TMA 4180 Optimeringsteori
AN INTRODUCTION TO INVERSE PROBLEMS

H. E. Krogstad, IMF, Spring 2007, Rev. May 18.

1 INTRODUCTION

The number of hits on Google on the term *inverse problems* has now exceeded 1.2 million. Yet, most of you have hardly met the term in any of the courses you have been through. The field of inverse problems is definitely a branch of applied mathematics, but deals with situations one tend to avoid in the traditional courses in mathematics. This includes less pleasant situations such as equations with no or infinitely many solutions, missing conditions in a problem, linear equations where the matrix is non-singular in theory, but where even multi-precision arithmetic is of no help. Although Wikipedia gives a reasonable definition of the field, the MathWorld discussion is still rather narrow.

The TV game *Jeopardy* is an example of an everyday inverse problem. If you are given the answer *"It was in 1905"*, your question will depend both on the circumstances and who you are:

- When did Einstein publish his Theory of Relativity?
- When did Robert Koch get the Nobel Prize in Medicine
- When was my grandmother born?
- When did Norway and Sweden split up from the union?

Readers should have no problems with adding other "correct" questions. Thus, the information we are given is incomplete and non-conclusive. Different input leads to the same result and the "most probable" input depends on occasion. All this is typical for inverse problems.

The origin of the term *inverse* problem is simple and mirrors what is called the *forward* (or *direct*) problem. In simple terms, the forward problem is the situation "given the question, find the answer", whereas the inverse problem is "given the answer, find the question".

A better way to express this would be to say that solving an inverse problem is to determine a cause from its effect. In some cases, there is no hope of ever being able to solve the forward problem in full generality.

Applications of inverse problem techniques abound in medical imaging, seismology, geosciences, and many other areas of sciences and engineering. It is probably fair to say that the majority of real world problems are inverse problems.

The French mathematician Jacques Hadamard introduced the term *well-posed* for a mathematical problem where

- a solution always exists

- the solution is unique
- a small change in the initial conditions leads to a small change in the solution

The opposite of a well-posed problem is an *ill-posed* problem:

- a solution may not exist
- there may be more than one solution
- a small change in the initial conditions leads to a big change in the solution

Inverse problems tend to be ill-posed.

With 1.2 million hits, the need for another note about inverse problems is not that urgent. However, the note will be tied up with topics we already have been touching in the optimization course. On the other hand, the topics are flavoured by this author's own experience, and open material from the Internet will be cited freely (but with references where available).

The note is based on a general presentation that has been given to various groups over the last years, but the field is too extensive to be covered in a short talk, and there are major parts of inverse problem theory that are not covered at all here.

2 FAMOUS INVERSE PROBLEMS

Before we dig into the theory in more details, let us briefly discuss some very famous inverse problems.

Can you hear the shape of a drum? This question was first posed by the Hungarian mathematician Mark Kac in 1966. It is a popular phrasing of the mathematical problem of whether the spectrum of the Laplace operator of a 2d domain (the drumhead!) is sufficient for determining its shape. The spectrum is here essentially the eigenfrequencies of the drum head relative to the fundamental frequency. The question raised a lot of interest and was not resolved before 1992, and then in the negative: There are indeed non-isomorphic plane domains with the same eigenfrequencies. One example of *iso-spectral* domains is shown in Fig. 1, and there are many other examples.

It is interesting that even if you can't here the shape of the drum, it is claimed that you can hear the number of holes and the length of the perimeter. The problem is thoroughly discussed on the *Internet*, see e.g. Wikipedia and American Mathematical Society's Web pages.

Computer tomography. The material in the human body has variable penetration to X-rays. By the exposing a cross section of the body to X-rays and measuring as illustrated on Fig. 2, one obtains measurements of the decay along straight lines through the body at varying positions and varying angles.

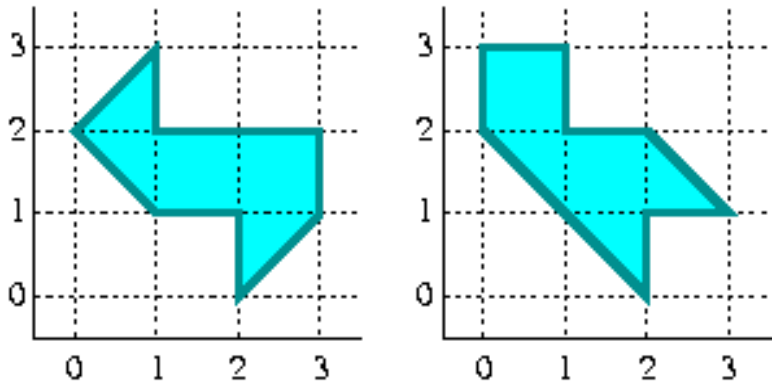


Figure 1: Simple examples of drum heads having identical eigenfrequencies (from Wikipedia article).

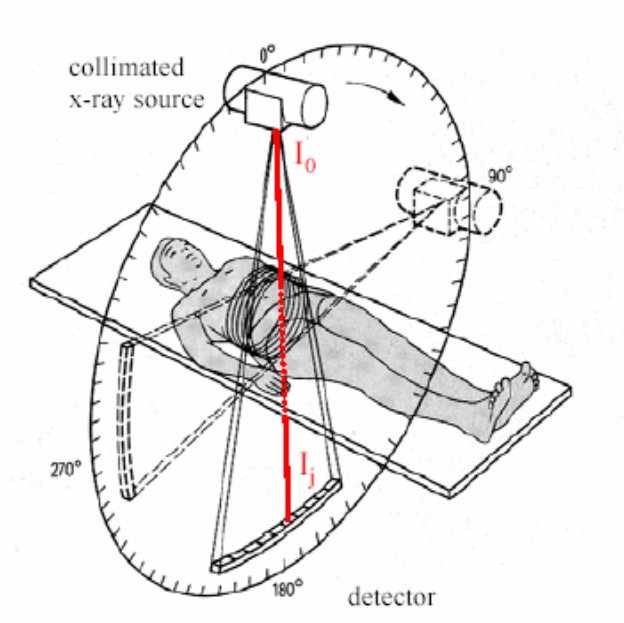


Figure 2: Typical arrangement of body, X-ray source and detectors for computer tomography (Figure copied from a presentation by Dr. G. Lauritsch, Univ. of Heidelberg).

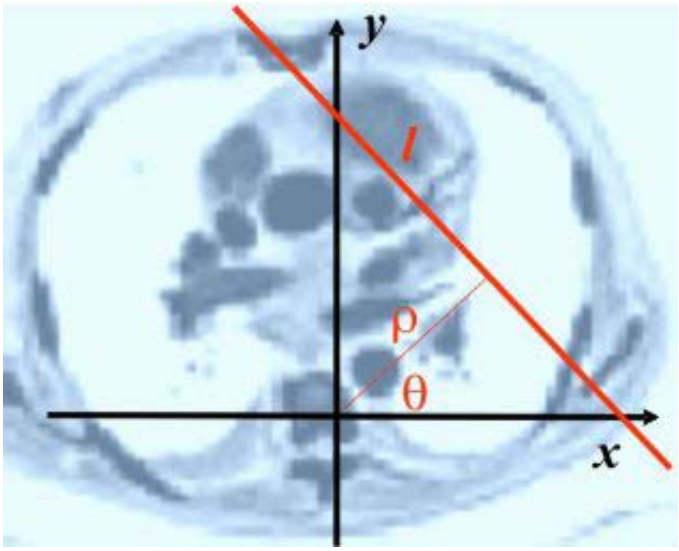


Figure 3: Scanning along a line l defined by ρ and θ.

Mathematically, the set up is as follows: We measure the body in a plane, $\mathbf{x} \in \mathbb{R}^2$, with a variable absorption μ per length unit, that is, $\mu = \mu(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^2$. The decay along a particular line may thus be expressed as

$$I_{\text{line}} = I_0 \exp\left(-\int_{\text{line}} \mu(\mathbf{x}(l))\, dl\right). \tag{1}$$

The line is parametrized in terms of its distance ρ from the origin and the angle θ defining the orientation of the line, as shown in Fig. 3. The integral in Eqn. 1 may then be written

$$\mathcal{R}\mu(\rho,\theta) = \hat{\mu}(\rho,\theta) = \int_{-\infty}^{\infty} \mu(\rho\cos\theta - t\sin\theta, \rho\sin\theta + t\cos\theta)\, \mathrm{d}t, \tag{2}$$

where the $\mathcal{R}$ stands for the *Radon transform*, and

$$t \longrightarrow \{\rho\cos\theta - t\sin\theta, \rho\sin\theta + t\cos\theta\}, \tag{3}$$

is a parametrization of the line. The problem of reconstructing a 2-dimensional function from integrals along line averaged was first considered by Johann Radon around 1917, and computer tomography thus involves an *inversion* of the Radon transform,

$$\mu(\mathbf{x}) = \mathcal{R}^{-1}(\hat{\mu}(\rho,\theta)). \tag{4}$$

Contrary to the Fourier transform, where the inverse transform is very similar to the forward transform, the inverse Radon transform is considerably more tricky, and the inversion formula exists in many different forms, each with their own merits (See Wikipedia article).

