Numerical Astronomy 1 – Part 3
Classical inversions and instability

Norman Gray

8 November 1998*

The simplest inverse problem is $g(x) = \int^x u(y)\,\mathrm{d}y$, with solution $u(y) = \mathrm{d}g/\mathrm{d}x|_{x=y}$. More generally, the kernel may be such that we can analytically obtain the source, or underlying, function $u(y) = \mathcal{K}^{-1}\{g(x); y\}$. This is a fine thing to manage, but the problem is not finished. Even if you start with a known, analytic, $g(x)$ (as in the 'control problem', for example), and can find an analytic $\mathcal{K}^{-1}$, you might still be unable to stably produce the numbers you require (this is straying into the territory of the evaluation of functions – see *Numerical Recipes* for further discussion). We have already seen that differentiation can be unstable. Thus, even before we start to consider the problem of dealing with measurement errors, we might have to analyse our problem as an inverse problem.

First, examine the first-kind Fredholm problem

$$\int_a^b K(x,y)u(y)\,\mathrm{d}y = g(x).$$

We will be given the data function $g(x)$ at a set of points $x_i, i = 1, \ldots, m$, so that

$$\int_a^b K_i(y)u(y)\,\mathrm{d}y = g_i, \tag{3.1}$$

where $g_i \equiv g(x_i)$ and $K_i(y) \equiv K(x_i, y)$.

Note that there is no *error* here, in the sense that we haven't lost any of the available information by making any approximation; however, some *information has been lost* by the fact of discretisation. A real function has an infinite number of degrees of freedom, but only m are constrained here. There is precisely one polynomial of degree $(m-1)$ (that is, one with m degrees of freedom) which will pass through all of the g_i, but infinitely many of degrees $\geq m$ (eg, for $m = 2$ there is one straight line which passes through two points, but infinitely many curves). Now, most of the solutions are unreasonable, and the methods below use that fact implicitly (in basic quadrature, for example, by assuming that $f(x)$ is a reasonable estimate for the functions value in the range $(x-h, x+h)$), but the distinctive feature of the classical methods of IP solution is that they do not use such *a priori* information explicitly.

Non-classical methods explicitly use prior information about a solution, such as assumptions of smoothness or positivity, to help constrain a solution, and we will deal with these in the next part. The classical methods, however, are more directly related to the analytic problem, and as such are more directly intelligible, and more clearly demonstrate the nature of the problem.

* Modified 1 December, to clarify example in section 2.3.

1 Classical solutions

1.1 Quadrature

Approximate the integral Eqn. (3.1) by

$$g_i = \int_a^b K_i(y)u(y)\,\mathrm{d}y \approx \sum_{j=1}^n K_i(y_j)w_j u(y_j) = \sum K_{ij}u_j, \tag{3.2}$$

defining $K_{ij} = K_i(y_j)w_j$. The w_i are weights obtained from some quadrature technique such as the trapezoidal rule or Simpson's rule.

For $n \ll m$, the recovery of the values u_j is *overdetermined*, in that we want n numbers u_j, with $m > n$ constraints. This will require a *best-fit* approximation to the solution, which will typically be some variation of least-squares.

As we increase n, the approximation Eqn. (3.2) improves. The operator $\mathcal{K}$ will typically be more-or-less singular (if it weren't, we wouldn't be bothering with the apparatus of IP), so that as $n \lesssim m$, K_{ij}^{-1} more closely approximates the unboundedness of $\mathcal{K}^{-1}$. We can attempt to limit the damage by choosing suitable mesh of points g_i and y_j, but this ameliorates the problem rather than properly dealing with it.

