Ill-posed inference and sampling from a high-dimensional sheet-like distribution

Alex Barnett

rough ideas, June 2002, updated December 2002

1 Introduction to the problem

Consider a general inverse problem. You have unknown parameter vector $\mathbf{x}$ of N dimensions. The forward model gives signal (data) vector $\mathbf{f}(\mathbf{x})$ of M dimensions, which is to be compared to the (experimentally) measured signal vector $\mathbf{y}$ (also M dimensional). In simple contexts (*e.g.* linear least squares) solving the inverse problem is often called 'fitting'. More generally it can be called 'inference' for the parameters $\mathbf{x}$. We happen to be interested in

1. what the 'fit' tells you about the distribution of parameters $\mathbf{x}$,
2. the resulting distribution of another function $z(\mathbf{x})$.

We are interested in handling the case where $\mathbf{x}$ is not well-determined by the signal. We will discuss a method which handles sampling of the distribution in $\mathbf{x}$ efficiently when the distribution is *sheet-like* (of dimensionality less than N). The basic idea is to use Markov chain Monte Carlo sampling in some directions and Gaussian approximation in others.

For David Boas' hemodynamic modeling example, $\mathbf{x}$ are hemodynamic parameters ($N = 13$), $f_i(\mathbf{x})$ are the M values of a time-series of [ΔHbT] computed from a ODE forward model, y_i are a corresponding set of experimental measurements of [ΔHbT]. Our function is $z(\mathbf{x})$, a ratio involving [$\Delta CMRO_2$], which can be computed from CBF, which itself comes from the forward model, and other experimental values including y. The forward model $\mathbf{f}(\mathbf{x})$ takes a few seconds per evaluation, apparently.

2 Going from a 'fit' to inference for $\mathbf{x}$

Here I explain the transition to a Bayesian approach, starting from familiar territory. We quantify differences between $\mathbf{f}(\mathbf{x})$ and $\mathbf{y}$ by a 'cost function'. If 'least-squares fitting' with constant known noise level σ is used then the cost function is

$$\chi^2(\mathbf{x}) \equiv \sum_{i=1}^{M} \frac{[f_i(\mathbf{x}) - y_i]^2}{2\sigma^2} \tag{1}$$

Minimizing this function in the space of $\mathbf{x}$ can give a single best fit $\mathbf{x}_{BF}$ where the value is χ^2_{BF}. However for ill-posed problems there are generally well-determined directions in $\mathbf{x}$-space (curvature of χ^2 is high) and ill-determined directions (curvature is low or zero). A simple example is *amplitude calibration parameters* in DOT, which are often much more tightly constrained than the optical model parameters. If $\mathbf{f}(\mathbf{x})$ is nonlinear then these directions change with $\mathbf{x}$, giving extended *sheet-like* valleys (gorges) where χ^2 is almost equally small, which can bend and fold. If there are some parameters which never have much effect on $\mathbf{f}$ then the (hyper-)plane of the sheet will be parallel to these parameter axes. Sometimes there are local minima at more than one location within the sheet.

How can we formalize the fact that many values of $\mathbf{x}$ explain the experimental data equally well or almost equally well? The best way is to imagine that the forward model together with the experimental data and the noise model define a probability distribution over $\mathbf{x}$ — the *likelihood*. The 'least squares fit' is equivalent to maximizing a likelihood function

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z_\sigma} e^{-\chi^2(\mathbf{x})}, \tag{2}$$

where the equivalence arises simply because the negative exponential is a monotonically-decreasing function. But now we can interpret this likelihood as follows: it corresponds to a model of the noise where the data y_i are produced by adding independent gaussian noise (of standard deviation σ) to the underlying model output values. The constant Z_σ depends only on σ and M, and ensures that $\int p(\mathbf{y}|\mathbf{x}) d\mathbf{y} = 1$. The likelihood function can be confusing in that it can be viewed as a function of $\mathbf{y}$ with $\mathbf{x}$ constant (probability of getting the data given underlying model output), or of $\mathbf{x}$ with $\mathbf{y}$ constant (likelihood of certain model parameters given an experimental measurement). From now on we stick to the latter.

