it is however true, with few exceptions, that direct problems are well-posed while the corresponding inverse problems are ill-posed. A definition of a well-posed inverse problem reads,

Definition 1 (Well-posed) *Let $\mathcal{Z}$ and $\mathcal{Y}$ be normed spaces and let $K : \mathcal{Z} \to \mathcal{Y}$ be a continuous operator from $\mathcal{Z}$ into $\mathcal{Y}$. The problem $y = K(z)$ is well-posed in the sense of Hadamard if the following three conditions are satisfied:*

1. *Existence: There exist a solution $z \in \mathcal{Z}$ for any $y \in \mathcal{Y}$ with $K(z) = y$*
2. *Uniqueness: There exist at most one solution $z \in \mathcal{Z}$ for any $y \in \mathcal{Y}$ with $K(z) = y$*
3. *Stability: For every positive number ϵ, there exist a positive number $\delta(\epsilon)$ such that any pair $z_1, z_2 \in \mathcal{Z}$ for which $\|K(z_1) - K(z_2)\| < \delta(\epsilon)$, $\|z_1 - z_2\| < \epsilon$*

Problems for which at least one of the three conditions above fails to hold are termed ill-posed.

Whether a problem is well-posed or not, depend both on the operator K and the function spaces $\mathcal{Z}$ and $\mathcal{Y}$.

The simplest case of an operator equation is, a matrix equation, $y = K\,z$, for which $\mathcal{Z} = R^n$, $\mathcal{Y} = R^m$ and K is a $m \times n$ matrix. The existence criterion then imply that the rank of K is equal to m, the uniqueness criterion imply that the rank of K is equal to n. Hence to assure both existence and uniqueness the matrix must be square and have full rank. These are also sufficient conditions for a matrix equation to be well posed. Any inverse problem formulated as a square matrix equation of full rank is hence stable in a strict mathematical sense. For matrix equations the criterion of stability relates to computational aspects of the inverse, K^{-1}. If a small change in y produce a large change in $z = K^{-1}y$, the system is said to be unstable. The standard example of such a matrix is

$$K = \begin{bmatrix} 1 & 1+\epsilon \\ 1 & 1 \end{bmatrix}$$

with ϵ being a small number. Let the supscript T denote matrix transpose. The solution for $y^T = [\,2\,,\,2\,]$ is $z^T = [\,2\,,\,0\,]$, while the solution for $y^T = [\,2+\epsilon\,,\,2\,]$ is $z^T = [\,1\,,\,1\,]$, hence a change in the input of order ϵ result in a change in the answer of order one. In unstable systems, some of the equations are almost linearly dependent. These systems are therefore hard to solve numerically, see Hansen (1998) for an extensive discussion.

2.2 Singular value expansion

Consider an operator equation

$$y = K\,z\,, \tag{3}$$

with $K : \mathcal{Z} \to \mathcal{Y}$ being a compact linear operator between two Hilbert spaces. In common notation $K^* : \mathcal{Y} \to \mathcal{Z}$ denotes the adjoint of K, and is defined by the requirement that for all $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$, $(K\,z, y) = (z, K^*y)$, with $(\cdot,\cdot)$ denoting inner products in $\mathcal{Y}$ and $\mathcal{Z}$ at the left and the right side of the equality respectively. For any compact linear operator $K : \mathcal{Z} \to \mathcal{Y}$, there exist a singular system $\{\sigma_i, v_i, u_i\}_{i=1}^{\infty}$, with σ_i being nonnegative numbers, $\{v_i\}_{i=1}^{\infty}$ and $\{u_i\}_{i=1}^{\infty}$ being complete orthonormal systems of basis elements for $\mathcal{Z}$, and $\mathcal{Y}$ respectively. That is, $z \in \mathcal{Z}$ and $y \in \mathcal{Y}$ can be represented by the generalized Fourier series, $z = \sum z_i\,v_i$ and $y = \sum y_i\,u_i$, with $z_i = (v_i, z)$ and $y_i = (u_i, y)$. The numbers σ_i are the singular values of K, these are usually ordered in a non increasing order, $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$. Singular systems resembles the eigensystem of compact self adjoint operators, indeed $\{\sigma_i^2, v_i\}_{i=1}^{\infty}$ and $\{\sigma_i^2, u_i\}_{i=1}^{\infty}$ are the eigensystems of the self adjoint operators K^*K and KK^* respectively.

The singular system defines the singular value expansion of K,

$$K\,z = \sum_{i=1}^{\infty} \sigma_i \,(v_i, z)\, u_i \;, \tag{4}$$

The singular value expansion diagonalize the problem such that the generalized Fourier coefficients of z can be solved independently, i.e.

$$Kz = y \;\Leftrightarrow\; \sigma_i\, z_i = y_i \;,\; i = 1, 2, \ldots \;. \tag{5}$$

The ill-posedness of a linear inverse problem is frequently related to the decay of the singular values. As $i \to \infty$, $\sigma_i \to 0$, hence the effect of z_i in Kz diminishes as $i \to \infty$. The rate of decay of the singular values can be used to classify linear ill-posed problems. A problem is termed mildly ill-posed if $\sigma_i \sim i^{-r}$ and $0 < r \leq 1$, moderately ill-posed if $\sigma_i \sim i^{-r}$ and $r > 1$ or severely ill-posed if $\sigma_i \sim \exp\{-ri\}$ or worse.