1.2 Product integration

Rewrite Eqn. (3.1) as

$$g(x_i) = \sum_{j=1}^n \int_{y_{j-1}}^{y_j} K(x_i,y)u(y)\,\mathrm{d}y \qquad ,i = 1,\ldots,m \tag{3.3}$$

setting $y_0 = a$ and $y_n = b$. Now we can approximate the underlying function $u(x)$ by $\bar{u}_j(x)$ in $[y_{j-1}, y_j]$. The functions $\{\bar{u}_j\}$ could be constant, or linear, or whatever was most appropriate. Taking, for example, $\bar{u}_j$ to be constant, with the value of $u(y)$ at the midpoint of the interval, we can write

$$K_{ij} \equiv \int_{y_{j-1}}^{y_j} K(x_i,y)\,\mathrm{d}y \tag{3.4}$$

so that

$$g(x_i) = \sum_{j=1}^n K_{ij}\bar{u}_j. \tag{3.5}$$

How is this different from Eqn. (3.2)? The principal difference is that by making the approximation for $u(x)$ we have sneaked in a smoothness constraint, and this implicit prior information is enough to improve the stability properties of the inversion.

1.3 Polynomial expansion

Another route is to expand the source function $u(x)$ in terms of a complete set of functions $\{\phi_i\}$, which are orthonormal with respect to an inner product

$$\langle\phi_i,\phi_j\rangle \equiv \int_a^b \phi_i(y)\phi_j(y)\,\mathrm{d}y = \delta_{ij}. \tag{3.6}$$

Expand the source function as

$$u(x) = \sum_{i=0}^\infty u_i\phi_i(x). \tag{3.7}$$

Taking the inner product of this with ϕ_j we obtain

$$u(x) = \sum_{i=0}^{\infty} \langle u, \phi_i \rangle \phi_i(x), \tag{3.8}$$

and if we drop this into the discretised problem Eqn. (3.1), and truncate the resulting series, we obtain

$$g(x_i) = \sum_{j=0}^{m} \langle u, \phi_i \rangle K_{ij}, \qquad K_{ij} = \int_a^b K(x_i, y) \phi_j(y) \, \mathrm{d}y. \tag{3.9}$$

This can be inverted to obtain the coefficients $u_i = \langle u, \phi_i \rangle$, which can reconstruct the source function from Eqn. (3.7).

This is potentially a very powerful method, but it does rely on the legitimacy of the truncation of the series at m terms. That means that it depends on the choice of the set $\{\phi_i\}$: you can get very good results if this is chosen suitably, but also spectacularly bad (but nonetheless unfortunately plausible) results if it is chosen badly.

1.4 Singular value decomposition – SVD

The discretised problem Eqn. (3.1) can be written, as in Eqn. (3.2), as

$$\mathbf{g} = \mathsf{K}\mathbf{u}, \tag{3.10}$$

where K is $m \times n$ ($m \geq n$). It follows that, formally,

$$\mathbf{u} = [\mathsf{K}^T \mathsf{K}]^{-1} \mathsf{K}^T \mathbf{g}, \tag{3.11}$$

where $A \equiv \mathsf{K}^T \mathsf{K}$ is square $n \times n$ and non-singular by assumption, so that the inverse exists. This reduces to $\mathbf{u} = \mathsf{K}^{-1}\mathbf{g}$ if K is square.

It is a theorem of linear algebra that, for a matrix A ($m \times n$), there exist matrices U ($m \times m$, orthogonal), V ($n \times n$, orthogonal) and $\mathsf{\Sigma}$ ($m \times n$, diagonal), such that

$$\mathsf{A} = \mathsf{U}\mathsf{\Sigma}\mathsf{V}^T. \tag{3.12}$$

If A is square $n \times n$, as in our case, then so are U, V, and $\mathsf{\Sigma}$, and $\mathsf{\Sigma} = \mathrm{diag}(\sigma_i)$. The $\{\sigma_i\}$ are known as the singular values, and Eqn. (3.12) is the *Singular Value Decomposition*. We can arrange that $\sigma_1 \geq \sigma_2 \geq \ldots > 0$ (all will be strictly positive if A is non-singular).