If the object happens to be radially symmetric and the scanning lines are parallel (which is sufficient for a radially symmetric body), the inversion reduces to the solution of what is called *Abel's Integral Equation*,

$$\mu(r) = 2\int_r^{\infty} \frac{d\mathcal{R}\mu(\rho,0)}{d\rho} \frac{d\rho}{\sqrt{\rho^2 - r^2}}. \tag{5}$$

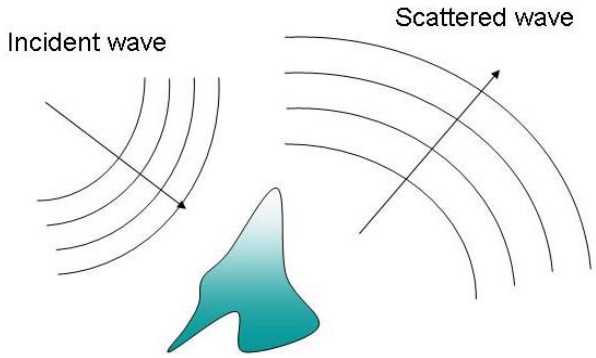


Figure 4: Inverse scattering: The (unknown) object is illuminated by incident waves, and the properties of the object are derived from the scattered wave field.

Actually, the Radon transform and its inversion were re-discovered in the sixties by Allen M Cormack and Godfrey N Hounsfield, who received the Nobel Prize in medicine for this work in 1979 (not without some controversy). F. Natterer's book about CT has become a classic.

Inverse scattering. Inverse scattering is a huge and quite inhomogeneous field. The basic principle consists of a transmitter, emitting some kind of waves, mostly electromagnetic or acoustic, into a medium that scatters the waves. In general, some of the wave is scattered, and some refracted through the medium due to a varying wave velocity. The scattered wave is recorded by one or a set of receivers, a *receiver array*. The situation is illustrated in Fig. 4. Depending on the problem, the angle of the incident wave may be varied, and the scattered waves recorded at many different locations.

A typical situation is often that an incident plane wave hits a small object. The scattered wave will, at large distances from the object, be about spherical, and reconstructing the shape (or the material properties) of the object is then based on the asymptotic properties of the so-called *far field* solution.

One of the most important and mathematically developed inverse scattering topic is seismic processing. Here acoustic waves are transmitted (typically a short wave pulse from some kind of explosive device) into the ground and the scattered acoustic waves are recorded by receivers located at various places on the surface. The scattering occurs mainly from discontinuities in the material properties. However, due to varying wave velocity in various rock types, the received data are severely distorted and require extensive computer processing in order to be usable. Fig. 5 shows an illustration (See Petty for a full tutorial in the reference list).

Today, seismic processing is used to map the whole world. In addition, acoustic arrays are capable to spot small events in the earth's mantle and discriminate between earthquakes and a nuclear explosions.

Based on the seismic recordings, and only hours after the tragic Sumatran tsunami on December 26, 2004, California Institute of Technology was able to provide the world with estimates of the sea floor motion at the continental rift causing the tsunami, see Fig. 6.

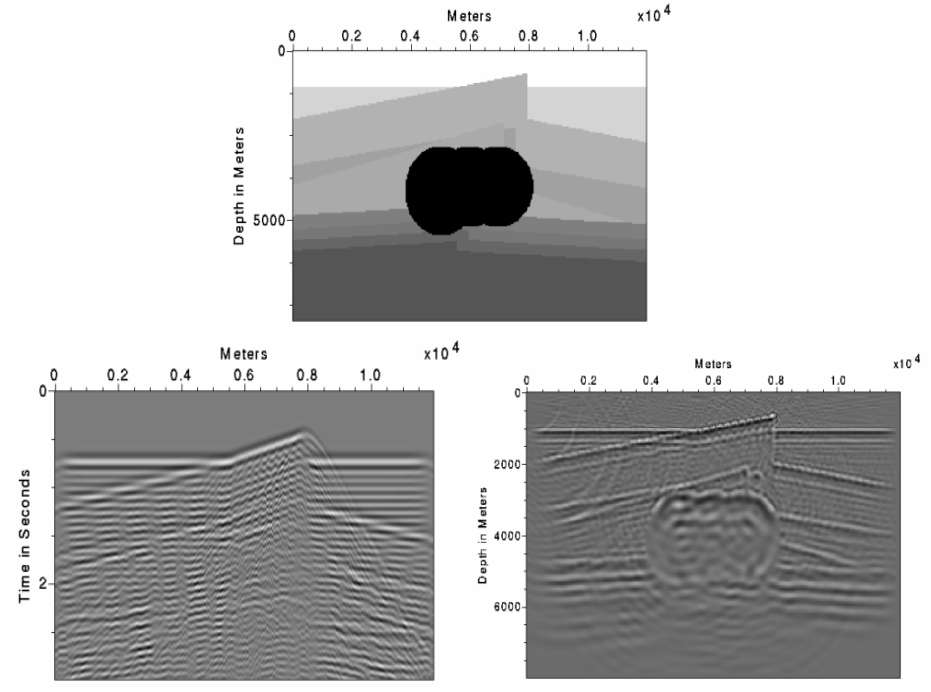


Figure 5: Simulated example demonstrating seismic inversion. Top: Model of the ground; lower left: raw scattered signals; lower right: processed data (Illustration copied from tutorial material prepared by Dr. R. Petty, see references).

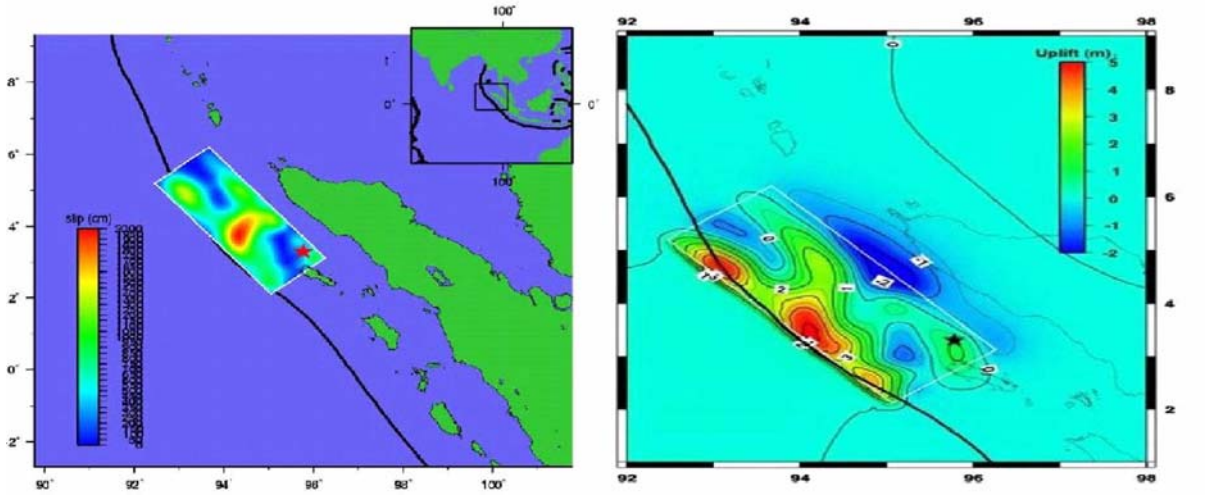


Figure 6: Sea bottom motion causing the Sumatrian tsunami (Graphs obtained from Prof. Bjørn Gjevik, UiO. Originally published by Caltech).

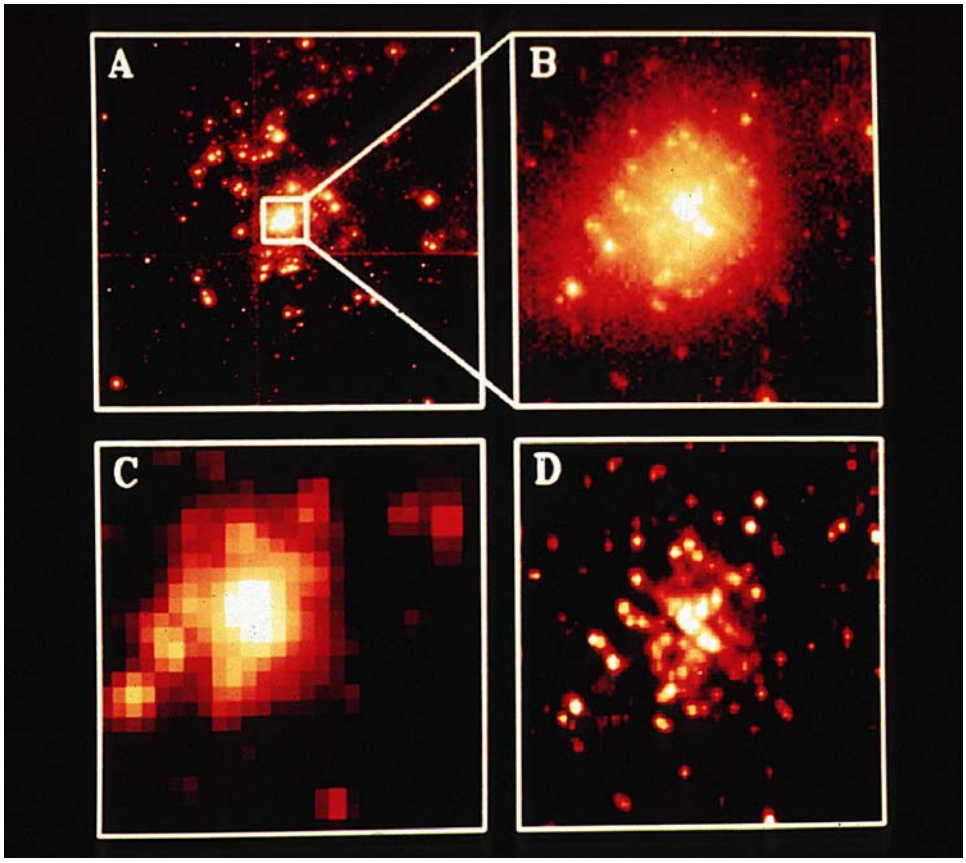


Figure 7: Old example of imagery from the Hubble Space Telescope before adding the correction lens. A: Original HST photo. B: Enlarged section. C: Ground telescope image. D: Digitally enhanced image (©ESA).

