Using Prior Knowledge in Inverse Problems

Charles Byrne
(Charles_Byrne@uml.edu)
http://faculty.uml.edu/cbyrne/cbyrne.html
Department of Mathematical Sciences
University of Massachusetts Lowell
Lowell, MA 01854, USA

March 11, 2013

Talk Available on Web Site

This slide presentation and accompanying article, with more detail and references, are available on my web site, http://faculty.uml.edu/cbyrne/cbyrne.html ; click on “Talks”.

Contents

Overview

It is important point to keep in mind when doing signal and image processing that,

- **1.** While the data is usually limited, the information we seek may not be lost.
- **2.** Although processing the data in a reasonable way may suggest otherwise, other processing methods may reveal that the desired information is still available in the data. The first figure illustrates this point.

Using Prior Knowledge

For under-determined problems, prior knowledge can be used effectively to produce a reasonable reconstruction.

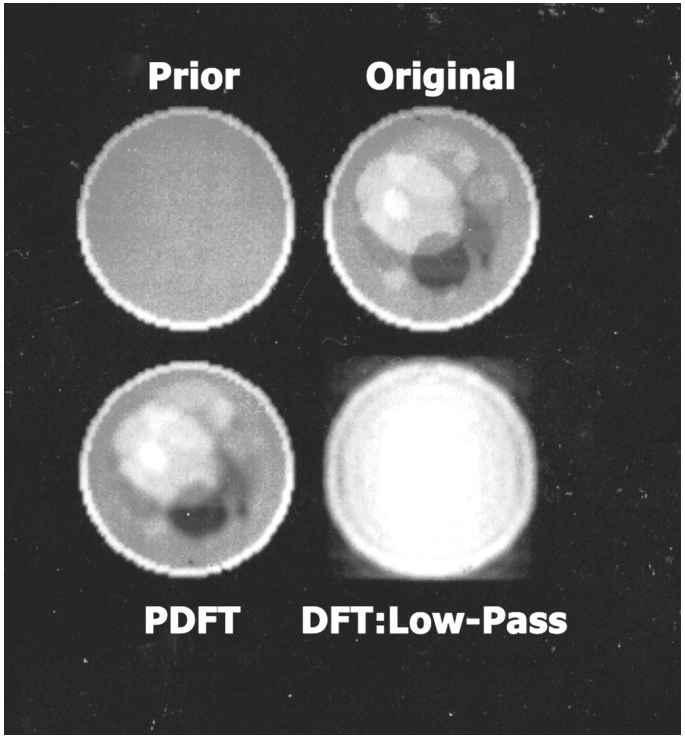

Figure : Minimum Norm and Minimum Weighted Norm Reconstruction.

The Basic Problem

We are concerned with problems of reconstruction of a function of one or more variables, call it $f(x)$, from limited measurements. The x is usually continuous initially, but we may choose to discretize x at the start. Either way, the problem is typically under-determined; there are infinitely many functions that agree with the measurements. How should we select one? Our prior knowledge about $f(x)$ should play a role.

Issues

- **1.** Model the data;
- **2.** Select an ambient space for $f(x)$;
- **3.** Determine hard and soft constraints;
- **4.** Choose appropriate distance measure;
- **5.** Select an objective function to optimize;
- **6.** Choose an appropriate algorithm.

Model the Data

We shall restrict our attention to measurements that are linear, so that the data are linear functional values of $f(x)$; the methods to be discussed here may also be applied to nonlinear problems, such as the **phase problem**, however. To model linear functional data we need an ambient space for $f(x)$.

Select an Ambient Space

Often the function $f(x)$ is viewed as a member of a Hilbert space, so that the measurements can be described in terms of inner products. An important, and often overlooked, issue is the selection of the Hilbert space. Prior knoiwledge can be usefully incorporated through the choice of the ambient space.

