Numerical Astronomy 1 – Part 4
Non-classical inversion

Norman Gray

27 November 1998

1 Overview

In previous parts of this course, we have implicitly used assumptions about the underlying functions we are trying to recover (for example, that they are smooth). These have been vital to the methods we have examined, since (as discussed at the beginning of part 3) they help limit the infinite number of solutions which are in principle compatible with the finite number of data points we have. The problem is that, because these assumptions are implicit, they enter our deliberations uncontrollably. That is, quadrature and the product integration method work well if the underlying function really is smooth, and polynomial expansion works if the underlying function is well-fitted by the expansion. You *can* play games and fiddle with the details of either of these methods, but if there's a real mismatch between your assumptions about the underlying function and the reality, then you're on a hiding to nothing.

SVD doesn't have this problem – it really is model-free, but it does require some insight into the problem when making the decision about where to cut off the singular values. There are very sophisticated guides to this decision, and choices which are optimal in various technical senses, but these are based on information brought to the problem from outside.

Non-classical methods explicitly add in other information, by allowing you to specify quite general constraints. You both choose the constraint and can control how strong to make it.

Non-classical methods boil down to a minimisation of some 'goodness of fit' measure, in a *very* general sense, subject to a bound on an equally general 'reasonableness' constraint. In the method of regularisation, we minimise (the norm of) the *residual* $\mathcal{K}\hat{u} - \hat{g}$ subject to a quadratic functional of $\hat{u}$ having a bound (here $\hat{u}$ represents a particular recovery of the underlying function, depending on a particular realisation $\hat{g}$ of the data – this is in contrast to u, which is the unknowable 'real' underlying function); in Backus-Gilbert, we simultaneously maximise the resolution and stability of an estimate of the recovered function; and in Maximum Entropy, we minimise the residual, subject to a bound on a particular non-linear functional of $\hat{u}$.

At one level, there are very strong connections between the classical and non-classical approaches. SVD, for example, is very strongly linked to zeroth order regularisation (where the prior assumption is that the underlying function is zero unless the data demands otherwise), but crucially the latter non-classical approach comes from a different point of view.

Without becoming overly philosophical about it, non-classical approaches spring from, and SVD fits into, an approach which doesn't attempt to recover the solution which is 'really' there, but instead attempts to find (a set of properties of) a solution which is *consistent with the data*, acknowledging that the fact that a residual is small does not imply, and should not be taken as a proxy for, the claim that $||u - \hat{u}||$ is small. The Backus-Gilbert method makes this explicit, to some extent, by discussing functionals of the solution, and the resolution and stability with which these can or cannot be recovered. To some extent, the process of finding the mean of a set of data can be regarded as a very unambitious inverse problem, obtaining a potentially very high-quality estimate of a property of the underlying function. Once you've found the mean, you can go on to find the variance and higher moments, and worry about whether they provide useful information about your function.

2 Regularisation

One way of stating the classical approach to the solution is to decide that a solution is acceptable provided $||\mathcal{K}\hat{u} - \hat{g}|| \leq ||\delta g||$, where δg is some measure of the error between g and $\hat{g}$. This might obtain the gross properties of the solution, but fail to recover the high-frequency properties with any plausibility, producing spurious, wildly-oscillating, solutions *which fit the data* as closely as you might wish.

We can remove such high-frequency artefacts by forming an appropriate linear functional of the underlying function, $\mathcal{H}u$, and finding the u which minimises the residual $||\mathcal{K}u - \hat{g}||$ subject to the constraint $||\mathcal{H}u||$ having some particular value or, equivalently, minimising $||\mathcal{H}u||$ subject to a bound on the value of the residual. Using lagrange multipliers, this turns into the prescription that we take as our recovery the u which minimises

$$||\mathcal{K}u - \hat{g}||^2 + \lambda||\mathcal{H}u||^2, \tag{4.1}$$

with λ termed the regularisation parameter. The choice of the stabilising operator $\mathcal{H}$ is where we feed in our prior suppositions about the nature of the underlying function. The simplest choice for $\mathcal{H}$ is the identity operator, so that the prescription in Eqn. (4.1) minimises the size of u and so in effect presumes that the solution will be approximately zero – this is zeroth order regularisation. First order regularisation presumes that the solution will be approximately constant (minimising the derivative), and second order regularisation (the original, and still common) minimises the norm of the second derivative:

$$||\mathcal{H}u||_2^2 = ||u''||_2^2 = \int_a^b [u''(y)]^2 \, \mathrm{d}y, \tag{4.2}$$

subject to the classical $||\mathcal{K}\hat{u} - \hat{g}|| \approx ||\delta g||$. The constraint need not be an integral one, but might weight the solution towards some prior estimate.