Image restoration, With the advent of strong computers, computer image enhancement and restoration has become feasible for everyone. Soon after the Hubble Space Telescope (HST) was put into operation, it was discovered that the telescope mirror was seriously malfunctioning. Images were blurred, and, even if they still were better than those obtained from ground based telescopes, the quality was far below expectations. Since the telescope was already in space, there was no immediate way of improving the situation, and this triggered an intense research in image enhancement techniques. Fig. 7 shows the state of the art in 1990. See the ESA article in the reference list for further information about this figure.

The most popular model describing blurring of images is to model the process by a spatially invariant *Point Spread Function* (PSF), f_{PS}, acting on the ideal image by convolution, that is, a weighted moving average. If we neglect the image boundaries, this may be written

$$BI\left(\mathbf{x}\right) = f_{PS} * I\left(\mathbf{x}\right) = \int_{\mathbb{R}^2} f_{PS}\left(\mathbf{x}-\mathbf{y}\right) I\left(\mathbf{y}\right) d\mathbf{y}, \tag{6}$$

where I is the ideal and BI the blurred image. Contrary to what one would assume, the Fourier transform,

$$\mathcal{F}h\left(\mathbf{k}\right) = \int_{\mathbb{R}^2} e^{-i\mathbf{k}\mathbf{x}} h\left(\mathbf{x}\right) d\mathbf{x}, \tag{7}$$

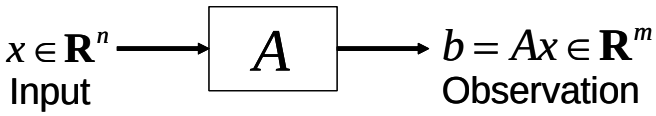


Figure 8: In system engineering, a linear equation is vizualized in terms of input-filter-output.

and the well-known identity

$$\mathcal{F}(BI)(\mathbf{k}) = \mathcal{F}(f_{PS})(\mathbf{k}) \times \mathcal{F}(I)(\mathbf{k}), \tag{8}$$

is seldom of much help. First of all, $\mathcal{F}(BI)(\mathbf{k}) / \mathcal{F}(f_{PS})(\mathbf{k})$ breaks down where the Fourier transform of f_{PS} has zeros, an even if that should not be the case, regions where $\mathcal{F}(f_{PS})$ is small will tend to magnify random errors in $\mathcal{F}(BI)$. The term Point Spread Function is reasonable: An image consisting of a single point represented by a δ-function at $\mathbf{x}_0$, $\delta_{\mathbf{x}_0}$, will be blurred into

$$\delta_{\mathbf{x}_0} * f_{PS}(\mathbf{x}) = f_{PS}(\mathbf{x} - \mathbf{x}_0). \tag{9}$$

In astronomy, a remote star may be used as an excellent δ-function.

3 LINEAR INVERSE PROBLEMS

Before we start, one should note that the form of the answers given below is not necessary expressions one would use in practice. In particular, the *Singular Value Decomposition* (SVD) is a nice mathematical tool, but not always feasible for large problems.

3.1 Noise-free problems

The linear, noise-free inverse problem is nothing but the familiar linear equation. System engineering considers a linear equation $Ax = b$ as a linear filter, where $x \in \mathbb{R}^n$ is the *input* and $b \in \mathbb{R}^m$ the *output*, or the observations, see Fig. 8. The matrix A is a *representation* of the filter, and not so seldom, the filter is only partially known and itself part of the problem.

As long as we are able to test the filter by entering suitable input signals, it is possible to find A by entering any set of n linearly independent input vectors, e.g. the n standard basis vectors.

The trivial case is of course when we know A and b and may just solve the equation, e.g. by standard Gaussian elimination. This leads into the following (typical) cases:

- $m < n$: There are many different solutions fitting the data
- $m = n$: There is one unique solution
- $m > n$: No solution fulfils the system exactly

Inverse problems are typically under-determined, $m < n$, and that is what we consider below, apart from some cases where $m = n$.

As discussed in the Least Square-optimization note, the easiest way to analyze what happens is to apply the Singular Value Decomposition of A. We recall that the (*full*) SVD of an arbitrary $m \times n$ matrix A is

$$A = U\Sigma V', \tag{10}$$

where $U = [u_1, u_2, ..., u_m]$ and $V = [v_1, v_2, ..., v_n]$ have orthogonal columns of norm 1, and the $m \times n$ matrix Σ contains the *singular values* $\{\sigma_i\}$ along the first diagonal. Singular values are non-negative, with $r = \text{rank}\,(A)$ of them being strictly larger than 0. We assume that the singular values are ordered as

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0. \tag{11}$$

Recall that the first r columns of U are a basis for $\mathcal{R}(A)$ (the *range* of A) and the last $n - r$ columns of V are a basis for $\mathcal{N}(A)$ (the *null-space* of A), and the SVD-representation can be reduced to

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i'.$$

The Moore-Penrose generalized inverse of A is the matrix

$$A^+ = \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i u_i'. \tag{12}$$

In terms of the singular value decomposition, the full solution of the Least Square problem

$$x^* = \arg\min_{z \in \mathbb{R}^n} \|b - Az\|_2 \tag{13}$$

is

$$x^* = \sum_{i=1}^{r} \frac{u_i' b}{\sigma_i} v_i + \sum_{i=r+1}^{n} \alpha_i v_i = \sum_{i=1}^{r} \frac{u_i' b}{\sigma_i} v_i + x^{\perp} \tag{14}$$

where $x^{\perp}$ is an arbitrary vector in $\mathcal{N}(A)$.

Many inverse problems deal with huge matrices. E.g., in image restoration, the number of unknowns is equal to the number of pixels, and the matrix of the Point Spread Function (in general varying over the image) will be of dimension equal to the number of pixels squared. As we mentioned when discussing Least Square problems, determining the rank of such a matrix is far from trivial. The common situation is that the singular values decrease gradually to 0, and it is impossible to say where to stop and say that is the rank of A. The nice thing about Eqn. 14 is that we can add one term at a time and stop when the solution starts to get unstable. Small singular values magnify random errors in b, U or V, and destabilize the solution. This is an therefore a typical *ill-conditioned* problem, and a characteristic feature of inverse problems.

A linear equation may also be considered as the finite dimensional analogue of an important class of equations in applied mathematics, namely the *Fredholm integral equations*. The model for the image blur in Eqn. 6 is one example of a Fredholm equation. More generally, we may write

$$y(t) = \int \kappa(t, s)\, x(s)\, ds, \tag{15}$$

where t and s belong to (possibly different) bounded or unbounded sets of real numbers. The κ-function is called the *kernel* of the equation. Typically, the functions x and y will be elements in Hilbert spaces, $\mathbb{H}_1$ and $\mathbb{H}_2$, respectively, and the equation then defines a linear operator, $K : \mathbb{H}_1 \to \mathbb{H}_2$, $y = Kx$. For nice and smooth kernels, the operator K will be *compact*. This means that the map of a bounded set by K will be pre-compact (the closure is compact). Compact operators have a generalized SVD $\{\sigma_k, v_k, u_k\}_{k=1}^{\infty}$, so that we may write

$$Kx = \sum_{k=1}^{\infty} \sigma_k \langle x, v_k \rangle u_k. \tag{16}$$

Moreover, to each distinct singular value σ_k different from 0, there are only finitely many v_k-s and u_k-s. If only finitely many singular values are different from 0, we say that the operator is *degenerate*. Otherwise, if we assume that same ordering as before, i.e. $\sigma_1 \geq \sigma_2 \geq \cdots$, we have

$$\lim_{k\to\infty} \sigma_k = 0,$$

and this is the only limit point.

All this makes the Fredholm integral equations with nice kernels a natural extension of the finite dimensional theory. E.g., the generalized inverse is again of the form

$$K^+ y = \sum_{k=1}^{\infty} \frac{\langle y, u_k \rangle}{\sigma_k} v_k. \tag{17}$$

Unless K happens to be degenerate, solving a Fredholm integral equation will always tend to be ill-conditioned. Engl et al. give an extensive discussion of this and more general operators between Hilbert spaces.

Example: The Hilbert matrix system. The so-called *Hilbert matrix, H*, is the Gram-Schmidt matrix for the linearly independent functions $\{t^n\}_{n=0}^{\infty}$ on the interval $[0, 1]$:

$$H_n = \{h_{ij}\}_{i,j=1}^{n} = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & & & \frac{1}{n+1} \\ \vdots & & & & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & & \cdots & \frac{1}{2n-1} \end{bmatrix}, \; h_{ij} = \int_0^1 t^{i-1} t^{j-1} dt. \tag{18}$$

The matrix is clearly non-singular, but very ill-conditioned: For $n = 20$, the condition number is 1.9×10^{19}, and the singular values and the solution, when using Matlab to solve $H_{20}x = b$, is shown in Fig. 9. Matlab has a carefully coded numerics and displays a strong warning against using the computed solution because of the large condition number. Nevertheless, the solution using only 10 terms in the SVD-expansion is very reasonable. The direct solution in this case is useless, and in fact, the actual outcome varies from computer to computer with different real number representations.