The advantage of the SVD is that A in Eqn. (3.12) is very easy to invert:

$$\mathsf{A}^{-1} = (\mathsf{U}\mathsf{\Sigma}\mathsf{V}^T)^{-1} = \mathsf{V}\mathsf{\Sigma}^{-1}\mathsf{U}^T, \tag{3.13}$$

and $\Sigma^{-1} = \mathrm{diag}(1/\sigma_i)$. We thus have the formal solution

$$\mathbf{u} = \mathsf{A}^{-1}\mathsf{K}^T\mathbf{g}, \tag{3.14}$$

but this is still ill-conditioned, since $\mathsf{\Sigma}^{-1}$ blows up when σ_i is too small, which has the effect of destructively amplifying random noise in the data vector $\mathbf{g}$. The extent of this ill-conditioning is given by the *condition number* $C_K = \sigma_1/\sigma_n$. We improve the conditioning by selecting a maximum condition number c, and using in Eqn. (3.14) not A but A_1, which is A with all the terms $1/\sigma_i$ larger than a threshold c/σ_1 suppressed, by setting them to zero.

The selection of the maximum condition number is potentially quite delicate, and might depend, for example, on the numerical accuracy available or required in your numerical calculation. Inversion using SVD is computationally expensive, but stable and robust.

2 The instability

In matrix form, the inverse problem is

$$\mathsf{K}\mathbf{u} = \mathbf{g}. \tag{3.15}$$

There will always be measurement errors in the data vector $\mathbf{g}$, so that what we measure is not $\mathbf{g}$ but $\mathbf{g} + \delta\mathbf{g}$. These induce errors in the recovered source function, and we recover $\mathbf{u} + \delta\mathbf{u}$. Since K is linear, we have

$$\delta\mathbf{u} = \mathsf{K}^{-1}\delta\mathbf{g}, \tag{3.16}$$

and we want to find the relationship between the 'size' of $\delta\mathbf{g}$ and $\delta\mathbf{u}$. Such 'sizes' we can discuss using vector and matrix norms.

2.1 Norms of vectors and matrices

Norms are real-valued functionals of vectors and matrices, with the properties

$$\left.\begin{array}{rl} ||\mathbf{v}|| \geq 0, & \text{with } ||\mathbf{v}|| = 0 \iff \mathbf{v} = \mathbf{0} \\ ||\alpha\mathbf{v}|| = |\alpha|||\mathbf{v}|| & \\ ||\mathbf{v} + \mathbf{w}|| \leq ||\mathbf{v}|| + ||\mathbf{w}|| & \end{array}\right\} \tag{3.17}$$

There are several possible definitions of the vector norm, of which one of the most useful is the 'p-norm':

$$||\mathbf{v}||_p = \left[\sum_{j=1}^{n} |v_i|^p\right]^{1/p} \tag{3.18}$$

from which

$$||\mathbf{v}||_\infty = \max_i |v_i|. \tag{3.19}$$

Matrix norms satisfy the conditions

$$\left.\begin{array}{ll} ||\mathsf{AB}|| \leq ||\mathsf{A}||||\mathsf{B}||, & \mathsf{A} \text{ and } \mathsf{B} \text{ square} \\ ||\mathsf{A}\mathbf{v}|| \leq ||\mathsf{A}||||\mathbf{v}|| & \end{array}\right\} \tag{3.20}$$

We can define p-norms in such a way that

$$||\mathsf{A}||_2 = (\max\lambda : \mathsf{A}^T\mathsf{A}\mathbf{v} = \lambda\mathbf{v})^{1/2} = \sigma_{\max}, \tag{3.21}$$

where $\sigma_{\max}$ is the maximum singular value of A (cf, Sect. 1.4), and

$$||\mathsf{A}||_\infty = \max_i \sum_{j=1}^{n} |A_{ij}|. \tag{3.22}$$

2.2 Instability in IP inversion

From Eqns. (3.15), (3.16) and (3.20), we have

$$\begin{array}{rcl} ||\mathbf{g}|| & \leq & ||\mathsf{K}||||\mathbf{u}|| \\ ||\delta\mathbf{u}|| & \leq & ||\mathsf{K}^{-1}||||\delta\mathbf{g}||. \end{array}$$