The singular value expansion is the infinite dimensional analog of the singular value decomposition of a matrix. In the case of K being a real $m \times n$ matrix, this decomposition reads,

$$K = U\Sigma V^T = \sum_{i=1}^{\min(n,m)} \sigma_i\, u_i\, v_i^T \;,$$

with $U = [u_1\, u_2 \cdots u_m] \in R^{m \times m}$ and $V = [v_1\, v_2 \cdots v_n] \in R^{n \times n}$ being matrices with orthonormal columns, and Σ being a $m \times n$ diagonal matrix with the singular values, $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_{min(n,m)}$, on the diagonal. From the equations $K^T K = V\Sigma^T \Sigma V^T$ and $KK^T = U\Sigma\,\Sigma^T U^T$ it is seen that the singular value decomposition of K is closely linked to the eigenvalue decomposition of $K^T K$ and KK^T.

2.3 Fundamental subspaces and the generalized inverse

The range of K, $\mathcal{R}(K)$, are those $y \in \mathcal{Y}$ that can be reached from a $z \in \mathcal{Z}$. This is in general not a proper subspace of $\mathcal{Y}$, but the closure of this set, $\overline{\mathcal{R}(K)}$, is so. $\overline{\mathcal{R}(K)}$ is spanned by the basis elements of $\mathcal{Y}$ that correspond to strictly positive singular values, i.e. $\{u_i\}_{\{i:\,\sigma_i > 0\}}$. The orthogonal complement of $\overline{\mathcal{R}(K)}$ in $\mathcal{Y}$ is termed the null space of K^*, denoted $\mathcal{N}(K^*)$, and it is spanned by the basis elements for which the corresponding singular values are zero, i.e. $\{u_i\}_{\{i:\,\sigma_i = 0\}}$. Similarly $\mathcal{Z}$ can be divided into two subspaces corresponding to whether the elements influence the output of K or not. From Expression (4) it is easy to see that the subspace that influence the output of K is spanned by the basis elements of $\mathcal{Z}$ for which the corresponding singular values are strictly positive, i.e. $\{v_i\}_{\{i:\,\sigma_i > 0\}}$. This space is the closure of the range of K^*, denoted $\overline{\mathcal{R}(K^*)}$. The orthogonal complement of $\overline{\mathcal{R}(K^*)}$ in $\mathcal{Z}$ is termed the null space

of K, denoted $\mathcal{N}(K)$, and is spanned by the basis elements of $\mathcal{Z}$ for which the corresponding singular values are zero, i.e. $\{v_i\}_{\{i:\sigma_i=0\}}$. It is more natural to relate to the operator K instead of the adjoint, hence it is common to define $\overline{\mathcal{R}(K^*)}$ as the orthogonal complement of $\mathcal{N}(K)$, i.e. $\overline{\mathcal{R}(K^*)} = \mathcal{N}(K)^\perp$.

The generalized inverse, $K^\dagger$, of a compact linear operator K can be defined using the singular system of K,

$$K^\dagger y = \sum_{\{i:\,\sigma_i>0\}} \frac{(u_i, y)}{\sigma_i} v_i \,, \tag{6}$$

when $y \in \mathcal{R}(K)$ this generalized Fourier series converge. The solution is easily found by solving the sequence problem in Expression (5). For a given $y \in \mathcal{Y}$ the convergence of the series in Expression (6), is equivalent with y satisfying the Picard criterion,

$$\sum_{\{i:\,\sigma_i>0\}} \frac{|(u_i, y)|^2}{\sigma_i^2} < \infty \,.$$

When the Picard criterion is fulfilled, the general solution to the inverse problem is characterized by the sum of one component from the null space of K and the generalized inverse of y, i.e. for any $z_0 \in \mathcal{N}(K)$,

$$z = z_0 + K^\dagger y \,. \tag{7}$$

This decomposition of the general solution as a sum of the homogeneous solution and a particular solution, is common in differential equations and matrix algebra.

2.4 Approximate solutions

In a real case the observations are prone to contain error hence the operator equation in Expression (3), should be replaced by

$$y = K\,z + \varepsilon \,, \tag{8}$$

with ε being an error term, see Expression (2). For most inverse problems the generalized inverse, $K^\dagger$, is unstable because $\sigma_i \to 0$ as $i \to \infty$. This imply that a small error ε will contribute significantly to the series in Expression (6) because the inner product (u_i, ε) is divided by σ_i. That is, the Picard criterion is usually not fulfilled for measured data. Since the true solution of Expression (8) cannot be obtained, an approximate solution is sought. An approximate solution is denoted $\hat{z}$. Some commonly used approximations are discussed below. All the approximate solutions are parameterized by a nonnegative number α that defines the degree of approximation. The parameter, α, is defined such that $\alpha = 0$ defines the exact solution. This parameter is briefly discussed below and more throughly in Section 3.