If we have knowledge or assumptions about $\mathbf{x}$ regardless of $\mathbf{y}$, these are called the *prior* distribution of $\mathbf{x}$. In our example each x_i has lower and upper bounds, but no other knowledge. If there is not a large ratio between lower and upper bounds for each parameter, it is then reasonable to take the prior on $\mathbf{x}$ to be uniform in the hyper-cuboid defined by the bounds.

The product of prior and likelihood gives the *posterior* distribution on $\mathbf{x}$, $p_{\text{post}}(\mathbf{x})$, which is everything we have learned about $\mathbf{x}$ so far.

In our example, the prior's only effect is to change the normalization of the likelihood, and to set it to zero outside the bounds. So

$$p_{\text{post}}(\mathbf{x}) \propto e^{-\chi^2(\mathbf{x})} \quad \text{inside bounds, zero otherwise.} \tag{3}$$

The posterior is large in the sheet-like regions (valleys of the cost function), and exponentially small away from the sheet. A change of $+1$ in χ^2 corresponds to $1/e$ decrease in probability. Decreasing σ relabels the contour surfaces of $p_{\text{post}}(\mathbf{x})$, making the sheet thinner, and causing the distribution to pull itself in along the valleys towards the global maximum (or global maxima if it is not unique).

We now can answer the usual inference questions, *e.g.*

- The expectation of $\mathbf{x}$ is $E[\mathbf{x}] = \int \mathbf{x} p_{\text{post}}(\mathbf{x}) d^N\mathbf{x}$.
- 1-dimensional marginal distributions in the parameters x_i are can be found by projecting $p_{\text{post}}(\mathbf{x})$ onto each axis i. Note that this requires $(N-1)$-dimensional integration.
- The variance (or, more robustly, the 25% and 75% quantiles) of x_i can be found from these 1-dimensional distributions.
- We can infer anything about any function $z(\mathbf{x})$...

3 Review of methods for handling high-dimensional posteriors

The basic inference questions about $\mathbf{x}$, as well as those about z to follow, formally seem to require integration of $p_{\text{post}}(\mathbf{x})$ over N dimensions. However evaluating $p_{\text{post}}(\mathbf{x})$ everywhere in $\mathbf{x}$-space is rarely possible for N more than about 3 or 4. (One forward model run is needed for each $p_{\text{post}}(\mathbf{x})$ evaluation).

The two usual methods to handle this for larger N are:

1. The posterior is approximated as a multivariate Gaussian distribution,

$$p_{\text{post}}(\mathbf{x}) \approx \frac{1}{Z} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{x}_0)^{\mathrm{T}}\Sigma^{-1}(\mathbf{x}-\mathbf{x}_0)}, \tag{4}$$

with normalizing constant

$$Z = \sqrt{(2\pi)^N \det \Sigma}. \tag{5}$$

This is justified if σ is small enough that $\mathbf{f}(\mathbf{x})$ is approximately linear in the $\mathbf{x}$-region with significant probability. In this case the covariance matrix is $\Sigma = H^{-1}$, where the hessian is

$$H_{ij} \equiv \left. \frac{\partial^2}{\partial x_i \partial x_j} \chi^2 \right|_{\mathbf{x}=\mathbf{x}_0} \tag{6}$$

and can be approximated, in this noise model, by

$$H \approx \frac{1}{\sigma^2} J^{\mathrm{T}} J \tag{7}$$

where the jacobean is $J_{mn}(\mathbf{x}) \equiv \partial f_m / \partial x_n|_{\mathbf{x}}$. In our example the maximum is at $\mathbf{x}_0 = \mathbf{x}_{BF}$ because the prior is flat. Integrals over $\mathbf{x}$ can now be performed fast, often analytically. J can be measured in $(N+1)$ forward model runs using finite differencing along each dimension, if no explicit derivative information if available from the forward model (see Griewank book).

There now remains the *optimization problem* of finding $\mathbf{x}_0$, and H at this point. This problem is much faster if derivative information J is available as the search proceeds.

A good analogy for optimization is of walking blindfolded on a mountainous terrain, taking steps in directions of your choosing, with the goal of finding the lowest point on the terrain.