Thus, immediately

$$\frac{||\delta\mathbf{u}||}{||\mathbf{u}||} \leq C_K \frac{||\delta\mathbf{g}||}{||\mathbf{g}||}, \tag{3.23}$$

with the *condition number* defined as $C_K \equiv ||\mathsf{K}||||\mathsf{K}^{-1}||$.

Recall that $||\mathsf{K}||_2 = \sigma_{\max}$. It follows that $||\mathsf{K}^{-1}||_2 = 1/\sigma_{\min}$, so that

$$C_K = ||\mathsf{K}||_2||\mathsf{K}^{-1}||_2 = \frac{\sigma_{\max}}{\sigma_{\min}} \tag{3.24}$$

(compare Sect. 1.4 above). When the condition number is large, Eqn. (3.23) tells us that the matrix K permits the recovered source vector $\mathbf{u}$ to have errors much larger than those in the data vector $\mathbf{g}$.

2.3 Examples

In Sect. 2-3, we discussed recovering a source function by differentiating the data function Eqn. (2.12). The condition number of the inversion Eqn. (2.14) depends on the matrix K, which has norm $\|\mathsf{K}\|_\infty = n \sim 1/h$. Thus $C_K = \|\mathsf{K}\|_\infty \|\mathsf{K}^{-1}\|_\infty = n^2$. Thus, although the quadrature Eqn. (2.13) becomes more accurate as n increases, the inversion becomes less stable, and beyond a point, we lose accuracy instead.

Secondly, consider

$$\mathsf{E} = \begin{pmatrix} 1 & 1 \\ 1 & 1+\epsilon \end{pmatrix}. \tag{3.25}$$

This is easily invertible, but its condition number $C_E \sim 1/\epsilon$ as $\epsilon \to 0$. This echoes the observation that, as $\epsilon \to 0$, the matrix rows become degenerate and K becomes singular. For all $(x, y) \in \mathrm{domain}(\mathsf{E}) = \mathbb{R}^2$, the set of points $\{(x, y) : x + y = c\}$ ismapped into the single point (c, c). That is, $Z_E = \{(x, y) : x = -y\}$. When $\epsilon \neq 0$, $\mathrm{range}\,\mathsf{E} = \mathbb{R}^2$ but, following Eqn. (3.23), a tiny change in position in $\mathrm{range}(\mathsf{E})$ is consistent wth a huge movement in $\mathrm{domain}(\mathsf{E})$.

3 Example: Inversion of Abel's equation

We will examine the product integration (Sect. 1.2) and polynomial expansion (Sect. 1.3) methods as applied to the problem

$$\int_0^x \frac{u(y)}{(x-y)^\alpha} = g(x), \tag{3.26}$$

which is Abel's equation. This follows Craig and Brown, section 5.4.

If we expand

$$u(y) = \sum_{j=1}^{\infty} u_j y^{j-1}, \tag{3.27}$$

then we swiftly find, using Eqn. (3.9), that

$$g(x_i) = \sum_{j=1}^{\infty} x_i^{j-\alpha} \beta(j, \alpha+1) u_j, i = 1, \ldots, n \tag{3.28}$$

The sum can be truncated at a suitable point; truncation at $m = n$ gives a lower-triangular system which is easily solved for u_j. This solution is analytically attractive, but it becomes unstable rapidly as n increases.