Filtering is a common way to obtain smoother solutions. In the context of inverse problems, filter factors may be introduced in the generalized Fourier series defining the generalized inverse, see Expression (6). The approximation may then be written as

$$\hat{z} = \sum_{\{i\,:\,\sigma_i>0\}} \phi_i(\alpha)\,\frac{(u_i, y)}{\sigma_i}\,v_i\ , \tag{9}$$

with $\phi_i(\alpha)$ being filter or shrinkage factors, and $\{\sigma_i, v_i, u_i\}_{i=1}^{\infty}$ being the singular system of K. The filter factors satisfy $0 \leq \phi_i(\alpha) \leq 1$ and $\phi_i(0) = 1$, and are defined such that the series in Expression (9) converge for $\alpha > 0$. Many different approximate solutions of Expression (8) have the form of Expression (9). In fact this expression is to general and a specific choice must be made for the filter factors $\phi_i(\alpha)$. An example is the truncated singular value expansion, which can be defined by $\phi_i(\alpha) = I\{\sigma_i > \alpha\}$, where $I\{\cdot\}$ is one if the event in the brackets is true, zero otherwise. In this case only the terms where the singular value exceed α are included. The singular value expansion of a problem is generally not known this complicates the approach in practical situations.

Tikhonov regularization exploits the fact that for any $y \in \mathcal{R}(K)$ the generalized inverse is the unique solution of the least squares problem

$$z = \arg\min_{z\,\in\,\mathcal{N}(K)^{\perp}} \|Kz - y\|^2. \tag{10}$$

When $y \notin \mathcal{R}(K)$, the solution $K^{\dagger}y$ is unbounded, i.e. $\|K^{\dagger}y\| = \infty$. Tikhonov regularization avoid this by adding a penalty term in the minimization to keep z bounded. The approximate solution, $\hat{z}$, is defined as the unique solution of

$$\hat{z} = \arg\min_{z\in\mathcal{Z}} \|Kz - y\|^2 + \alpha J(z), \tag{11}$$

with $J(z)$ being a suitable penalizing functional; and α being a positive number determining the trade off between the mismatch to the data and the penalizing term. The most common choice of penalizer is the squared norm in $\mathcal{Z}$, i.e. $J(z) = \|z\|^2$. In this case the approximate solution have the form of Expression (9), with $\phi_i(\alpha) = \sigma_i^2/(\sigma_i^2 + \alpha)$. Other choices are Sobolev norms, L^1 norm and maximum entropy. The methodology can also be generalized by using other measures for deviations in the data. Tikhonov regularization may be formulated in different ways, such as minimizing the error in the data subject to an upper bound on the penalizing functional, or as minimizing the functional subject to an upper bound on the error. For any given data y there is an one to one connection between the different formulations, where the bounds can be computed in terms of α and y.

Landweber iteration is an algorithmically defined approximate solution. It can be regarded as steepest decent algorithm with a fixed step length, $\omega < 1/\sigma_1^2$. The approximate solution is defined iteratively by

$$z^n = z^{n-1} - \omega K^*(Kz^{n-1} - y)\ , \tag{12}$$

with starting point $z^0 = 0$. After $m = 1/\alpha$ iterations, the solution obtained have the form in Expression (9) with $\phi_i(\alpha) = 1 - (1 - \omega\sigma_i^2)^m$. Note that when $y \notin \mathcal{R}(K)$ the true solution is unbounded, since Expression (12) converge to the true solution it is not beneficial to iterate the expression too many times. The amount of approximation hence lies in the number of iterations.

Conjugate gradient is another iterative technique for approximating the solution in Expression (10). The conjugate gradient identify the best solution in the Krylov subspace space of order m after m iterations. The Krylov subspace of order m is defined by

$$\mathcal{K}_m(K^*K, K^*y) = \text{span}\{K^*y, (K^*K)K^*y, \ldots, (K^*K)^{m-1}K^*y\}.$$

In certain applications the Krylov subspace of order m is an approximation to the subspace spanned by the first m singular vectors, i.e. $\text{span}\{v_i\}_{i=1}^m$. In these cases the conjugate gradient method stopped after $m = 1/\alpha$ iterations, can be seen as an approximation to the truncated singular value expansion with $1/\alpha$ terms. Note again that the amount of approximation is determined by the number of iterations.

Landweber and conjugate gradient iterations as defined above are used as means to define approximate solutions, for this purpose a finite number of iterations is required. The iterations can also be used as numerical schemes to solve well posed problems such as Expression (11). For such cases the the number of iterations is a purely numerical question. Practical implementation of the methods above require discretization. Different discretization schemes can also be used to define approximate solutions of the continuous problem. Most of the methods above can be described as filtering of the singular system. The singular value expansion is however not directly accessible for a given problem. In a given situation the Landweber iteration is hence much easier to implement than the general filtering scheme.

The approximate solutions above are not fully specified but depend on the parameter α that determines the tradeoff between data adaption and the boundedness of the approximate solution, $\hat{z}$. Many techniques are developed for choosing the parameter α. The L-curve is a helpful tool in understanding the impact of a particular choice, and can be used in various ways to pick a particular value of α. Cross validation and generalized cross validation (Wahba 1990) are also used for this purpose. The parameter choice is formalized in the regularization theory to be discussed next.

3 Inversion by regularization

The basic idea of regularization theory, is that the approximate solution should be stable with respect to small deviations in the observations. The problem can be seen as a game between a scientist and a malicious opponent. For given

bounds on the parameter $z \in C$ and the error $\|\varepsilon\| < \delta$, the scientist can choose the approximate solution, $\hat{z}$, depending only on the data y. The subset $C \subset \mathcal{Z}$ should at least exclude the components of z that does not influence the data, i.e. $\mathcal{N}(K)$. The opponent can chose the parameter, z, within the restriction C. The pay off in the game is the maximum deviation between approximate solution and the parameter for errors within the error bound

$$\sup_{\|y-K\,z\|<\delta} \|\hat{z} - z\|^2$$

This measure of deviation can be interpreted as allowing the opponent to chose the error in addition to the parameter. The ultimate goal for the scientist is hence to find an approximate solution that minimize the worst case deviation.