When doing a sequential SVD solution by introducing one term at a time,

$$x^{(k)} = \sum_{i=1}^{k} \frac{u_i' b}{\sigma_i} v_i, \;\; k = 1, \cdots, n, \tag{19}$$

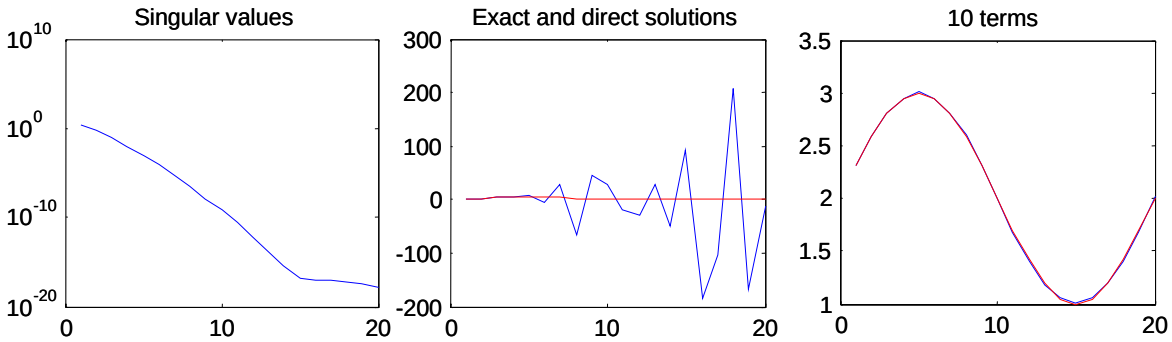


Figure 9: Behaviour of the Hilbert matrix system when $n = 20$. The left plot shows the singular values, which reach the machine accuracy around singular value no. 15. The middle plot shows the exact solution (red) and the solution computed by Matlab (blue). Note the scale on the ordinate axis. Finally, the exact solution (a simple sine-shaped function) and the SVD-solution using only 10 terms is shown to the right.

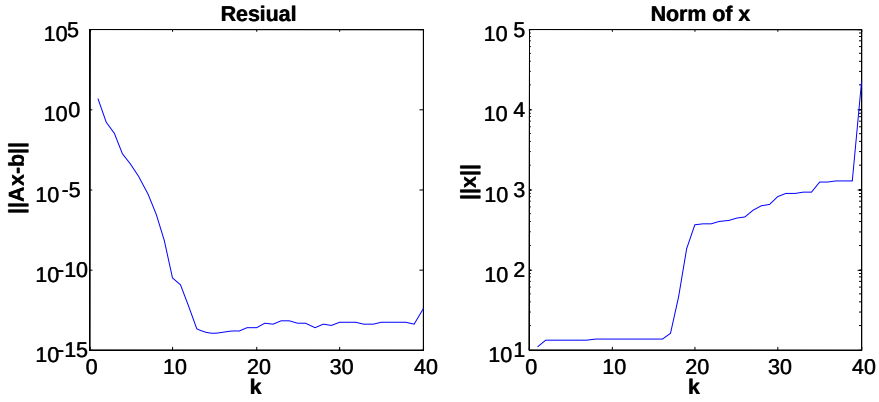


Figure 10: Norms of residuals and solutions for the sequential SVD solution of the Hilbert system for $n = 40$.

it is important to be aware that *a small residual is not a useful criterion*, as shown in Fig. 10 for the solution of $H_{40}x = b$. Even if the residual is small (equal to the machine accuracy), for all k-s larger than about 13, the solution runs terribly off as k increases.

It is mandatory to get control over this instability, and it is now time for introducing the most important and new concept for inverse problems, namely *regularization*.

3.1.1 Regularization based on the SVD

Regularization means to make the solution more regular, which often is the same as more reasonable. There are different ways of applying regularization depending on the problem. To truncate the SVD at a suitable point as indicated above, is a very straightforward way of doing

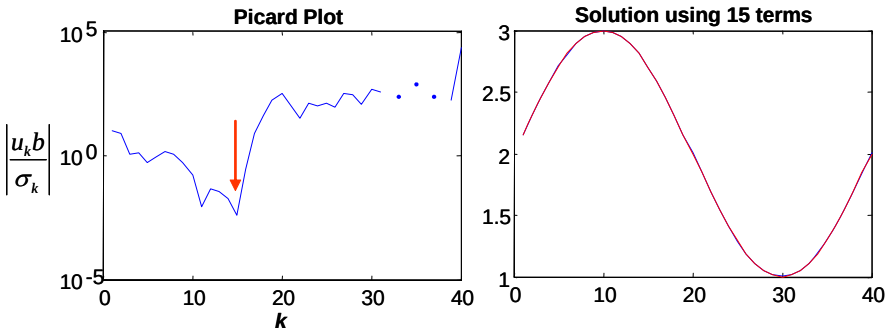


Figure 11: Optimally truncated SVD solution found by means of the Picard Plot (left). Exact (red) and numerical (blue) solution to the right.

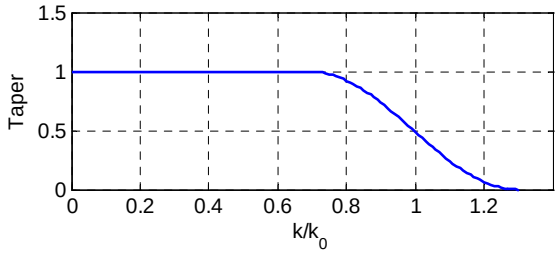


Figure 12: The cosine-bell taper. The tapering zone extends in this case from 0.7 to 1.3.

a regularization. Since the solution is expressed in terms of an orthogonal series,

$$x^* = \sum_{k=1}^{n} \alpha_k v_k, \tag{20}$$

where $\|v_k\| = 1$, and $\alpha_k = u_i' b/\sigma_i$, it is reasonable to display the absolute value of the Fourier coefficients, $|\alpha_k|$, and stop when these are at their smallest. Such a plot is called a *Picard plot*, and an example for the Hilbert problem treated above is shown in Fig. 11.

Experience from Fourier theory has showed that cutting the series abruptly is not always the best way to do it. The sharp cut-off generates so-called *Gibb's oscillations* near discontinuities of the solution (look this up in the literature if it is not familiar). The cure is to introduce a *tapering* function, which is a function of k starting at the value 1 for $k = 1$ and then dropping gently to 0 where we want to have the truncation. A typical taper, the so-called *cosine bell*, $T(k/k_0)$, shown in Fig. 12.

The truncated solutions may now be written

$$x^{(k)} = \sum_{i=1}^{n} T\left(\frac{i}{k}\right) \frac{u_i' b}{\sigma_i} v_i. \tag{21}$$

There is an extensive theory about the performance of various tapers (also called *windows*) in signal processing, e.g. how to choose the width of the tapering zone.

It is sometimes more convenient to express the taper as a function of σ, tending to 1 when σ gets large and to 0 when σ tens to 0 (See the taper derived for the Tikhonov regularization below).

3.1.2 Tikhonov regularization

There are many other ways of introducing regularization into an inverse problem. A rather old idea is to extend the minimization in Eqn. 13 by some penalty term which "punishes" bad behaviour.

Consider again an under-determined linear problem where $A \in \mathbb{R}^{m \times n}$, $m < n$, and assume that we know that the solution should not be that different from x_0. The vector x_0 is sometimes called our *a priori* belief about the outcome. It would then be an idea to modify Eqn. 13 as:

$$x^* = \arg\min_x \left\{ \|Ax - b\|^2 + \mu \|x - x_0\|^2 \right\}, \tag{22}$$

where the positive constant μ determines how much we believe in x_0. Note that the two norms are generally taken in different spaces, so that μ also has to take care of physical dimensions. This approach is called *Tikhonov regularization* and appears to be introduced by Andrey Nikolayevich Tikhonov already in the 1940s (In mathematics, Tikhonov is spelled Tychonoff and famous for his theorem about the topological product of compact spaces).

Problem: *Show that the solution of the problem in Eqn.* 22 *may be written in the alternative forms*

$$\begin{aligned} x^* &= \left(A'A + \mu I\right)^{-1} \left(A'b + \mu x_0\right) \\ &= x_0 + \left(A'A + \mu I\right)^{-1} \left(A'b - A'Ax_0\right) \\ &= x_0 + \sum_{k=1}^{m} \frac{\sigma_k}{\sigma_k^2 + \mu} u_k' \left(b - Ax_0\right) v_k \\ &= x_0 + \sum_{k=1}^{r} \frac{\sigma_k^2}{\sigma_k^2 + \mu} \frac{1}{\sigma_k} u_k' \left(b - Ax_0\right) v_k \end{aligned} \tag{23}$$

Observe that Tikhonov regularization introduces the taper

$$T(\sigma) = \frac{\sigma^2}{\sigma^2 + \mu}, \tag{24}$$

which dampens singular values less than about $\mathcal{O}\left(\mu^{1/2}\right)$. The function is plotted in Fig. 13.

This regularization has some less convenient properties. If we know that x has a strong peak and our a priori belief is slightly wrong about its position, this will show up in the solution as a spurious peak spoiling the quality of the solution. Tikhonov regularization is sometimes used with $x_0 = 0$, thus only punishing the size of $\|x\|$.

3.1.3 Smoothing operators

Regularization is often about punishing irregularities in the solution, and a popular operator for this purpose is the discrete Laplace operator (∇^2). For a one-dimensional problem, the

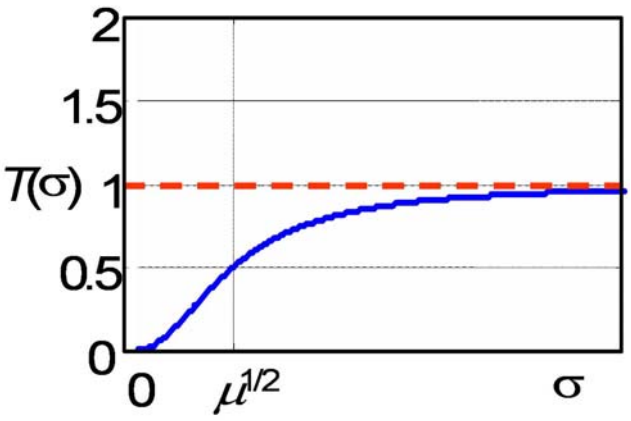


Figure 13: The taper $T(\sigma)$ for Tikhonov Regularization

operator will be proportional to the 2nd difference, D^2x, defined from $\mathbb{R}^n$ to $\mathbb{R}^{n-2}$ as

$$\left(D^2x\right)_{k-1} = -x_{k-1} + 2x_k - x_{k+1},\ k = 2, \cdots, n-1. \tag{25}$$

For higher-dimensional data, the operator is defined accordingly.