Constraints

Clearly, the measurements provide constraints, which may be taken as hard (exact solutions required) or soft (approximate solutions acceptable). We usually know more about $f(x)$, though. We may know that $f(x) \geq 0$, or that $f(x) \in [a, b]$. We may have a good idea of the support of $f(x)$. We may have a prior estimate of $|f(x)|$. We may know that $f(x)$ is smooth, or that it tends to be spiky. If $f(x)$ is an image, we may want sharp edges. We may wish to allow a large dynamic range.

Linear Functional Data

The measured values are linear functionals of $f(x)$, that is, our data are the finitely many inner products

$$d_n = \langle f, h_n \rangle,$$

where, for $n = 1, ..., N$, the $h_n(x)$ are known functions. The inner product is intentionally unspecified.

Distances

If our ambient space is a Hilbert space, the norm for that space provides a distance measure. We may then seek the minimum-norm solution. If we have additional constraints, we may seek to minimize the norm over the constraint sets. This usually involves projection methods. The **projected Landweber** method for solving $Ax = b$ is a good example.

Other Distances

The Hilbert space norm, or 2-norm, may not always be the best choice. The use of the L_1 or 1-norm is another popular choice in image reconstruction. For non-negative x the Kullback-Leibler, or cross-entropy, distance is another useful choice.

Choosing an Objective Function

We may simply want to minimize deviation from the measured data, or to minimize the 2-norm, subject to agreement with the data. If additional constraints are included, we may minimize a proximity function based on orthogonal projections. When the data are noisy, which is really all the time, some regularization may be included.

Choosing the Algorithm

There are general-purpose algorithms, such as Newton-Raphson and its various approximations. There are also special-purpose methods, such as entropy maximization through MART or likelihood maximization using the EM algorithm. For practical use, storage requirements and computational time become important. Finally, since the goal is a reconstruction that serves the practical purpose, mathematical convergence may be of secondary importance.

The Minimum-2-Norm Solution

The minimum-norm solution has the algebraic form

$$\hat{f}(x) = c_1 h_1(x) + ... + c_N h_N(x),$$

where the c_n are chosen to make the reconstruction $\hat{f}(x)$ agree with the data.

Calculating Coefficients

Taking inner products with a fixed $h_m(x)$ on both sides, we get

$$d_m = \langle f, h_m \rangle = \sum_{n=1}^{N} c_n \langle h_n, h_m \rangle.$$

To find the c_n we must solve this N by N system of linear equations, which we write as $d = Hc$.

Ghosts

The true $f(x)$ can be written uniquely as

$$f(x) = \Big(c_1 h_1(x) + ... + c_N h_N(x)\Big) + g(x),$$

where

$$\langle g, h_n \rangle = 0,$$

for $n = 1, ..., N$. Since the $g(x)$ is a *ghost function* whose presence cannot be detected by our sensing system, it would seem that the only way for us to proceed is to accept $\hat{f}(x)$ as our reconstruction and end the discussion.

Ghost Busters

We intentionally left the inner product unspecified because the inner product is not unique; we have the freedom to select the particular inner product we wish to use, and this alters our reconstruction.

Examples

Suppose, initially, that we have data that we can describe as

$$d_n = \int_a^b f(x) g_n(x) dx.$$

Then we can define the inner product of any real functions $u(x)$ and $v(x)$ to be

$$\langle u, v \rangle = \int_a^b u(x) v(x) dx.$$

With this inner product, we have

$$h_n(x) = g_n(x),$$

for each n, and our reconstruction is a linear combination of the functions $g_n(x)$:

$$\hat{f}(x) = c_1 g_1(x) + ... + c_N g_N(x).$$

A New Inner Product

However, for any positive function $p(x)$ on $[a, b]$, we can also write

$$d_n = \int_a^b f(x) g_n(x) p(x) p(x)^{-1} dx.$$

Suppose we define the inner product of any $u(x)$ and $v(x)$ to be

$$\langle u, v \rangle = \int_a^b u(x) v(x) p(x)^{-1} dx.$$

Then, for this inner product, we have

$$h_n(x) = g_n(x) p(x);$$

the PDFT reconstruction takes the form

$$\hat{f}(x) = p(x) \Big(c_1 g_1(x) + ... + c_N g_N(x) \Big).$$

When $p(x)$ is selected as our prior estimate of $|f(x)|$, we incorporate our prior information about $f(x)$, such as its support, into the reconstruction.