In order to reduce the complexity of the problem an indexed family of continuous operators $R_\alpha : \mathcal{Y} \to \mathcal{Z}$ is considered. A family of operators is denoted a regularization strategy if $R_\alpha K z \to K^\dagger K z$ for all $z \in \mathcal{Z}$ when $\alpha \to 0$. That is $R_\alpha K$ converge pointwise to the projection operator onto $\mathcal{N}(K)^\perp$. All of the approximate solutions listed in Section 2.4 are valid regularization strategies. A regularization method consist of a regularization strategy, R_α, and a rule for selecting the index α. A selection rule that only depend on the error bound, δ, is denoted an apriori selection rule, if the index also depend on the data, y, it is termed an aposteriori selection rule. A pair (R_α, α) is called an admissible or convergent regularization method if

$$\sup_{\|y-K\,z\|<\delta} \alpha(\delta, y) \to 0 \quad \text{as } \delta \to 0\,,$$

and

$$\sup_{\|y-K\,z\|<\delta} \|R_{\alpha(\delta,y)} y - K^\dagger K z\| \to 0 \quad \text{as } \delta \to 0\,.$$

That is, as the error in the data tends to zero, so should the amount of regularization and the error of the approximation. By choosing a regularization strategy, the original problem has been reduced to a one dimensional problem, selecting a value of α for a given value of δ and y.

The most widespread procedure for selection of α is the discrepancy principle of Morozov, which defines the value of the regularization parameter α to be the one that yields $\|KR_\alpha y - y\| = \delta$, with δ being the maximum bound on the error.

Regularization methods are evaluated by the rate of convergence as the error in the observation approach zero. In this respect the goal is to obtain an uniform convergence rate in $\mathcal{Z}$, this is however impossible for most inverse problems. For this reason attention is drawn to subsets of $\mathcal{Z}$ of the form

$$\mathcal{Z}_{B,\rho} = \{\, z = Bw,\ \|w\| < \rho \,\}, \tag{13}$$

with $B : \mathcal{W} \to \mathcal{Z}$ being a bounded linear operator; and ρ being a finite number. This can be interpreted as an abstract smoothness constraint on z. Typical

theorems for inversion by regularization consist of two results, the first result state the optimal convergence rate in $\mathcal{Z}_{B,\rho}$, the second result prove that one particular regularization method have the optimal rate of convergence in $\mathcal{Z}_{B,\rho}$. A typical theorem is hence that the approximate solution found by Tikhonov regularization using a quadratic regularizer and a selection rule given by the discrepancy principle, is obtained by an admissible regularization method and obtain the optimal convergence rate in $\mathcal{Z}_{K^*,\rho}$.

Inversion by regularization is a general approach and can be used to solve nonlinear inverse problems. There are however few general results for this type of problems. Most convergence theorems are of local type, or assume that the global solution of a minimization problem can be found. Further, the uniqueness problem is not as easily decomposed as for linear systems, see Expression (7). To avoid problems in this respect a new origin, z_0, that represent the best prior guess is selected. When several solutions can be chosen, the one closest to z_0 is preferred. Tikhonov regularization, with penalizer $J(z) = \|z - z_0\|^2$ is widely used in nonlinear inverse problems. Since there are few rigorous results in this area there is no general optimality results for the solution obtained in most practical cases.

4 Statistical aspects of inverse problems

In the current presentation, inverse problems will be discussed in the context of point estimation (Lehmann and Casella 1998). There are many other statistical aspects of inverse problems than those discussed here. Most important are statistical methods for estimating the regularization parameter α, for a regularization strategy R_α, without having a prior bound on the error, see O'Sullivan (1986), Wahba (1990) and Hansen (1998) for a discussion of some of these methods, see also Stark (2000) for an insightful discussion of inverse problems as statistics.

From a statistical point of view an inverse problem, as phrased in Expression (2) and (8), is no different from any other estimation problem. A parameter z in a parameter space $\mathcal{Z}$ is to be estimated based on observations, y, in the data space $\mathcal{Y}$. The statistical link between the parameter, z, and the observations, y, is described by the likelihood, $p(y|z)$. Here, and in what follows, $p(\cdot)$ is being used as a generic probability distribution. A parameter z, or a feature of z, is said to be unidentifiable if it does not influence the likelihood, otherwise it is identifiable. An estimator, $\hat{z}$, for z is a measurable function of the data, $\hat{z}(y)$, or in general an operator, $\hat{z} : \mathcal{Y} \to \mathcal{Z}$. To evaluate an estimator a loss function, $L(z, \hat{z})$, is defined. A common choice, that will be used in what follows, is the squared L^2 norm, i.e. $L(z, \hat{z}) = \|z - \hat{z}\|^2$. The statistical philosophy is that if the experiment conducted to give the observations y is repeated, a new sample y' from $p(y|z)$ is obtained. Hence the error, ε, in Expression (2) and (8) is given a random variable interpretation. The objective is now to identify the estimator

that minimizes the expected loss when observations are sampled according to the likelihood. The expected loss of an estimator, $\hat{z}(y)$, is denoted the risk, $r_{\hat{z}}(z)$, and is defined pointwise in $\mathcal{Z}$ by,

$$r_{\hat{z}}(z) = \int_{\mathcal{Y}} \|z - \hat{z}(y)\|^2 \, dp(y|z) = \mathrm{E}_{Y|z} \left\{ \|z - \hat{z}(Y)\|^2 \right\}.$$