2. Monte Carlo (*i.e.* random) sampling of $p_{\text{post}}(\mathbf{x})$ to give a small number S of independent samples $\mathbf{x}_s$ from the distribution. The integrals can then be estimated by (possibly weighted) sums over s. The estimates converge like $1/\sqrt{S}$, independent of the dimension N. The samples are drawn by either

   (a) choosing independent random $\mathbf{x}_s$ then rejecting (or weighting) based on $p_{\text{post}}(\mathbf{x}_s)$. This is called importance sampling. Or,

   (b) building a *Markov Chain* where each $\mathbf{x}$ is updated by a random step from the previous value; samples are then taken sufficiently far apart along the chain as to be independent. The Metropolis walk is an example. General name is MCMC.

See MacKay's excellent introduction to Monte Carlo sampling methods.

The great thing about sampling methods is that they give you the marginal distributions over parameters x_i trivally (by plotting only the i coord of the samples).

The not-so-great thing is their potential to be extremely inefficient unless performed intelligently (and therefore often problem-specifically).

In the case where the posterior is extended over a bending, winding sheet-like structure, both above methods have difficulties:

1. The Gaussian approximation is bad, so inferences based on local curvature are wrong.
2. Conventional Monte Carlo methods are inefficient:
   (a) Importance-type sampling is inefficient because random samples almost never hit the sheets, because being thin in many dimensions, they have very small volume fraction, and
   (b) Markov Chain methods take prohibitively long to explore the sheet by random walking using small step sizes. This can be cured using 'hybrid MCMC' which simulates hamiltonian motion in the cost function potential well — this requires derivative information $(\partial(\chi^2)/\partial x_i)$ at each $\mathbf{x}$ in the chain.

4 Now we know something about $\mathbf{x}$, what can we know about $z(\mathbf{x})$?

z is any function of $\mathbf{x}$.
Practicalities aside, $p_{\text{post}}(\mathbf{x})$ gives you full distribution $p(z)$.
Formulae for inference about z.
Given posterior over $\mathbf{x}$, expectation is

$$E[z] = \int z(\mathbf{x}) p_{\text{post}}(\mathbf{x}) d^N\mathbf{x}. \tag{8}$$

Variance is

$$\text{Var}[z] = \int (z(\mathbf{x}) - E[z])^2 p_{\text{post}}(\mathbf{x}) d^N\mathbf{x}. \tag{9}$$

Full distribution is more messy,

$$p(z) = \int \delta(z - z(\mathbf{x})) p_{\text{post}}(\mathbf{x}) d^N\mathbf{x}, \tag{10}$$

which boils down to the usual jacobean relation $p(z)dz = p(x)dx$ in the 1d monotonic case.

If you have access to independent samples $\mathbf{x}_s$ from $p_{\rm post}(\mathbf{x})$, then the set $z(\mathbf{x}_s)$ are independent samples from the distribution of z. This (histogram-type) method is good way to get $p(z)$.

Crude 'importance' sampling — an example of a weighting method. You don't have access to independent samples. Draw $\mathbf{x}_s$ from a *uniform* distribution in $\mathbf{x}$-space, and assign them weights $w_s = p_{\rm post}(\mathbf{x}_s)$. You then can estimate the expectation value of z via

$$E[z] \approx \hat{E}[z] = \frac{\sum_s w_s z(\mathbf{x}_s)}{\sum_s w_s}. \tag{11}$$

Similarly, the full distribution $p(z)$ can be estimated by histogramming the $z(\mathbf{x}_s)$ using the weights w_s to add to the histogram rather than the usual 1. Note that the convergence still goes like $1/\sqrt{S}$ but there is an extra factor of the volume fraction now too. (The effective number of samples is only S times the volume fraction). For sheet-like structures, this ruins the efficiency.

5 Scheme for inference with a sheet-like posterior

We propose a combination of the two general schemes from Section 3, which handles the thin dimensions of the sheet with Gaussian approximations and the extended dimensions using Monte Carlo sampling.

There is an easier way to do this than setting up a curvilinear coordinate system inside the sheet (which would be a mess). If you take a (hyper-)sheet $p_{\rm post}(\mathbf{x})$ and keep projecting out (integrating over) whichever is the current tightest-constrained coordinate, it will eventually lose sheet-like character and start to occupy a non-small volume fraction in the space of the remaining coords. Imagine a curvy line in 3d space: if you project once, you get a line in 2d (still 'sheet-like', *i.e.* small volume), but if you project again, get a distribution in 1d which occupies a large volume fraction.