Let us denote the smoothing operator by L. The regularized problem takes the form

$$x^* = \arg\min_x \left\{ \|Ax - b\|^2 + \mu \|Lx\|^2 \right\}, \tag{26}$$

with the obvious solution

$$x^* = \left(A'A + \mu L'L\right)^{-1} A'b. \tag{27}$$

Example: The Hodrick-Prescott Filter. We consider yearly mean temperatures in the form of time series with a time step of 1 year, and write the measurements as $\{X_i\}$, where X_i is the measured value at time t_i, $i = 1, \cdots, n$.

The data in this example have been obtained from freely available climatic temperature data at the web location `rimfrost.no`.

In a trend analysis, X_i is expressed as a sum of two parts,

$$X_i = T_i + r_i, \tag{28}$$

where T_i is the *trend* and the remainder, $r_i = X_i - T_i$, is called the *residual*. An example of a trend curve and the residuals for the temperatures from Blindern, Oslo, is shown in Fig. 14. The trend curve is slowly varying, and the residuals spread out evenly around the trend. This is what we appreciate for a good trend curve.

The trend curve in Fig. 14 is produced by applying an irregularity penalty along with a least square deviation from the trend. We are looking for the trend T in the same points as we have the data. Since the trend curve should be centred in the middle of the data, it is first of all reasonable to require that

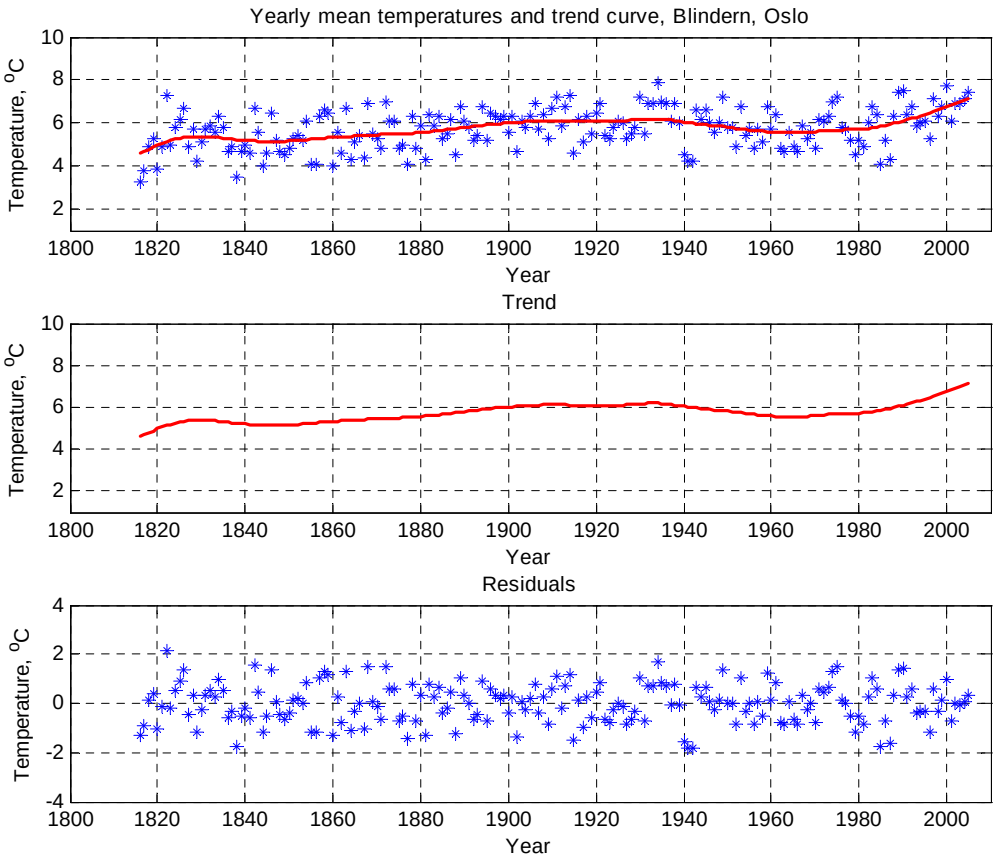


Figure 14: Data, trend and residuals for yearly mean temperatures from Blindern, Oslo. The trend curve is computed by means of the Hodrick-Prescott filter.

$$\sum_{i=1}^{n} (X_i - T_i)^2 \tag{29}$$

is small, but not so small that the curve becomes too irregular. Let D^2T be the operation in Eqn 25. If

$$\sum_{i=1}^{n-2} \left(D^2T\right)_i = 0, \tag{30}$$

all points of T lie on a straight line. It is therefore reasonable to consider the sum of the terms in Eqns. 29 and 30,

$$HP(T,\mu) = \sum_{i=1}^{n} (X_i - T_i)^2 + \mu \sum_{i=1}^{n-2} \left(D^2T\right)_i, \tag{31}$$

and solve

$$T_\mu = \arg\min_T HP(T,\mu). \tag{32}$$

The trend curve ranges from the trivial $T = X$ for $\mu = 0$, to the mean square linear regression when $\mu \to \infty$. Apart from the limiting cases, the solution of the linear least square problem in Eqn. 31 cannot be solved analytically for n of some size. However, no numerical problems was experienced in this example, even if the dimension of the sparse linear system is equal to the number of data points.

Data and trend curves for varying values of μ are shown in Fig. 15. It is not easy to say what is the best.

These trend curves are called *Hodrick-Prescott* curves, and the algorithm is called the Hodrick-Prescott filter, named after the people who introduced the method to the economists in the 1990s (E. C. Prescott got the Nobel Prize in economics for 2004 together with the Norwegian Finn Kydland). Nevertheless, the method is much older, dating back at least to the 1920s.

3.2 Problems containing noise

Real measurements are always suffering from various kinds of noise. The noise is unavoidable and may added to the signal, multiplied to the signal, or interfering in some other and more complicated way. The analysis of random noise requires elements of stochastic signal analysis which will not be covered in detail here.

The standard and very common situation is to have some additive noise in the data, as illustrated in Fig. 16. A common source of noise is introduced by the digitalization of analog signals. The noise may come from time or space digitization, as well as from the discretization of the data values (called *quatization noise*). A CD or video recording is a good example.

Let us consider the effect of the noise on the SVD-solution $x^* = \sum_{j=1}^{r} \frac{u_j' b}{\sigma_j} v_j$. Using the basis

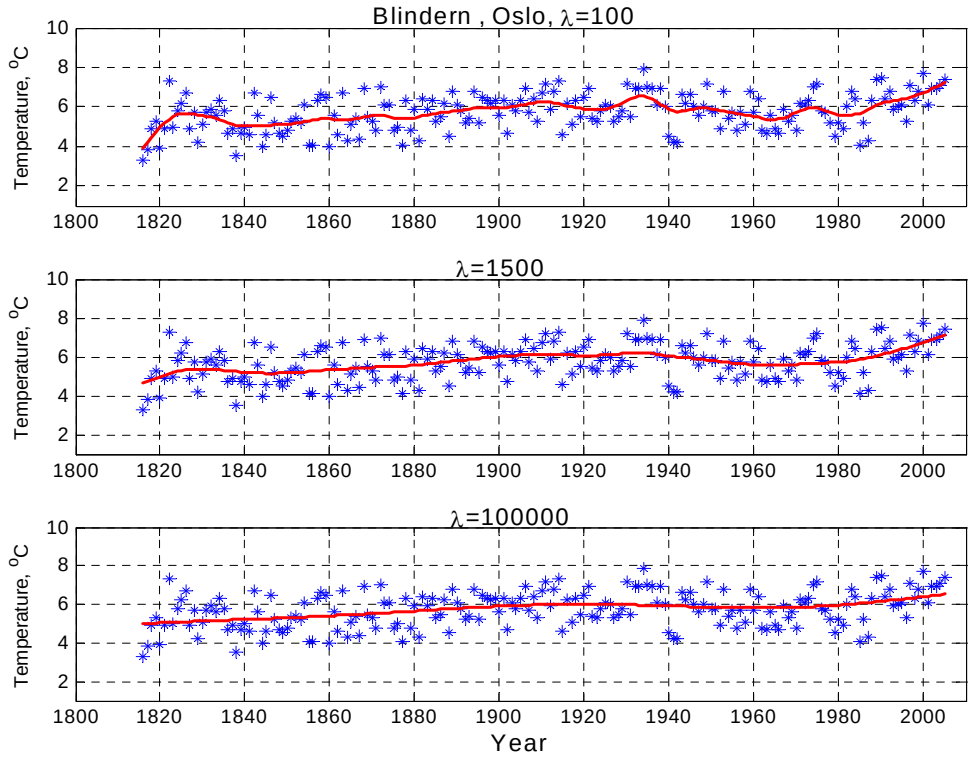


Figure 15: Examples of trend curves for varying values of μ $(= \lambda)$ in the Hodrick-Prescott filter.