An estimator is said to be admissible if no other estimator can improve the risk uniformly in $\mathcal{Z}$. The risk is defined pointwise in $\mathcal{Z}$, but the estimator must be chosen without knowledge of z, hence in some way or other the estimator must take the risk for all $z \in \mathcal{Z}$ into account. In the minimax risk approach, the maximum risk over $\mathcal{Z}$ is used to compare estimators. An estimator is optimal if it has the least maximum risk in comparison to any other estimator. This philosophy is used for inverse problems in Section 5 below. In the average risk approach a measure is defined on $\mathcal{Z}$ and the the optimal estimator is defined as the one that minimizes the average risk according to this measure. The average risk approach is the fundament of Bayesian statistics which is further developed in Section 6 below.

The principles of estimation are the same for inverse problems as for any other statistical problem. Most inverse problems do however have some characteristics that distinguish them from the classical statistical theory. In inverse problems the number of parameters will frequently be of the same order, most often larger, than the number of observations. This can be seen from the sequence problem in Expression (5). Inverse problems of this type have closer resemblance to problems where the number of parameters grow together with the number of observations, than to the classical large sample theory (Lehmann 1999). Stein (1956) showed that the celebrated maximum likelihood estimator is inadmissible in a sequence model when there is an equal number of observations and parameters, larger than two. It is hence not likely that the maximum likelihood methodology will succeed in solving inverse problems. Further, in inverse problems the parameter is observed through a transform, see Expression (2), and not directly as in the traditional statistical theory of function estimation.

5 Statistical minimax inversion

In the minimax approach the estimates are evaluated by the maximum risk in $\mathcal{Z}$. The problem can be seen as a game between the scientist and a malicious opponent. For given bounds on the parameter $z \in C$ and a specified likelihood model, $p(y|z)$, the scientist can choose the estimator, $\hat{z}$, depending only on the data y. The opponent can chose the parameter, z, within the restriction $C \subset \mathcal{Z}$. The pay off in the game is the risk for the opponents choice of parameter, that is the expected loss under the likelihood model,

$$r_{\hat{z}}(z) = \mathrm{E}_{Y|z} \left\{ \|z - \hat{z}(Y)\|^2 \right\}.$$

The subset C may be a smoothness constraint such as Expression (13). The ultimate goal for the scientist is hence to find an estimator that minimize the worst case expected loss.

Estimators in the minimax approach are frequently evaluated by the rate of convergence as the information content of the date increase, the zero noise limit is common. Typical theorems for statistical minimax inversion consist of two results. First the optimal rate of convergence in C is obtained next an optimal estimator is found. The case where the set C is of the quadratic type, see Expression (13) is treated in Johnstone and Silverman (1990,1991), in which a rate optimal estimator is defined. The estimator correspond to filtering of singular values, see Expression (9). The estimator truncate the singular value expansion and shrink the remaining coefficients.

In some cases a smoothness constraint on the parameters such as Expression (13) can be limiting. The resulting estimators are always linear or almost so. Resent developments in the field of computational harmonic analysis allow for using the notion of sparsity rather than smoothness. The resulting estimator being represented by the wavelet-vaguelette decomposition (Donoho 1995). The main idea is that wavelets give a sparse representation of functions. Since the functions sought have few large coefficients, the focus can be directed towards which coefficients that should be estimated, instead of trying to estimate all. The typical result for these type of estimators is that the minimax rate of convergence is obtained adaptively within a logarithmic term. The adaptivity is in contrast to the traditional approaches where the smoothness must be defined prior to the estimation. The estimators based on wavelet-vaguelette decomposition has been particular successful for mildly and moderately ill-posed inverse problems.

The concern in minimax estimation is to get the best possible estimator for z, not to assess the uncertainty. There are however statistical results that deal with the uncertainty of the estimates also in this case, Stark (2001) considers confidence intervals for linear estimators of linear functionals, and report the methods of strict bounds (Backus 1989) and minimax confidence intervals (Donoho 1994), in both cases under the assumption of K being a compact linear operator.

The minimax approach is a general principle for estimation, and would apply also to nonlinear inverse problems. It is however a complex machinery and to the knowledge of the author, which may be limited, there has been no extensive study of minimax estimation for general nonlinear inverse problems.

6 Bayesian inversion

In the Bayesian approach knowledge and uncertainty regarding the parameter, z, is summarized in probability distributions. The prior distribution, $p(z)$, represent the knowledge of z prior to observations. The average risk, commonly

denoted the Bayes risk, of an estimator, $\hat{z}(y)$, is the expected risk under the prior measure,

$$B_{\hat{z}}[p(z)] = \mathrm{E}_Z \left\{ r_{\hat{z}}(Z) \right\} = \mathrm{E}_Z \left\{ \mathrm{E}_{Y|Z} \left\{ \|Z - \hat{z}(Y)\|^2 \right\} \right\} . \tag{14}$$

The objective in Bayesian estimation is to find the estimator that minimizes the Bayes risk, $B_{\hat{z}}[p(z)]$, for a given prior $p(z)$. When the Bayes risk is finite, the order of integration in Expression (14) can be interchanged. The Bayes estimator, $\hat{z}_B : \mathcal{Y} \to \mathcal{Z}$, is then formally defined by

$$\hat{z}_B = \arg\min_{\hat{z}} \mathrm{E}_Y \left\{ \mathrm{E}_{Z|Y} \left\{ \|Z - \hat{z}(Y)\|^2 \right\} \right\}$$