Upon discretisation of the problem, we again find ourselves solving a matrix equation, this time finding the solution of the equation

$$(\mathsf{K}^T\mathsf{K} + \lambda\mathsf{H})\mathbf{u} = \mathsf{K}^T\hat{\mathbf{g}}, \tag{4.3}$$

where H is a smoothing matrix depending on the functional $\mathcal{H}$. When $\lambda = 0$ we recover the classical solution, and as λ increases we force the solution more towards our prior estimate. There is no mechanical prescription for choosing the value of the smoothing parameter λ; there are choices which are optimal in technical senses, but few robust improvements on the brute-force method of choosing the value of λ which best recovers simulated data.

3 Other non-classical techniques

Variants of regularisation dominate, but do not exhaust, the range of non-classical inversion techniques. The Backus-Gilbert method, and the method of Maximum Entropy have the same general approach, in that both aim to minimise some measure of goodness-of-fit subject to some criterion on the acceptability of the recovery.

3.1 Backus-Gilbert

The Backus-Gilbert method[1] concentrates, not on finding estimates for the underlying function, as such, but instead on finding *integrals* of that function which can be recovered with confidence. Because of its different approach it allows us a qualitative understanding of, and thus an explicit quantitative control over, the compromise between bias and stability in our inversion.

The underlying function $u(r)$ is related to the data g_i though

$$g_i = \int K_i(y) u(y)\,\mathrm{d}y + n_i, \tag{4.4}$$

where n_i is a random admixture of noise. For the Backus-Gilbert method, we suppose that the underlying function u and our estimate $\hat{u}$ of it are related by an *averaging kernel* $\hat{\delta}(r, r')$, through

$$\hat{u}(y) = \int \hat{\delta}(y, y') u(y')\,\mathrm{d}y'\,. \tag{4.5}$$

Since we do not know the underlying function, the averaging kernel is of no use to us directly; however we can study its properties, and use our data g_i in such a way as to optimise those properties, and so optimise the dependence of the estimate $\hat{u}(y)$ on the underlying function and the noise. Specifically, we will aim to minimise the width of $\hat{\delta}(y, y')$, and so maximise its resolution, subject to the conflicting demand that the averaging kernel be wide enough that the estimate is not unduly sensitive to noise. We seek a set of *response kernels* $q_i(y)$, which produce an estimate of the underlying function through

$$\hat{u}(y) = \sum_i q_i(y) g_i. \tag{4.6}$$

By substituting Eqn. (4.4) into Eqn. (4.6) and comparing with Eqn. (4.5), we obtain an expression for $\hat{\delta}(y, y')$ in terms of $q_i(y)$ and $K_i(y)$. Using this, we can form some measure of the *width* of $\hat{\delta}(y, y')$ such as

$$\mathcal{A} = \int (y - y')^2 [\hat{\delta}(y, y')]^2\,\mathrm{d}y', \tag{4.7}$$

which depends on q_i and K_i; and we can form a measure of the stability of Eqn. (4.6) such as $\mathcal{B} = \mathrm{Var}\,\hat{u}(y)$, which depends on q_i and the covariance matrix of the noise n_i.

The Backus-Gilbert method consists of finding those $q_i(y)$ which minimise

$$\mathcal{A} + \lambda\mathcal{B} = \int (y - y')^2 [\hat{\delta}(y, y')]^2\,\mathrm{d}y' + \lambda\,\mathrm{Var}\,\hat{u}(y), \tag{4.8}$$

for some selected parameter λ. The nature of the trade-off is clear: in order to improve the stability of the recovery, we choose response kernels q_i which make $\hat{\delta}(y, y')$ broader, and so extend the weighted average over a greater number of the data points g_i. The cost of this is that the estimate of the recovered point will be biased by the inclusion of the extra data, and this will be more marked when the underlying function is rapidly varying. The minimisation problem Eqn. (4.8) has an explicit analytic solution for $\mathbf{q}_\lambda(y)$ in terms of the parameter λ, the noise covariance matrix, and integrals of the K_i, and these different solutions, when combined with the data g_i using Eqn. (4.6), give different reconstructions $\hat{u}_\lambda(y)$.