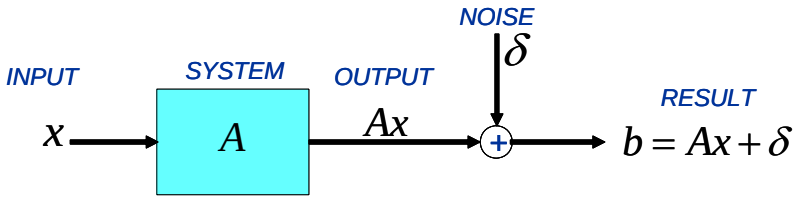


Figure 16: System with additive noise in the result

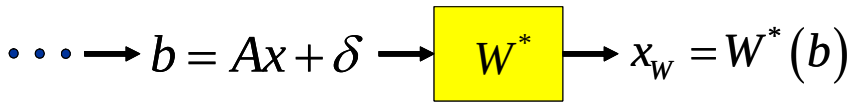


Figure 17: The Wiener filter puts the data through an additional filter which tries to reproduce the input as closely as possible in an averaged sense.

vectors in V and U, we may write

$$x = \sum_{i=1}^{n} \alpha_i v_i, \tag{33}$$

$$\delta = \sum_{j=1}^{m} \eta_j u_j. \tag{34}$$

Since

$$Ax + \delta = \sum_{k=1}^{m} (\sigma_k \alpha_k + \eta_k) u_k, \tag{35}$$

and since the u_i-s are orthogonal, $u_j' u_k = \delta_{jk}$, this gives

$$x^* = \sum_{i=1}^{r} \frac{u_i' (Ax + \delta)}{\sigma_i} v_i = \sum_{i=1}^{r} \left(\alpha_i + \frac{\eta_i}{\sigma_i} \right) v_i. \tag{36}$$

Noise on the components where σ_i is small is therefore particularly serious, although what matters is the relative size of α_i compared to η_i/σ_i.

This introduces a connection between the regularization and the noise level: It is necessary to choose the regularization, say the cut-off in the SVD expansion in Eqn. 36, *before* the noise starts to dominate.

3.3 Choosing the Right Regularization

Since the amount of necessary regularization depends on the noise level, inverse problem theory has developed many rules which, for a given noise level, determine the amount of regularization. In the following, we discuss some of the most popular methods.

3.3.1 The Wiener Filter

The *Wiener filter*, introduced by Norbert Wiener in the 1940s, is based the idea of continuing the graph in Fig.16 as shown in Fig. 17. The idea is to choose the filter W^* so that $W^*(b)$ is close to x in an *average* optimal way,

$$W^* = \arg\min_W \|W(Ax + \delta) - x\|_S, \tag{37}$$

where $\|\|_S$ is a norm that averages over the signals we have at our disposal.

We shall here only derive the filter for a very simple case, where we again utilize the SVD of the A-matrix, $A = U\Sigma V'$.

Assume that the collection (*ensemble*) of all input signals are multivariate stochastic variables in the form

$$x = \sum_{i=1}^{n} \alpha_i v_i, \tag{38}$$

where $\{\alpha_i\}$ are independent stochastic variables with zero mean (for convenience) and variance $\mathrm{Var}(\alpha_i) = s_i^2$. Similarly, we assume that the noise δ is written

$$\delta = \sum_{j=1}^{m} \eta_j u_j, \tag{39}$$

with $\mathsf{E}\eta_j = 0$ and $\mathrm{Var}(\eta_j) = n_j^2$. Typically, n_j is assumed to be constant, or what is called *white noise*. Thus, we have

$$\mathsf{E}\|x\|^2 = \sum_{i=1}^{n} \mathrm{Var}(\alpha_i) = \sum_{i=1}^{n} s_i^2, \tag{40}$$

$$\mathsf{E}\|\delta\|^2 = \sum_{j=1}^{m} n_j^2. \tag{41}$$

The $\|\|_S$-norm averaging over the signals is therefore

$$\|x\|_S = \mathsf{E}\|x\|^2. \tag{42}$$

From the SVD of A we obtain

$$Ax + \delta = \sum_{j=1}^{m} (\sigma_j \alpha_j + \eta_j) u_j. \tag{43}$$

Since we consider the Wiener filter as some kind of generalized inverse, we look for an optimum filter of the form

$$W = V\Sigma_W U', \tag{44}$$

where $(\Sigma_W)_{jj} = w_j$. Thus,

$$W(Ax + n) - x = \sum_{i=1}^{m} \{w_i(\sigma_i\alpha_i + \eta_i) - \alpha_i\} v_i - \sum_{i=m+1}^{n} \alpha_i v_i. \tag{45}$$

The last term in the sum, belonging to $\mathcal{N}(A)$, is beyond our control. For the first term we have

$$\mathsf{E}\left\|\sum_{i=1}^{m} \{w_i(\sigma_i\alpha_i + \eta_i) - \alpha_i\} v_i\right\|^2 = \sum_{i=1}^{m} \left[(w_i\sigma_i - 1)^2 s_i^2 + w_i^2 n_i^2\right]. \tag{46}$$

Since each term is independent of the others, we can minimize the sum term by term. We leave to the reader to show that the optimum values of w_j are

$$w_j = \frac{\sigma_j s_j^2}{\sigma_j^2 s_j^2 + n_j^2}, \ j = 1, \cdots, m \tag{47}$$

Applying the optimal Wiener filter, we thus obtain

$$x_{WF} = \sum_{j=1}^{m} \frac{\sigma_j s_j^2}{\sigma_j^2 s_j^2 + n_j^2} \left(u_k' b\right) v_k \tag{48}$$

The coefficients reduce to the standard generalized inverse when $n_j = 0$, and we also see the similarity with the Tikhonov regularization taper when $\{s_j\}$ and $\{n_j\}$ are constant.

The Wiener filter is an important tool in signal processing. Look up the Wikipedia article about the Wiener filter for other formulations depending on the system.

3.3.2 Morozov's Discrepancy Principle

If we know that our noisy data vector b_δ deviates from the exact data vector b at most with an amount δ,

$$\|b_\delta - b\| \leq \delta, \tag{49}$$

we have to accept all solutions x where

$$\|Ax - b_\delta\| \leq \delta. \tag{50}$$

However, not all these x-s are equally reasonable. As we saw above for the ill-conditioned Hilbert matrix problem, x can be quite large even if the residual is very small (In an infinite-dimensional Hilbert space, the set of x-vectors defined by the condition in Eqn. 50 is even unbounded, unless $\mathcal{R}(A)$ is closed).

Assume that we have, as in Tikhonov regularization, a method with parameter μ determining the amount of regularization. The idea of *Morozov's Discrepancy Principle* is to choose the largest possible (i.e. the most cautious) μ, where x_μ satisfies

$$\|Ax_\mu - b_\delta\| \leq \delta. \tag{51}$$

Recall the solution of the Tikhonov regularization problem with an a priori assumption x_0:

$$x_\mu = x_0 + \sum_{k=1}^{m} \frac{\sigma_k^2}{\sigma_k^2 + \mu} \frac{1}{\sigma_k} u_k' \left(b_\delta - Ax_0\right) v_k. \tag{52}$$

We assume that

$$\|Ax_0 - b_\delta\| > \delta, \tag{53}$$

so that x_0 is not an acceptable solution from the start. It is obvious that the last term in Eqn. 52 tends to 0 when $\mu \to \infty$, so that 51 will be violated for large enough μ-s. However, it is *not* obvious that there exists a μ so that x_μ satisfies 51. This will definitely not hold if the distance from $\mathcal{R}(A)$ to b_δ is larger than δ, that is,

$$\min_{x \in \mathbb{R}^n} \|Ax - b_\delta\| > \delta. \tag{54}$$

We leave to the reader to show, using the expression in Eqn. 52, that

$$\|Ax_\mu - b_\delta\| \underset{\mu \to 0}{\longrightarrow} 0 \tag{55}$$

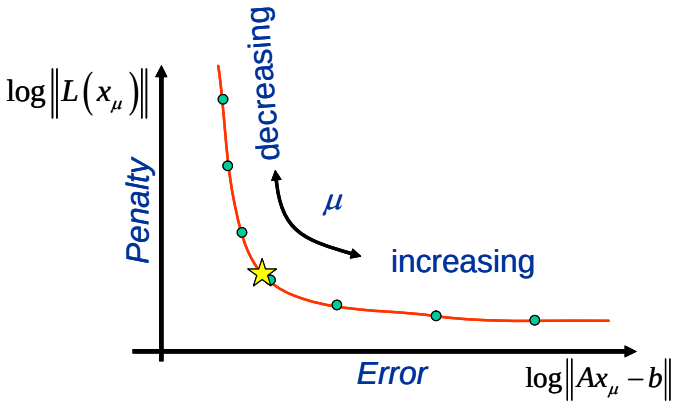


Figure 18: Definition of the L-curve.