The problem can be solved for each y separately by minimizing

$$\hat{z}_B(y) = \arg\min_{\hat{z}} \mathrm{E}_{Z|y} \left\{ \|Z - \hat{z}(y)\|^2 \right\} \tag{15}$$

The major advantage of this expression is that the estimator only need to be found for the observation, y, actually obtained. The unique minimizer of Expression (15) is known to be the posterior expectation, that is

$$\hat{z}_B(y) = \mathrm{E}_{Z|y}\{Z\} . \tag{16}$$

This is the classical Bayes estimator. The averaging measure in Expression (15) and (16) is denoted the posterior distribution and can formally be written as

$$p(z|y) = \frac{p(y|z)p(z)}{p(y)}. \tag{17}$$

For the Bayesian analyst the posterior distribution is the answer to the inverse problem, since this contains his updated knowledge regarding the parameter. The knowledge can be used to produce the best estimate of a parameter according to a general loss function and to assess uncertainty regarding the parameter.

Expression (16) and (17) look quite convenient, but computation of these quantities can be difficult. In order to evaluate expectations under the posterior distribution in the general case various types of Monte Carlo integration can be used. The most common approach is Markov chain based techniques like Metropolis-Hastings (Robert and Casella 1999). One important special case is however analytically tractable and will be describe in grater detail below. In the special case the observations are related to a compact linear operator with additive error, see Expression (8), and the parameter, z, and the error, ε, are modeled as Gaussian random functions.

A Gaussian random function, Z, in a separable Hilbert space can be represented by the Karhunen-Loève expansion, see Yaglom (1987),

$$Z = \sum_{i=1}^{\infty} Z_i \, v_i \ ,$$

with $\{Z_i\}_{i=1}^{\infty}$ being independent Gaussian random variables with mean μ_i and variance γ_i^2; and $\{v_i\}_{i=1}^{\infty}$ being the corresponding basis elements of unit length. The pairs $\{\gamma_i^2, v_i\}_{i=1}^{\infty}$ is the eigensystem of the covariance operator of Z. This is the infinite dimensional equivalent of the eigenvalues and eigenvectors of the covariance matrix. For simplicity let $\{Z_i\}_{i=1}^{\infty}$ be centered, i.e. $\mu_i = 0$, $\forall\, i$. The observations are $y = Kz + \varepsilon$, see Expression (8), with $K : \mathcal{Z} \to \mathcal{Y}$ being a compact operator; and ε being an error term, modeled as a Gaussian random function. Assume further that ε have the Karhunen-Loève expansion

$$\varepsilon = \sum_{i=1}^{\infty} \varepsilon_i u_i$$

with $\{\varepsilon_i\}_{i=1}^{\infty}$ being centered independent Gaussian random variables with variance λ_i^2, and for presentational simplicity that K have the singular system $\{\sigma_i^2, v_i, u_i\}_{i=1}^{\infty}$, with v_i and u_i being identical to the basis elements in the Karhunen-Loève expansion of Z and ε respectively. The posterior random function $(Z|Y = y)$ can then be represented by the same Karhunen-Loève expansion as the prior, only with different coefficients. Defining $y_i = (u_i, y)$ this reads

$$(Z|Y = y) = \sum_{i=1}^{\infty} (Z_i|Y_i = y_i)\, v_i\,, \tag{18}$$

with $\{(Z_i|Y_i = y_i)\}_{i=1}^{\infty}$ being independent Gaussian random variables with mean $y_i\sigma_i/(\sigma_i^2 + \lambda_i^2/\gamma_i^2)$ and variance $\gamma_i^2[1 - \sigma_i^2/(\sigma_i^2 + \lambda_i^2/\gamma_i^2)]$. The optimal estimator can in this case be found explicitly as

$$\hat{z}(y) = \sum_{i=1}^{\infty} \frac{\sigma_i^2}{\sigma_i^2 + \lambda_i^2/\gamma_i^2} \frac{(u_i, y)}{\sigma_i} v_i$$

Note that this result is of the form in Expression (9).

Nonlinear inverse problems fits equally well into the Bayesian methodology, as linear. The optimal estimator under quadratic loss is again the conditional expectation, and the uncertainty is again described by the posterior distribution. There is no additional problem with identifiability since the posterior is a measure on the parameter space. There is however a computational cost which may be a severe obstacle.

7 Comparison of methodologies

In the previous sections inversion by regularization, statistical minimax inversion and Bayesian inversion, are presented as methodologies to solve inverse problems. In the current section these methodologies are compared.

In the presentation above inversion by regularization is presented as a mathematical approach whereas the other two are presented as statistical approaches.

This classification focuses the observation error, ε. In the mathematical approach the error is chosen by a malicious opponent that always makes the least favorable choice, whereas in the statistical approaches the error is considered random, hence it will change if the experiment is repeated.

Historically the mathematical and statistical approaches are developed separately and different languages have emerged. The result is that different names have been given to the same effect, and similar names have been given to different effects. The first is exemplified by uniqueness in the mathematical language and identifiability in the statistical. An example of the latter is that an admissible regularization strategy relate to an effect in the zero noise limit, while an admissible estimator relates to the performance of a particular estimator regardless of the noise level.