Alternatively, we can use the product integration method to expand Eqn. (3.26) as

$$g(x_i) = \sum_{j=1}^{n} \overline{u}_j \int_{x_{j-1}}^{x_j} (x_i - y)^{-\alpha}\, \mathrm{d}y, \qquad i = 1, \ldots, m. \tag{3.29}$$

We can recover the values $\overline{u}_j$ by a process similar to above. This is in principle less accurate, as it implicitly makes the potentially crude assumption that the source function value may be taken to be constant in each interval $[x_{j-1}, x_j]$, whilst the polynomial expansion method will be exact for polynomial $u(x)$ up to degree $(n-1)$. However, the payoff for this lower accuracy is substantially improved stability.

Examples

Section 1.4

Consider the problem $g = \mathsf{K}u$, where

$$\mathsf{K} = \frac{1}{2} \begin{pmatrix} 2 & 0 & 0 \\ 0 & 1+\epsilon & 1-\epsilon \\ 0 & 1-\epsilon & 1+\epsilon \end{pmatrix}. \tag{3.30}$$

Confirm that the SVD decomposition of this is $\mathsf{K} = \mathsf{U}\Sigma\mathsf{U}^T$, where

$$\mathsf{U} = \frac{1}{\sqrt{2}} \begin{pmatrix} \sqrt{2} & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & -1 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & \epsilon \end{pmatrix}, \tag{3.31}$$

showing that the problem K has singular values $(1, 1, \epsilon)$. Satisfy yourself that $\mathsf{K}^{-1} = \mathsf{U}\Sigma^{-1}\mathsf{U}^T$ is indeed the inverse of K. If the source vector is $\mathbf{u} = (1, 1, 1)$, then show that $\mathbf{g} = (1, 1, 1)$. If, however, we were to obtain the data $\hat{\mathbf{g}} = (1 + \delta_1, 1 + \delta_2, 1)$, with δ_1 and δ_2 both small, like ϵ, then obtain $\hat{\mathbf{u}} = \mathsf{K}^{-1}\hat{\mathbf{g}}$ (there is no significance, by the way, in my having omitted any error on u_3 – this is simply to make the calculation fit on one page!). Note that if ϵ is small, then the solution $\hat{\mathbf{u}}$ is dominated by the noise δ_2 (but the noise term δ_1, not near the range of the null space of $\mathcal{K}$, does not have a big effect).

Now remove the troublesome singular values by removing the term $1/\epsilon$ in Σ^{-1}, to obtain $\Sigma^{-1\prime} = \mathrm{diag}(1, 1, 0)$, and thus obtain $\mathsf{K}^{-1\prime} = \mathsf{U}\Sigma^{-1\prime}\mathsf{U}^T$. Obtain $\hat{\mathbf{u}}$ using this new $\mathsf{K}^{-1\prime}$, and notice that the small noise term δ_2 produces only a small effect on the recovered solution.

Section 2.3

Consider the problem $\mathbf{g} = \mathsf{E}\mathbf{u}$. Show directly that the inverse of E is

$$\mathsf{E}^{-1} = \begin{pmatrix} 1 + 1/\epsilon & -1/\epsilon \\ -1/\epsilon & 1/\epsilon \end{pmatrix}. \tag{3.32}$$

Show that an source vector of $\mathbf{u} = (1/2, 1/2)$ produces a data vector $\mathbf{g} = (1, 1+\epsilon/2)$, and that a data error $\delta\mathbf{g} = (0, \delta/2)$ produces an error in the recovery of $\delta\mathbf{u} = (-\delta/2\epsilon, \delta/2\epsilon)$. Find the condition number of the problem E (using the ∞-norm in Eqn. (3.22)), and show that this is consistent with the ratios $||\delta u||_\infty/||u||_\infty$ and $||\delta g||_\infty/||g||_\infty$. Finally, pick values such as $\epsilon = 0.1$ (ie, a condition number of forty-ish – a rather well-conditioned problem, really!) and $\delta = 0.1$ (ie, data errors of 5%), evaluate $\mathbf{g} + \delta\mathbf{g}$ and the $\mathbf{u} + \delta\mathbf{u}$ it gives rise to, and note the respective differences between these and $\mathbf{g}$ and $\mathbf{u}$.