Computational Issues

To calculate the coefficients c_n we must first generate the entries of the matrix H, which are now

$$H_{mn} = \langle h_n, h_m \rangle$$

$$= \int_a^b (g_n(x)p(x))(g_m(x)p(x))p(x)^{-1}dx = \int_a^b g_n(x)g_m(x)p(x)dx.$$

This can be a difficult step that we may want to avoid.

The Far Field

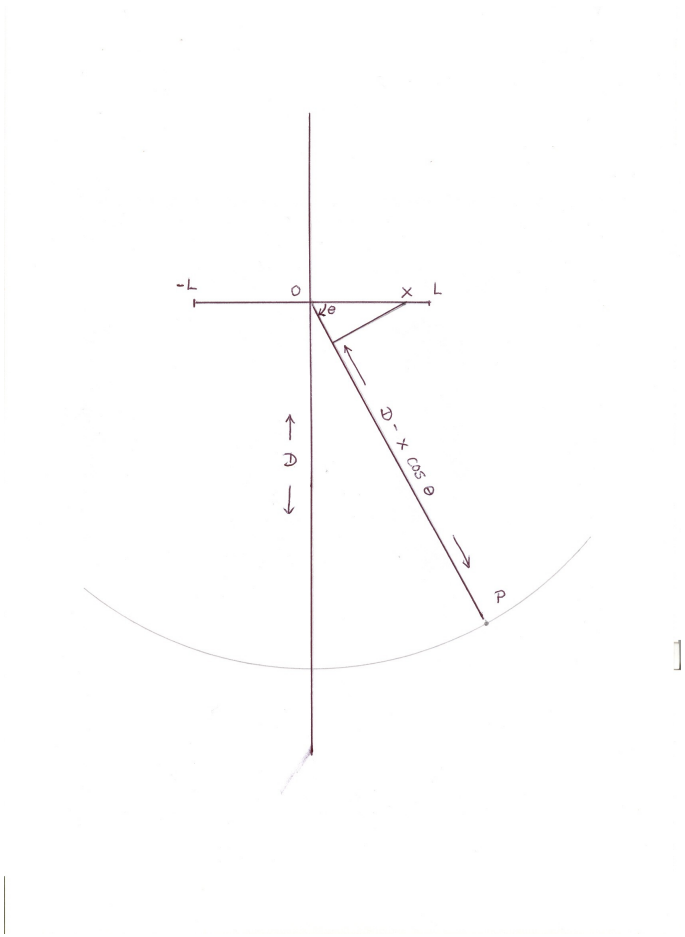

Figure : Farfield Measurements.

The Fourier Transform as Data

Let P lie on a circle of radius $D >> L$, centered at the origin, with the ray from O to P making an angle θ clockwise from the positive x-axis, as shown in the Figure. Each point x in the interval $[-L, L]$ sends out signal

$$f(x)\exp(i\omega t),$$

for known ω and unknown $f(x)$. We want to estimate $f(x)$ from measurements at various points P.

Farfield Measurements

Using the farfield approximation of the distance from x to P, we say that the signal from x to P is delayed by $\frac{1}{c}(D - x\cos(\theta))$, where c is the speed of propagation. Therefore, the signal measured at P at time t is

$$\exp(i\omega(t - D/c)) \int_L^L f(x) \exp(ix\omega\cos(\theta)/c)dx$$

and our measurement at P provides a value of the Fourier transform of $f(x)$.