From a mathematical point of view the important notion for the solution is stability and convergence, which is implied by an upper bound on the estimation error in terms of the observation error. These bounds are seldom tight such that tight uncertainty bounds for the approximate solution can be derived. In a discrete problem the maximum estimation error of an approximate solution, can in theory be found using constrained optimization. This is however a hard problem to solve numerically. In the statistical literature the two estimation approaches justify two different strategies to assess the uncertainty. In the minimax approach few techniques are able to assess the uncertainty in the setting of inverse problems, the few rigorous methods stated above are limited to linear inverse problems. In the Bayesian approach the uncertainty is described by the posterior distribution, any probabilistic uncertainty statement regarding the parameter can be deduced from this distribution. Stark (1992) denote the Bayesian uncertainties as formal uncertainties, because they are based on an apriori assumption about the parameter that cannot be verified.

In many respects it is more natural to classify the methodologies by their view on the parameter. Inversion by regularization and statistical minimax inversion regard the parameter as a fixed quantity, while in the Bayesian approach it is considered to be random. In Donoho (1994) a related problem is investigated, a deep connection between two fixed parameter approaches corresponding to those above is found.

Results in any of the three methodologies, require that additional information is given. In the fixed parameter approaches this is done by imposing bounds on the parameter space such as Expression (13). In the Bayesian approach the information is given in terms of a probability measure on $\mathcal{Z}$. The Bayesian approach hence requires stronger assumptions, since the relative importance of any two elements in $\mathcal{Z}$ can be measured.

Consider also the achievement of the methodologies, within their own standard. In the Bayesian approach the optimal estimator under quadratic loss is well defined, i.e. $\mathrm{E}_{Z|y}\{Z\}$, hence it is a computational question to obtain the solution for given set of data. The fixed parameter approaches are more ambitious, but

only rate optimality is established.

Tikhonov regularization is frequently given a Bayesian interpretation, by defining

$$p(z) = const \times \exp\{-\tau^2 \alpha J(z)/2\}$$

and

$$p(y|z) = const \times \exp\{-\tau^2 ||Kz - y||^2/2\}$$

with *const* being a generic normalizing constant; τ being a scaling factor; and $J(z)$, $||Kz-y||^2$ and α being as as for Expression (11). The posterior distribution is then

$$p(z|y) = const \times \exp\{-\tau^2(||Kz - y||^2 + \alpha J(z))/2\}$$

The value of z that maximizes this distribution is denoted the maximum posterior estimate. This is identical to the solution found by Tikhonov regularization, see Expression (11). Although this estimate for computational reasons is commonly used in Bayesian analysis, it is not a proper Bayes estimate in the cases considered here, since no proper loss function correspond to this estimator.

There is also a connection between the statistical minimax inversion and Bayesian inversion. The minimax approach can be seen as a game version of the Bayesian approach. In this game the scientist pick the estimator, $\hat{z}$, whereas the opponent may pick the prior distribution, $\pi(z)$, within a restricted class of distributions. The pay off in this game is the Bayes risk with the prior from the opponent, i.e. $B_{\hat{z}}[\pi(z)]$. This connection is an essential part in the theory of statistical minimax inversion.

8 Conclusions

The three methodologies are all successful for linear inverse problems, and the solutions look surprisingly similar.

The philosophical difference between mathematical approach and the statistical approaches is the nature of the observation error. The mathematical approach considers the worst case error. The statistical approaches regard the error as random. To choose one over the other based on this criterion is a philosophical debate of the nature of the error, ε. The error is caused by many sources. If all of the sources are of equal strength, the central limit theorem, can be used to argue the case for random errors. If some of the sources are dominant, this will produce a systematic error hence the mathematical philosophy would be preferable.

A practical difference between the mathematical approach and the two statistical approaches, is that stability is focused in the mathematical approach whereas uncertainty is focused in the statistical approach. There is a fundamental difference between the two notions, i.e. a solution can be stable and have large

uncertainty. In the authors opinion assessment of uncertainty is an important issue, hence he tends to favor the statistical approaches.

The Bayesian choice of prior distribution is usually criticized in traditional statistics. The critique is not as severe when it comes to the inverse problems considered here since information about the parameter must be included apriori in any case. Non informative prior distributions is in the author opinion only of interest for hyper parameters when inverse problems are considered, since inverse problems requires additional structure to be enforced. The Bayesian methodology achieves more and is more widely applicable, but the Bayesian assumptions regarding the parameters are stronger than that of minimax estimation. Whether the Bayesian achievements are worth the price of stronger assumptions is for the practitioner to decide.

From a purely statistical point of view the minimax estimator usually have better properties and should be preferred. On the other hand minimax estimators are only found for special cases, and are hence generally not available. Phenomena that are studied in inverse problems frequently have spatio-temporal structures, hence modeling the prior by random fields seem natural and may give the Bayesian estimates an advantage. The Bayesian methodology also guarantee the estimator to be admissible. Hence the Bayesian estimators can be used also by non-Bayesian that do not fully believe in the posterior distribution.

When it comes to aspects of uncertainty, the question is whether formal uncertainties are acceptable or not, keeping in mind that the alternative might be no assessment at all. In the authors opinion formal uncertainties are acceptable in any engineering application, but can be questioned for scientific purposes.