For integrals of $z(\mathbf{x})$ over this posterior, we have to consider if the variation of z across the thickness of the sheet in the projected dimensions is significant. We will assume it is not, *i.e.* z is constant, and we can use the mean value.

Call $\mathbf{x}^{\rm T} = (\mathbf{a}^{\rm T}, \mathbf{b}^{\rm T})$, where $\mathbf{a}$ is the subspace of the dimensions to integrate out, and $\mathbf{b}$ are the remaining ones. $N_a = \dim(\mathbf{a})$. $N_b = \dim(\mathbf{b})$. $N_a + N_b = N$. We need to find a split such that :

1. The conditional $p(\mathbf{a}|\mathbf{b})$ is approximately Gaussian as a function of $\mathbf{a}$, with a single Gaussian peak, for all (or most) relevant $\mathbf{b}$. This can be checked by plotting slices of $p_{\mathrm{post}}(\mathbf{x})$, or of the cost function, in the $\mathbf{a}$ directions. We also need that $z(\mathbf{a},\mathbf{b})$ is close to constant over the gaussian in $\mathbf{a}$, which for smooth z and thin sheet is likely to be true (note that only even terms in Taylor expansion of z about the peak will contribute to lowest order, due to symmetry of the Gaussian).

2. The marginal $p(\mathbf{b})$ is not sheet-like in $\mathbf{b}$-space. For weighting-type methods this means it should occupy an acceptably large fraction of the volume in $\mathbf{b}$-space (*e.g.* if it's 1% then 100 forward models have to be run per effective sample).

If the sheet doesn't fold back on itself, then this should be easy (need pictures here!). The choice should just be governed by which conditional distributions (slices) are tightest constrained. With luck there will be many ways (compare a couple of choices to validate).

Marginalizing via a Gaussian integral in $\mathbf{a}$ gives

$$p_{\mathrm{post}}(\mathbf{b}) \;=\; \int p_{\mathrm{post}}(\mathbf{a},\mathbf{b})\, d^{N_a}\mathbf{a} \;\approx\; \sqrt{\frac{(2\pi)^{N_a}}{\det H_a(\mathbf{a}_0)}} \cdot p_{\mathrm{post}}(\mathbf{a}_0,\mathbf{b}) \tag{12}$$

where $\mathbf{a}_0$ is the $\mathbf{a}$ where posterior is maximized (*i.e.* χ^2 is minimized in our example), with $\mathbf{b}$ held fixed at a given value. $H_a(\mathbf{a}_0)$ is the hessian at this point, which can either be found numerically or by Eq. (7) restricted to the $\mathbf{a}$ subspace.

Let's assume you choose the crude importance sampling scheme from the previous Section. Say for instance it turns out that the split of $\mathbf{x}$-space into the two subspaces is $\mathbf{a}^{\mathrm{T}} = (x_1, x_2, x_3)$ and $\mathbf{b}^{\mathrm{T}} = (x_4, x_5, \ldots, x_{13})$, for the $N = 13$ dimensional example from the Introduction. The recipe is the same as before except you restrict to $\mathbf{b}$-space, and for each sample divide the maximum posterior found (via 'fitting' in $\mathbf{a}$-space) by the square root of the $\mathbf{a}$-space hessian to get the weight of that sample. The value of z for each sample should be taken at $(\mathbf{a}_0, \mathbf{b})$, since we assumed it is constant within the $\mathbf{a}$-peak. (Formalise this better, starting from integral for $E[z]$, etc).

The overall effort to get the equivalent of 1 independent sample is about (1/volume fraction in $\mathbf{b}$ sampling space)$\times N_{its} \times (N_a + 1)$ forward model runs. N_{its} is the typical number of iterations to find $\mathbf{a}_0$ in $\mathbf{a}$-space — this will be less than 10 if we use gradient and hessian information (Newton-type method). The final factor (N_a+1) accounts for the measurement of J in the $\mathbf{a}$ subspace. The hessian should be estimated with Eq. (7) restricted to the

a subspace. To estimate $E[z]$ or $\mathrm{Var}[z]$ you will need ~ 10 such independent samples. To estimate the full distribution $p(z)$ you'll need ~ 100 or more.

If you want more advanced weighting-type methods, MacKay article has them.

I think MCMC would be great for **b**-space sampling.