if $b_\delta \in \mathcal{R}(A)$ (First show this when $x_0 = 0$. Then prove that the part in Eqn. 52 involving x_0 tends to 0 when multiplied by A). From this, it is easy to see that there will be acceptable solutions as long as the distance between $\mathcal{R}(A)$ and b_δ is a little less than δ. In this case, Morozov's Discrepancy Principle suggests that the proper regularization is

$$\mu_{MDP} = \arg \max_{\|Ax_\mu - b_\delta\| \leq \delta} \mu \tag{56}$$

3.3.3 The L-curve

The *L-curve* as a general method for selecting the best regularization has been introduced by Per Chr. Hansen, DTU. The idea is simple and again illustrated by means of Tikhonov regularization in the form of an error term,

$$\|Ax_\mu - b\|, \tag{57}$$

and a penalty term

$$\|Lx_\mu\|. \tag{58}$$

The L-curve is simply the trace of $\{\log(\|Ax_\mu - b\|), \log(\|Lx_\mu\|)\}$ when μ varies, and the optimal value is found where the curve has its largest curvature. An example where the L-curve works extremely well has been copied from P.C. Hansen's book. The equation is the Fredholm integral equation

$$\int_{-\pi/2}^{\pi/2} \kappa(t,s)\, x(s)\, ds = b(t), \tag{59}$$

with the kernel

$$\kappa(t,s) = (\cos s + \cos t) \frac{\sin^2(\pi(\sin s + \sin t))}{\pi^2(\sin s + \sin t)^2}, \tag{60}$$

and a right hand side b computed from the prescribed solution

$$x_{sol}(t) = a_1 \exp\left(-\frac{(t-t_1)^2}{c_1^2}\right) + a_2 \exp\left(-\frac{(t-t_2)^2}{c_2^2}\right) \tag{61}$$

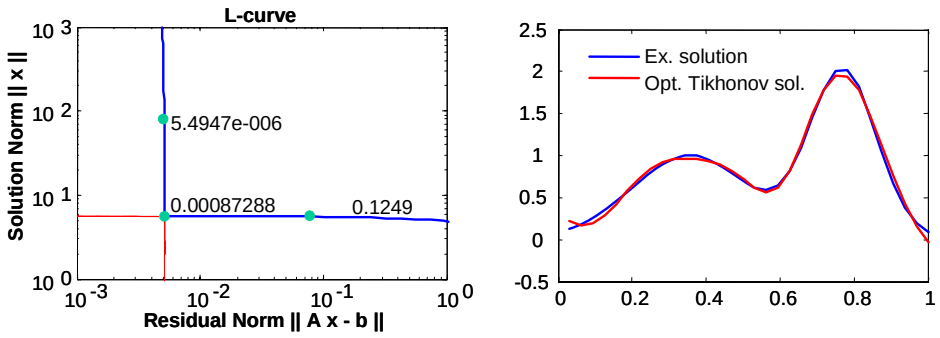


Figure 19: The L-curve to the left with 3 values of μ indicated. The optimal point is very well defined in this case.

The Tikhonov Regularization is carried out with a penalty for a large solution norm,

$$x_\mu = \arg\min \left\{ \|Kx - b\|^2 + \mu \|x\|^2 \right\}, \tag{62}$$

and the result is shown in Fig. 19.

Unfortunately, the L-curve does not always work well. In fact, it appears to be of little value for the Hodrick-Prescott filter. Per Chr. Hansen's home page (see the references) contains several notes related to the L-curve and many other interesting inverse problems.

4 ITERATIVE METHODS

Iterative methods find the solution in a step-wise manner, and in traditional courses in mathematics, the focus is on convergence. The iteration should converge towards a solution, preferably independent of the starting point. The same methods may also be used for inverse problems, but in this case, one does not care about the convergence of the series. It is even sometimes used where it is easy to see that the method finally diverges.

The trick is to start the iteration, and then stop at the right step, somewhat similar to computing with asymptotic series.

If we return to $Ax = b$ for huge matrices, such matrices are often sparse, and whereas it may be quite fast to carry out a matrix-vector product, say Ay for an arbitrary y, carrying out an LU-factorization will be slow or even practically impossible if the storage of the full matrices in core is impossible. This situation calls for some iterative solution, e.g. the Conjugate Gradient (CG) method when $A > 0$. The Conjugate Gradient method is discussed as an iterative method for the normal equations in Engl et al., Ch. 7.

For a quadratic system, $Ax = b$ where $A > 0$, the simplest iteration is

$$x_{k+1} = x_k + \omega (b - Ax_k), \tag{63}$$

where the *relaxation parameter,* ω, is chosen so that

$$\|I - \omega A\| < 1. \tag{64}$$

Equation 63 is recognized as the fixed-point iteration for

$$x = F(x) = x + \omega(b - Ax), \tag{65}$$

and Eqn. 64 as a sufficient condition for

$$\sup_{x,y} \frac{\|F(x) - F(y)\|}{\|x - y\|} < 1. \tag{66}$$

Problem: *Prove that with* $A > 0$ *and* $x_0 = 0$,

$$x_k = \sum_{j=0}^{k-1} (I - \omega A)^j (\omega b), \tag{67}$$

and $\|x_k - A^{-1}b\| \underset{k\to\infty}{\longrightarrow} 0$ *when* $\|I - \omega A\| < 1$ (*Hint:* $A^{-1} = [I - (I - \omega A)]^{-1}\omega$).

The corresponding iteration for the normal equations $A'Ax = A'b$ has the form

$$x_{k+1} = x_k + \omega A'(b - Ax_k), \tag{68}$$

and for $\omega = 1$, this iteration is called *Landweber Iteration* (Engl et al., Ch. 6).

The iteration defined by Eqn. 63 is used in many different contexts, also for nonlinear problems where an analytical analysis is out of reach. When data and input are of the same dimension, the problem at hand will be to solve a nonlinear equation, which we write, in order to avoid confusion with the linear case,

$$M(x) = d. \tag{69}$$

Exactly as above, it is now reasonable to try

$$x_{k+1} = x_k + \omega(d - M(x_k)). \tag{70}$$

This type of iteration is particularly attractive for large problems when the forward problem, $x \to M(x)$, is relatively easy to solve, as will be the case for convolution operators where we can utilize the Fast Fourier Transform. The method is popular, sometimes said to be "so popular that it is re-invented twice a year" and in many application areas known under the name of the *Van Cittert Deconvolution Method*.

In electromagnetic, electron, and NMR spectroscopy, the spectrum (as a function of frequency) is highly irregular with typically high narrow (δ-function like) peaks. However, since the instrument is not perfect, the peaks are "blurred", and will potentially hide smaller peaks that are neighbours to large peaks. The issue it then to "de-blur" or *deconvolve* the data and obtain a more accurate result with more details.

The example in Fig. 20 illustrates this for simulated data. The blurring is a gaussian moving average, and the "measurement", m_b, shown in the top plot, is simply the convolution between the gaussian, $g(\nu)$, and the ideal data, $m(\nu)$,

$$m_b(\nu) = g * m(\nu) = \int_{\nu'} m(\nu')\, g(\nu - \nu')\, d\nu'. \tag{71}$$

In the middle, we have the carried out 2000 iterations according to the formula

$$X_{k+1}(\nu) = \max[0, X_k + \omega(m_b(\nu) - g * X_k)] \tag{72}$$

with $\omega = 1$. The iteration breaks down for ω a little larger than 2. Also, if we know that the exact function should be nonnegative, we may enforce that during the iterations. Compared to the original measurements, the deconvolution has recovered the two small peaks around 0.1, the broad peak at 0.2 and also the remaining narrow peaks. Around the most prominent peak at 0.44, the solution has however become quite irregular. This kind of "over-compensation" is common for Van Cittert deconvolution and related to the Gibbs oscillations in Fourier theory. In conclusion, the de-blurring has only been partly successful.

It should be added that a Gaussian blur is a simple form of blurring, being monotone in the sense that if the original function increases, then so does the blurred function. Van Cittert deconvolution probably works best under such circumstances.

Here we have also assumed that the blurring function is known. There are, however, special techniques for deriving the form of the blurring function (Point Spread Function) along with the deconvolution, called *blind deconvolution*, which we not go into here.

In the following example we show the Van Cittert method applied to the most famous picture in image processing, *Lena*. The lady is Lena Söderberg (born Sjöblom) from Sweden and the picture is from *Playboy*, November 1972. Interestingly enough, and perhaps typical for the pre-Internet period, Lena Söderberg was not aware that her picture was extensively used in image processing research before 1988! The history of the picture is available on `http://ndevilla.free.fr/lena/`.

The 255 gray level picture here has 512×512 pixels and the blurring function is an approximately two-dimensional Gaussian bell $g(\mathbf{x})$ shown in Fig. 21. The deconvolution formula is similar to Eqn. 72.

The result from the deconvolution using up to 20 iterations with $\omega = 2$ is shown in Fig. 22. Increasing ω beyond 2 leads to instabilities as seen in Fig. 23.

Finally, Fig. 24 shows the original, the blurred and the restored picture after 20 iterations.

5 THE MAXIMUM ENTROPY PRINCIPLE

This, still somewhat controversial, principle for solving incompletely defined problems was introduced by E.T. Jaynes in 1957. Read about the fascinating and diverse history of *entropy* in physics and informatics on Wikipedia!

This section requires some rudimentary knowledge about variational calculus.