Fourier Coefficients

The function $f(x)$ has Fourier series

$$f(x) = \sum_{n=-\infty}^{\infty} c_n \exp(-in\pi/L),$$

for

$$c_n = \frac{1}{2L} \int_{-L}^{L} f(x) \exp(in\pi/L) dx.$$

If the angle at P satisfies

$$\cos(\theta) = \frac{n\pi c}{\omega L} = n\frac{\lambda}{2L},$$

where λ is the wavelength, then we have c_n.

Limited Data?

We can get c_n provided that

$$|n| \leq \frac{2L}{\lambda},$$

so there is a limit to how many of the c_n we can measure; the larger the ratio $\frac{2L}{\lambda}$, the more c_n we can get. For small objects we need short wavelengths to obtain good resolution. But clearly, there are other points P at which we can measure the signal. The issue is: What do we do with these other measurements?

Over-Sampled Data

When we take measurements at additional points P on the farfield circle we are said to be **over-sampling**. Complex-function theory tells us that there is information available through measurements within any interval on the farfield circle sufficient to recover $f(x)$ completely; the issue is noise and precision of measurement. Some amount of over-sampling can be effectively used, if we process this data wisely.

Example: Reconstruction from Fourier Transform Values

A basic problem in signal processing is the estimation of the function

$$f(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) e^{-ix\omega} dx \tag{1}$$

from finitely many values of its inverse Fourier transform $F(\omega)$. The discrete Fourier transform (DFT) is one such estimator. As we shall see, there are other estimators that are able to make better use of prior information about $f(x)$ and thereby provide a better estimate.

Choosing the Prior

Suppose the data is $F(n\Delta)$, for $n = 1, ..., N$. Our PDFT reconstruction has the form

$$\hat{f}(x) = p(x) \sum_{n=1}^{N} c_n e^{in\Delta x},$$

with the c_n chosen to make $\hat{f}(x)$ data consistent. If we know $f(x) = 0$, for $|x| > A$, then one choice for $p(x)$ is $\chi_A(x)$, the *characteristic function* that is one for $|x| \leq A$ and zero otherwise.

Over-Sampling

Suppose that $f(x) = 0$ for $|x| > A$, where $0 < A < \pi$. The Nyquist sample spacing is then $\Delta = \pi/A$. In many applications we can take as many samples as we wish, but must take them within some fixed interval. If we take samples at the rate of $\Delta = \pi/A$, we may not get very many samples to work with. Instead, we may sample at a faster rate, say $\Delta = 1$, to get more data points. How we process this over-sampled data is important.

Choosing the Hilbert Space

If we use as our ambient Hilbert space $L^2(-\pi, \pi)$, the minimum-norm reconstruction wastes a lot of effort reconstructing $f(x)$ outside $[-A, A]$, where we already know it to be zero. Instead, we use $L^2(-A, A)$ as the ambient Hilbert space.

The DFT and the MDFT

For the simulation in the figure below, $f(x) = 0$ for $|x| > A = \frac{\pi}{30}$. The top graph is the minimum-norm estimator, with respect to the Hilbert space $L^2(-A, A)$, called the *modified* DFT (MDFT); the bottom graph is the DFT, the minimum-norm estimator with respect to the Hilbert space $L^2(-\pi, \pi)$. The MDFT is a non-iterative variant of Gerchberg-Papoulis band-limited extrapolation.