Non of the theories are fully developed in the nonlinear case. In the Bayesian approach the optimal estimator is known in theory, but there is no general way to compute it. The choice in the nonlinear case is frequently between Tikhonov regularization and the Bayesian approach, i.e. the maximum aposteriori estimate and the conditional expectation. The authors personal preference is the conditional expectation since this account for many reasonable solutions and is the one where loss criterion carry through to the final estimate also for nonlinear problems.

9 The thesis in the current context

The thesis is fully within the framework of Bayesian inversion, but parts has been inspired by work within the other solution frameworks.

The focus in the introduction is mainly on linear estimators for linear inverse problems. Nonlinear inverse problems and nonlinear estimators are presently subject to extensive research interest, but a complete theory for these type of problems is still lacking.

Paper I is devoted to a particular type of nonlinear inverse problem an practicalities for Bayesian inversion in this case. Paper II and III are specific applications of Paper I. A feature that is of particular interest in Paper III is that the the conditional expectation does not honor the data to the same extent as individual samples. This imply that the maximum posterior estimator is better adapted to the data than the conditional expectation. The reason for this is that the conditional expectation take all of the parameter space into account when producing the final estimator. The maximum posterior estimator identify only one extreme case among many possible. The conditional expectation hence produce a more robust summary of the posterior distribution.

In Paper IV the prior is formulated by smoothing independently scattered random measures. For the Gaussian case this relates to Expression (13) with B being the smoothing kernel. When the prior measure is discretized this correspond to the version presented in Neumaier (1998). The use of both a Gaussian random process and a Cauchy process may be seen as a Bayesian version of basis selection (Donoho and Stark 1989; Donoho and Huo 2001). The choices being a spike basis defined through the Cauchy process or a harmonic basis defined through the Gaussian processes.

The essence of the Paper V is contained in Expression (18). By using the Fourier transform each set of frequency components may be solved independently. The only difference to Expression (18) is that in Paper V several parameters are solved simultaneously using information from several angles. Hence a block version of the result is required. In the context of seismic inversion Expression (18) is obtained when the impedance is estimated from the zero offset data. The fact that the fast Fourier transform correspond to the singular vectors of the discretized problem makes the approach highly efficient.

References

Backus, G.E. "Confidence set inference with a prior quadratic bound", Geophys.J. 97:pp 119-150.

Bertero, M. (1989), "Linear inverse and ill-posed problems", Advances in electronics and electron physics, Vol. 75, pp 1-120.

Donoho, D.L. and Stark, P.B. (1989) "Uncertainty principles and signal recovery", SIAM J. Appl. Math. 49, no. 3, 906–931.

Donoho, D.L. and Huo, X. (2001) "Uncertainty principles and ideal atomic decomposition", IEEE Trans. Inform. Theory 47, no. 7, 2845–2862.

Donoho, D.L. (1994) "Statistical estimation and optimal recovery", Ann. Statist. Vol 22, no. 1, pp 238-270.

Donoho, D.L. (1995) "Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition", Appl. Comput. Harmon. Anal. Vol 2, no. 2, pp

101-126.

Engl, H.W., Hanke, M. and Neubauer, A. (1996) "Regularization of inverse problems", Mathematics and its Applications, 375. Kluwer Academic Publishers Group, Dordrecht.

Hadamard, J. (1902) "Sur les problèmes aux dérivées partielles et leur signification physique", Bull. Univ. Princeton 13, 49.

Hadamard, J. (1923) "Lectures on Cauchy problem i linear partial differential equations", Yale Univ. Press, New Haven.

Hansen, P.C. (1998) "Rank-deficient and discrete ill-posed problems", SIAM Monographs on Mathematical Modeling and Computation. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Johnstone, I.M. and Silverman, B.W. (1990) "Speed of estimation in positron emission tomography and related inverse problems", Ann. Statist. Vol 18. no. 1, pp 251-280.

Johnstone, I.M. and Silverman, B.W. (1991) "Discretization effects in statistical inverse problems", J. Complexity, Vol 7, no. 1, pp 1-34.

Kirsch, A. (1996) "An introduction to Mathematical theory of inverse problems", Vol. 127 of Applied Mathematical Series, Springer, New York.

Lehmann, E.L. and Casella, G. (1998) "Theory of point estimation", Second edition. Springer Texts in Statistics. Springer-Verlag, New York.

Lehmann, E.L. (1999) "Elements of large-sample theory", Springer Texts in Statistics. Springer-Verlag, New York.

Neumaier, A. (1998) Solving ill-conditioned and singular linear systems: A tutorial on regularization,SIAM Rev. 40 (1998), no. 3, 636–666

Robert, C.P. and Casella, G. (1999) "Monte Carlo statistical methods", Springer Texts in Statistics. Springer-Verlag, New York.

Stark P.B. (1992) "Minimax confidence intervals in geomagnetism", Geophys. J. Int. 108:pp 329-338.

Stark, P.B. (2000) "Inverse problems as statistics", Surveys on solution methods for inverse problems, pp 253-275, Springer, Vienna.

Stein, C. (1956) "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution", Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Vol. I, pp. 197–206. University of California Press, Berkeley and Los Angeles.

Wahba, G. (1990) "Spline models for observational data", CBMS-NSF Regional Conference Series in Applied Mathematics, 59. Society for Industrial and Applied Mathematics (SIAM), Philadelphia.

Yaglom, A. M. (1987) "Correlation theory of stationary and related random functions", Vol. I. Basic results. Springer Series in Statistics. Springer-Verlag, New York