The Backus-Gilbert method is expensive, because each minimisation recovers $u(y)$ at only a single point y. It also suffers from bias, because of the spread of the average in Eqn. (4.6). Both of these problems can be addressed by sophisticated variants of the method, but the method is rarely used for actual data recovery. The approach is most valuable as a theoretical tool – it is supremely valuable for *exploring the problem*, and allowing you to make statements about which features of the underlying function are and are not reliably recoverable.

[1] G E Backus and F Gilbert, *Geophysical Journal of the Royal Astronomical Society*, **16**, 169–205 (1968) and G E Backus and F Gilbert, *Philosophical Journal of the Royal Society of London*, **A266**, 123–192 (1970).

3.2 Maximum entropy

The Maximum Entropy Method (ME) is simply stated: it is regularisation with the regularising functional $\mathcal{H}\mathbf{u} = \sum u_i \ln u_i$ (with a slight abuse of notation: here $\{u_i\}$ is the set of recovered points), and the constraint that $||(\mathcal{K}\hat{\mathbf{u}} - \hat{\mathbf{g}})/\delta g||_2 = N$, which is the constraint that $\chi^2(\hat{\mathbf{u}})$ be equal to its statistical expectation N, the number of data points.

This choice of regularisation has several consequences. One is that $\mathcal{H}\mathbf{u}$ diverges as any u_i goes to zero, so that this functional implicitly imposes positivity on the solution. Secondly, because $\mathcal{H}\mathbf{u}$ is non-linear, it is much harder to solve, and the solution must be obtained iteratively. Thirdly, the functional is by itself maximised when $\mathbf{u}$ is flat, so that this is the prior information expressed in this choice.

ME is most often used in image processing (when the u_i refer to pixels). It has the good features that it very effectively removes high-frequency noise, whilst enhancing the resolution of features within the image ('superresolution'). The cost of this is that it suffers significantly from bias, and that it can be difficult to obtain estimates of the uncertainties in, for example, recovered intensity.

3.3 Bayes theorem

Bayes' Theorem, stated for an inverse problem, is

$$\mathrm{Prob}(\mathbf{u}|\mathbf{g}) = \mathrm{Prob}(\mathbf{g}|\mathbf{u}) \frac{\mathrm{Prob}(\mathbf{u})}{\mathrm{Prob}(\mathbf{g})}, \tag{4.9}$$

and a Bayesian approach to inverse problems consists of finding that u which maximises this *posterior probability*, given some estimate for the *prior probability* $\mathrm{Prob}(\mathbf{u})$.

This is an illuminating approach to inverse problems in general, but it fits particularly naturally into a discussion of ME. If you have a certain number of photons which you know arrived at your CCD, then there is a large number of possible arrangements of those photons on the CCD consistent with that, just as there is a large number of arrangements of atoms in configuration space, consistent with a box of atoms having a certain temperature. This can be used to form a measure of the 'entropy' of your image $\mathbf{u}$, and thus swiftly to a measure of that image's prior probability as $\mathrm{Prob}(\mathbf{u}) \propto \sum u_i \ln u_i$.

This argument has been used to suggest that ME is the only 'consistent' approach to inverse problems. The claim is interesting and useful to examine, but probably overstated.

Examples

Section 2

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

$$\mathsf{E} = \begin{pmatrix} 1 & 1-\epsilon \\ 1 & 1+\epsilon \end{pmatrix}$$

Using

$$\mathsf{M} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \Rightarrow \mathsf{M}^{-1} = \frac{1}{|\mathsf{M}|}\begin{pmatrix} d & -b \\ -c & a \end{pmatrix},$$

write down the inverse matrix E^{-1}, and hence the $\mathbf{u}$ which results from the data vectors $\mathbf{g} = (1,1)^T$ and $\hat{\mathbf{g}} = (1, 1+\delta)^T$. The latter is the vector $\mathbf{g}$ with noise. Note that the recovery is unstable if ϵ is small. You should obtain

$$\begin{aligned} \mathbf{u} &= \mathsf{E}^{-1}\mathbf{g} = \frac{1}{2\epsilon}\begin{pmatrix} 1+\epsilon & \epsilon-1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \\ \hat{\mathbf{u}} &= \mathsf{E}^{-1}\hat{\mathbf{g}} = \frac{\delta}{2}\begin{pmatrix} 1 + 2/\delta - 1/\epsilon \\ 1/\epsilon \end{pmatrix}. \end{aligned}$$