30 Times Over-Sampled Data

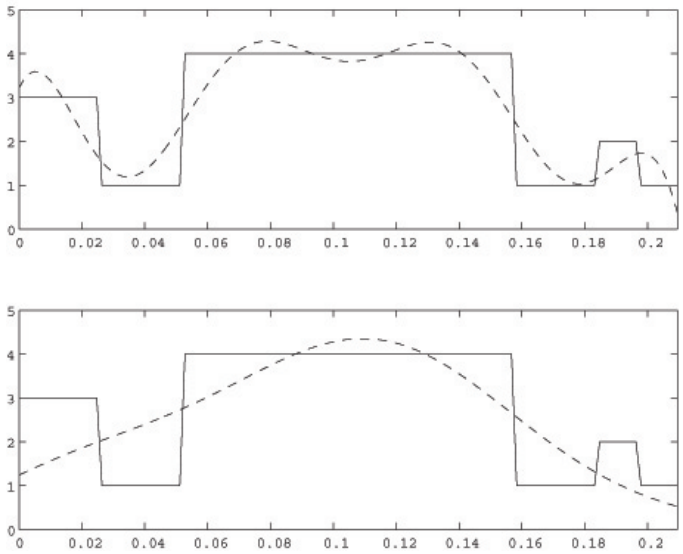

Figure : The non-iterative band-limited extrapolation method (MDFT) (top) and the DFT (bottom); 30 times over-sampled.

Using Other Prior Information

The approach that led to the MDFT estimate suggests that we can introduce other prior information besides the support of $f(x)$. For example, if we have some idea of the overall shape of the function $f(x)$, we could choose $p(x) > 0$ to indicate this shape and use it instead of $\chi_A(x)$ in our estimator. This leads to the PDFT estimator.

Discretizing the Problem

Suppose we select $J > N$ and replace the functions $f(x)$ and $g_n(x)$ with finite (column) vectors,

$$\mathbf{f} = (f_1, ..., f_J)^T,$$

and

$$\mathbf{g}^n = (g_1^n, ..., g_N^n)^T,$$

and model the data as

$$d_n = f_1 g_1^n + ... + f_N g_N^n.$$

Then a vector $\mathbf{f}$ is data consistent if it solves the under-determined system

$$A\mathbf{f} = d,$$

where the entries of the matrix A are

$$A_{n,j} = g_j^n.$$

Minimum-Weighted Norm Solutions

The PDFT estimator minimizes the weighted two-norm

$$\int |f(x)|^2 p(x)^{-1} dx,$$

subject to data consistency. In the discrete formulation of the reconstruction problem, we seek a solution of a system of equations $A\mathbf{f} = d$ for which the weighted two-norm

$$\sum_{j=1}^{J} |f_j|^2 w_j^{-1}$$

is minimized, where $\mathbf{w}$ is a discretization of the function $p(x)$. This can be done using, say, the algebraic reconstruction technique (ART), without forming the matrix H.

Minimum-Two-Norm Solutions

When a system of linear equations $Ax = b$ is under-determined, we can find the solution that minimizes the two-norm,

$$||x||_2^2 = \sum_{j=1}^{J} x_j^2.$$

One drawback is that relatively larger values of x_j are penalized more than smaller ones, leading to somewhat smooth solutions.

Minimum-One-Norm Solutions

If we want a sparse solution of $Ax = b$, we may seek the solution for which the one-norm,

$$||x||_1 = \sum_{j=1}^{J} |x_j|,$$

is minimized. This is important in *compressed sensing* (Donoho; Candès, et al.).

Comparison with the PDFT

If our weights w_j are reasonably close to $|x_j|$, then

$$\sum_{j=1}^{J} |x_j| = \sum_{j=1}^{J} |x_j|^2 |x_j|^{-1} \approx \sum_{j=1}^{J} |x_j|^2 w_j^{-1}.$$

Our goal is not sparsity, but we do wish to reduce the penalty on larger entries.

Sequential Re-weighting

We may obtain a sequence of PDFT solutions, each time using weights suggested by the previous estimate (M. Fiddy and students, 1983). The same idea has recently been applied in *re-weighted one-norm minimization* (Candès, Wakin and Boyd).

Other Applications

- **1.** The non-linear indirect PDFT (IPDFT): extending Burg's nonlinear high-resolution maximum entropy method to include prior information, with application to SONAR signal processing (CB, R. Fitzgerald, M. Fiddy).
- **2.** Phase retrieval: minimizing extrapolated energy as a function of chosen phases, to reconstruct from magnitude-only Fourier data (CB, M. Fiddy).
- **3.** Tomographic imaging: reconstruction from "line" integrals, using a prior estimate of the object (CB, M. Shieh).
- **4.** Mixture problems: estimating combining probabilities from photon frequency counts (CB, B. Levine, J.C. Dainty (1984), CB, D. Haughton, T. Jiang (1993)).