Consider the collection of probability distributions $p(t)$ defined on a set Ω,

$$\mathcal{D} = \left\{ p(t); t \in \Omega,\ p(t) \geq 0, \int_\Omega p(t)\, dt = 1 \right\}. \tag{73}$$

Note that $\mathcal{D}$ is convex. The entropy (also called *Shannon Entropy* or *Information Entropy*) of

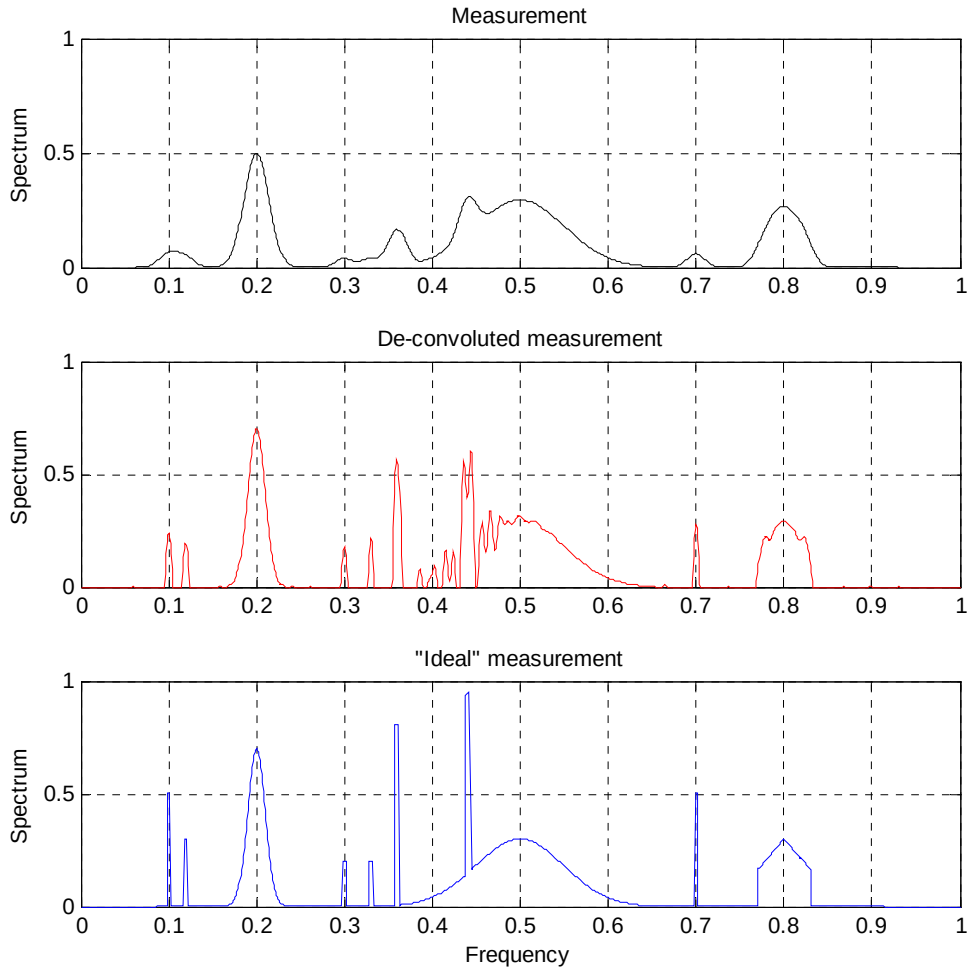


Figure 20: Numerical simulation of deconvolution using the Van Cittert method. Top: "Measured" signal; Middle: Deconvoluted signal; Lower: Ideal signal.

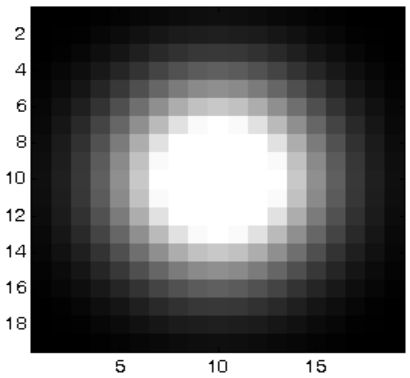


Figure 21: Point Spread Function used in the Lena example.



Figure 22: Van Cittert deconvolution applied to a blurred version of *Lena*.



Figure 23: Same as previous figure, but with $\omega = 2.3$. The deconvolution is no longer stable, resulting in over-compensation. In this case the iteration clearly has to stop at the right place.



Figure 24: Original image (center), blurred image to the left, and deconvoluted image ($\omega = 2$, 20 iterations) to the right.

a probability distribution was introduced by Claude E. Shannon in 1948,

$$H(p) = -\int_{\Omega} p(t) \ln p(t)\, dt. \tag{74}$$

A similar expression applies for a discrete distribution. The entropy expresses the degree of *uncertainty* about a stochastic variable: Consider a binary variable taking the value 0 with probability a and the value 1 with probability $1-\alpha$. Compare H for *no* uncertainty ($\alpha \to 0$ or 1) with *maximum* uncertainty ($\alpha = 1/2$).

If $\int_{\Omega} dt = 1$, $H(p)$ is maximized for $p(t) \equiv 1$. Try to figure out your own argument, or consider the convex Lagrangian

$$\mathcal{L}(p, \lambda) = \int_{\Omega} p(t) \ln p(t)\, dt + \lambda \left(\int_{\Omega} p(t)\, dt - 1 \right) \tag{75}$$

and solve $\delta L(p, v, \lambda) = 0$.

Assume that measurements of the stochastic variable X have given us some information about its probability distribution p. Quite often we'll have estimates about integral properties like

$$\int_{\Omega} g_j(t)\, p(t)\, dt = d_j, \;\; j = 1, \cdots, m. \tag{76}$$

Recall that the measured mean and the mean of the squares are of this form. In addition, we have the requirement $\int_{\Omega} p(t)\, dt = 1$. The question is now: What probability distribution p does this fit? Let us maximize the entropy (equivalent to minimizing the negative entropy) subject to the constraints enforced by the data equations in Eqn. 76.

$$\min_{p \in \mathcal{D}} \mathcal{L}(p, \lambda) = \min_{p \in \mathcal{D}} \left\{ \int_{\Omega} p(t) \ln p(t)\, dt - \sum_{j=1}^{m} \lambda_j \left(\int_{\Omega} g_j(t)\, p(t)\, dt - d_j \right) \right\} \tag{77}$$

Obviously, allowed variations $v(t)$ in p have to satisfy $p(t) + \nu(t) \geq 0$ and $\int_{\Omega} v(t)\, dt = 0$. If we compute the derivative, and set this equal to 0

$$\delta L(p, v, \lambda) = \int_{\Omega} \left\{ \ln p(t) + 1 - \sum_{j=1}^{m} \lambda_j g_j(t) \right\} v(t)\, dt = 0, \tag{78}$$

and

$$p(t) = \exp\left(\sum_{j=1}^{m} \lambda_j g_j(t) \right). \tag{79}$$

It remains to find $\lambda_1, \cdots, \lambda_m$, and this amounts to solving the m nonlinear equations

$$\int_{\Omega} g_j(t) \exp\left(\sum_{j=1}^{m} \lambda_j g_j(t) \right) dt = d_j, \;\; j = 1, \cdots, m. \tag{80}$$

Problem: *For $\Omega = R$, solve the problem*

$$p^* = \arg\min_{p \in \mathcal{D}} \{-H(p)\}, \tag{81}$$

when

$$\int_{\mathbb{R}} p(t)\,dt = 1, \tag{82}$$

$$\int_{\mathbb{R}} tp(t)\,dt = 0, \tag{83}$$

$$\int_{\mathbb{R}} t^2 p(t)\,dt = 1. \tag{84}$$

The maximum entropy functional serves as a regularization for severely under-determined inverse problems. If we have some a priori information we would like to incorporate into the regularization, this may be done by using the *cross-entropy*, defined as

$$H(p) = -\int_{\Omega} p(t) \ln\left(\frac{p(t)}{p_0(t)}\right) dt. \tag{85}$$

Show that the global maximum of H is now obtained for $p = p_0$.

6 EPILOGUE

This summary of inverse problems is quite incomplete. In particular, probabilistic and Bayesian reasoning for inverse problems is a fast growing field which has been left out. The freely available book of Tarantola in the reference list discusses this in detail and contains many additional references.

Another omission is the inverse scattering theory for nonlinear partial differential equations developed during the last 50 years.

Internet encyclopedias such as Wolfram MathWorld and Wikiperia are important sources for further information.

7 REFERENCES

Engl, H.W., M. Hanke and A. Neubauer: *Regularization of Inverse Problems*, 1996, Kluwer, Dordrecht.

ESA: Hubble Space Telescope Photographs Extragalactical Stellar Nursery, Aug. 13, 1990. `http://sci.esa.int/science-e/www/object/index.cfm?fobjectid=37715`

Golub, G.H. and C.F. VanLoan: *Matrix Computations*, Johns Hopkins University Press, 1989.

Gordon, C., D. Webb, and S. Wolpert: Isospectral plane domains and surfaces via Riemannian orbifolds, *Inventiones mathematicae,* **110** (1992) 1–22.

Hansen, P.C. : *Rank-Deficient and Discrete Ill-Posed Problems: Numerical Aspects of Linear Inversion*, SIAM Monograph Math. Mod. and Comp. Vol. 4.

Hansen, P. C.: *Home page*: `http://www.imm.dtu.dk/~pch/`

Kac, M: Can one hear the shape of a drum? *Am. Math. Monthly* **73**(4), 1966, Part II, 1–23.

Morozov, V.A.: *Regularization Methods for Ill-Posed Problems,* CRC Press, Florida, 1993.

Natterer, F.: *The Mathematics of Computerized Tomography*, Classics in Applied Mathematics, SIAM, Vol. 32, 2002.

Petty, R: Seismic Wave Propagation Modeling and Inversion. Power Point presentation at `http://www.mgnet.org/~douglas/Classes/cs521-s00/asdf/asdf.ppt`

Tarantola, A.: *Inverse Problem Theory and Methods for Model Parameter Estimation* (free PDF version at `http://www.ipgp.jussieu.fr/~tarantola/Files/Professional/Books`), SIAM, 2005. ISBN 0-89871-572-5.

Wikipedia: `http://en.wikipedia.org/wiki/Hearing_the_shape_of_a_drum`

Wikipedia: `http://en.wikipedia.org/wiki/Radon_Transform`

Wikipedia: `http://en.wikipedia.org/wiki/Tikhonov_regularization`

Wikipedia: `http://en.wikipedia.org/wiki/Wiener_filter`

Wikipedia: `http://en.wikipedia.org/wiki/Entropy`