Now consider using zeroth-order regularisation to recover the vector $\hat{\mathbf{u}}$. That is, from Eqn. (4.3)

$$\hat{\mathbf{u}} = (\mathsf{E}^T\mathsf{E} + \lambda\mathcal{I})^{-1}\mathsf{E}^T\hat{\mathbf{g}}.$$

Show that

$$\mathsf{A}_\lambda \equiv \mathsf{E}^T\mathsf{E} + \lambda\mathcal{I} = \begin{pmatrix} 2+\lambda & 2 \\ 2 & 2(1+\epsilon^2)+\lambda \end{pmatrix},$$

which has inverse

$$\mathsf{A}_\lambda^{-1} = \frac{1}{|\mathsf{A}_\lambda|}\begin{pmatrix} 2(1+\epsilon^2)+\lambda & -2 \\ -2 & 2+\lambda \end{pmatrix},$$

where

$$|\mathsf{A}_\lambda| = (2\epsilon^2 + \lambda)(2+\lambda) + \lambda.$$

Show that $\mathsf{A}_\lambda^{-1}\mathsf{E}^T$ reduces to E^{-1} when $\lambda = 0$, and persuade yourself that you expected that.

Thus, the source vector recovered from a data vector $\hat{\mathbf{g}}$ is

$$\hat{\mathbf{u}}_\lambda \equiv \mathsf{A}^{-1}\mathsf{E}^T\hat{\mathbf{g}}.$$

Given a data vector $\hat{\mathbf{g}} = (1, 1+\delta)^T$, with noise δ (there is no significance to the first component being noise-free – I just want your calculations to fit on one ream of paper), show that

$$\hat{\mathbf{u}}_\lambda = \frac{1}{|\mathsf{A}_\lambda|}\begin{pmatrix} 2(2+\delta)\epsilon^2 - 2\delta\epsilon + (2+\delta)\lambda \\ 2\delta\epsilon + (2+\delta+\delta\epsilon)\lambda \end{pmatrix}.$$

As checks, show that $\hat{\mathbf{u}}_{\lambda=0}$ is the unstable $\hat{\mathbf{u}}$ obtained above, and that for $\delta = 0$, $\hat{\mathbf{u}}_0 = (1,0)^T$, as above. Note that

$$\hat{\mathbf{u}}_\lambda \to \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \text{as } \lambda \to \infty,$$

illustrating the way that zeroth-order regularisation pulls a solution towards the prior estimate $(0,0)^T$.

Evaluate $\hat{\mathbf{u}}_\lambda$ for $\epsilon = 0.1$, $\delta = 0.1$ and $\lambda = 0.01$, 0.1, and 1. Note that none of these produce a particularly good recovery with this level of noise, but that the one with $\lambda \approx \epsilon$ is probably least unreasonable. Evaluate $\hat{\mathbf{u}}_l$ again with $\epsilon = 0.1$, $\delta = 0.01$ and $\lambda = 0.01$, 0.1 and 1. Again, the recovery with $\lambda \approx \epsilon$ appears to be the best compromise between resolution and stability, with the $\lambda = 0.01$ being undersmoothed, and the $\lambda = 1$ oversmoothed.

If you wish, try rewriting A_λ, substituting $\lambda = \lambda'\epsilon$. Take ϵ small (ie, discard positive powers of ϵ), and $\lambda' \sim 1$ (ie, taking $\lambda \sim \epsilon$), and obtain an expression for $\delta\hat{\mathbf{u}} = \mathsf{A}_\lambda^{-1}\mathsf{E}^T\delta\hat{\mathbf{g}}$, where $\delta\hat{\mathbf{g}} = (0, \delta)^T$. You can thus see that, *in this particular problem*, the error in the recovery is of order the error in the data when λ of order ϵ.

In a real problem, you would choose an appropriate value of λ for your problem by some more sophisticated technique. Also, remember that the explicit inversion used here is for illustration only; you would *not* do this in a real problem, but instead use an appropriate numerical technique. See *Numerical Recipes* for suitable algorithms.