Non-Linear Indirect PDFT

Suppose that $r(x) \geq 0$, for $|x| \leq \pi$, and we want to reconstruct its *additive causal part*,

$$r(x)_+ = \sum_{n=0}^{\infty} R(n) e^{inx},$$

from data $R(n)$, for $n = 0, 1, ..., N$. We use the prior $p(x)$ and the PDFT, obtaining the estimate

$$\hat{r}(x) = p(x) \sum_{n=0}^{N} c_n e^{inx}.$$

Obtaining the Coefficients

To obtain the c_n we need to solve the system

$$\begin{bmatrix} P(0) & P(-1) & \dots & P(-N) \\ P(1) & P(0) & \dots & P(-N+1) \\ \cdot & \cdot & \dots & \cdot \\ \cdot & \cdot & \dots & \cdot \\ P(N) & P(N-1) & \dots & P(0) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \cdot \\ \cdot \\ c_N \end{bmatrix} = \begin{bmatrix} R(0) \\ R(1) \\ \cdot \\ \cdot \\ R(N) \end{bmatrix}.$$

Suppose now that we switch the roles of $r(x)$ and $p(x)$, "estimating" $p(x)_+$ using $r(x) \geq 0$ as the prior.

Switching Roles

Now we need to solve the system

$$\begin{bmatrix} R(0) & R(-1) & \dots & R(-N) \\ R(1) & R(0) & \dots & R(-N+1) \\ . & . & \dots & . \\ . & . & \dots & . \\ R(N) & R(N-1) & \dots & R(0) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ . \\ . \\ c_N \end{bmatrix} = \begin{bmatrix} P(0) \\ P(1) \\ . \\ . \\ P(N) \end{bmatrix}.$$

Since $R(-n) = \overline{R(n)}$, we know all the entries of the matrix.

The "Estimate" of $p(x)_+$

Our "estimate"of $p(x)_+$ is then

$$\hat{p}(x)_+ = r(x) \sum_{n=0}^{N} c_n e^{inx} = r(x)c(x).$$

The additive causal part of the right side is

$$\Big(r(x)c(x)\Big)_+ = r(x)_+ c(x) + \sum_{m=0}^{N-1} \Big(\sum_{k=1}^{N-m} R(-k)c_{m+k} \Big) e^{imx}$$

$$= r(x)_+ c(x) + j(x).$$

The IPDFT

From

$$\hat{p}(x)_+ \approx r(x)_+ c(x) + j(x),$$

we get

$$r(x)_+ \approx q(x) = \frac{p(x)_+ - j(x)}{c(x)}.$$

Our IPDFT estimate of $r(x)$ is then

$$\hat{r}(x) = 2\text{Real}(q(x)) - R(0).$$

The IPDFT is real-valued. If $c(x)^{-1}$ is causal, that is,

$$c(x)^{-1} = d_0 + d_1 e^{ix} + d_2 e^{2ix} + ...,$$

then our estimate $q(x)$ of $r(x)_+$ is causal and the IPDFT is consistent with the data. It is not guaranteed to be non-negative, but seems to be, most of the time. When $p(x) = 1$ for all x we get Burg's maximum entropy estimator.
Open Problem: When is $c(x)^{-1}$ causal?

Poisson Mixture Problems

A *compound Poisson* probability function on the non-negative integers has

$$p(n) = \frac{1}{n!} \int_0^\infty c(\lambda) e^{-\lambda} \lambda^n d\lambda,$$

as the probability that the non-negative integer n will occur; here the non-negative function $c(\lambda)$ is the *compounding probability density function*. Measured counts provide estimates of $p(n)$, for $n = 0, 1, ..., N$. On the basis of this data we want to estimate the function $c(\lambda)$. Both the PDFT and IPDFT approaches can be used for this purpose.

The End

THE END