A comparative study of 'inversion', 'optimization', and 'inference' as frameworks for imaging

Colin Fox[1] **and Nicholas Dudley Ward**[2]

[1] Electronics Group, Department of Physics, University of Otago, PO Box 56, Dunedin, New Zealand
[2] Groundwater Resources Research, 54 Dudley Street, Christchurch, New Zealand

Email: fox@physics.otago.ac.nz & nick@groundwater.org.nz

Abstract: We give four examples that examine the features, and foibles, of performing imaging by *direct inversion*, *optimizing an objective function*, and *model-based inference*.

Keywords: Imaging, regularization, Bayesian inference, heat transfer, hydrology

1. INTRODUCTION

Millman and Grabel noted in their 1987 textbook [1], that the diverse activities encompassed by the term 'electronics' share the common property of processing information. In fact, a definition of modern electronics could be 'the processing of information represented as electrical signals'.

A central part of information processing is the analysis of measurements. We often interpret measurements in terms of physical models as a means of probing the world around us. The question we address in this paper is: what framework should guide us in doing that analysis? We present four examples that display the features, and foibles, of three common frameworks. We use the term 'imaging' to refer to data analysis because the primary unknown in our work is usually a spatially-varying quantity that is usefully displayed as an image.

The three frameworks we compare in this paper are *direct inversion*, *optimizing an objective function*, and *model-based inference*. The tools used in 'signal processing' largely come from the first two of these, and have become essential knowledge in electrical and electronic engineering. The third framework is now seen as the 'gold standard' in imaging, and we are starting to see the first measurement equipment built around inferential methods.

It might come as a surprise to practitioners in electronics that statisticians have played a key role in developing these frameworks, and are largely responsible for the framework of signal processing as practised in electrical engineering. For example, it was the statistician John Tukey who brought us the Cooley-Tukey FFT algorithm, and also the word 'bit'.

The connection between imaging, and statistics arises because every measurement is corrupted to some degree by noise. Hence any measurement process is naturally described using probability, and the task of estimating the unknown true image is naturally one of statistical inference. In fact the discipline of statistics has its origins in such problems, with Laplace's work in astronomy, and the later work by Jeffreys in geophysics [2].

2. INVERSION

2.1 Example 1: Image deblurring

Figure 1 shows a test image, with 200×200 8-bit gray scale pixels, and the result after blurring using the Matlab command `blurred = conv2(test,ones(5,5)/25,'same');`

Figure 1: A gray scale test image (left) and a blurred version (right).

Most students of electronics will have taken some courses in mathematics, and so have been exposed to the idea of inversion of a function. This is by far the most often suggested framework for imaging by those who have little or no experience in practical data analysis. In this section we look at using direct inversion to perform deblurring, and see why it will (always) fail for practical problems.

In this case the blurring, or forward map, is given by a convolution, so the image f and data d are related by

$$d = f * h$$

when h is the 'point spread function'. A standard result is then that the Fourier transforms F, D, and H of f, d, and h, respectively, are related simply by

$$D = F \times H, \text{ or } F = D/H.$$

In Matlab this deblurring by Fourier division looks like `f = real(ifft2(D./H));` and the result is shown in the left-hand panel of Figure 2. As you can see, the result is not good.

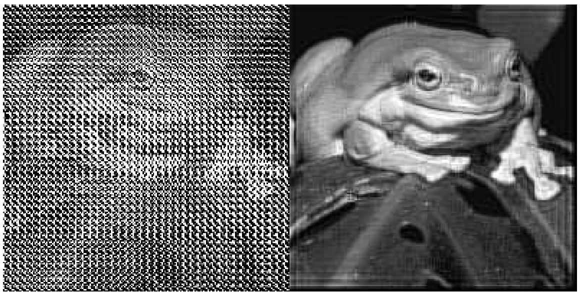

Figure 2: Left: image evaluated by direct inversion. Right: image evaluated by Tikhonov regularized least-squares inversion, using $\lambda = 0.14$.

2.2 *What went wrong with the inverse?*

In this case the forward map is invertible, so the route of direct inversion should work. The magnitude of the components of H (which are also the singular values of the forward map) are plotted in Figure 3, and though some singular values are small, none are actually zero or even near machine precision. We can't blame the failure of direct inversion on noise in the data since we added no noise (apart from round-off error) to the blurred image, so everything should work, right? Well, no. The problem is that the forward map is not *exactly* a convolution, because we trimmed the data to have the same size as the image. More than 5 pixels from the edge of the picture the forward map is exactly a convolution, but it is not quite for a band near the edges. That small *model error* is enough to make direct inversion near useless, with the effect that mis-modelling at the edges propagates catastrophically through the image.

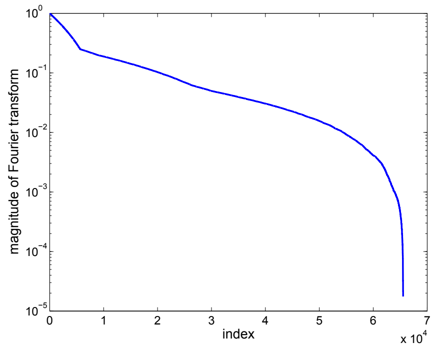

Figure 3: The singular values of the forward map, sorted in descending order.

When the data does contain noise, direct inversion will also fail because the small singular values ensure that the reconstructed image is dominated by noise.

3. OPTIMIZATION OF AN OBJECTIVE FUNCTION

A common route around the problems of direct inversion is to apply a regular approximation to the inverse of the forward map A. The usual way to do that is to minimize an objective function such as

$$Q(f) = \|Af - d\|^2 + \lambda^2 \|f\|^2$$

that balances the data misfit $\|Af - d\|^2$ against a measure of image quality – in this case the Tikhonov regularizing functional $\|f\|^2$. The balance is controlled by the regularizing parameter λ. When $\lambda = 0$ we get usual least-squares fitting to data, which reduces to direct inversion when A is invertible. More generally the minimum is the solution of the *generalized deconvolution equation*

$$\left(A^{\mathrm{T}}A + \lambda I\right) f = A^{\mathrm{T}} d.$$

In signal processing this linear equation is known as the Wiener filter.

The usefulness of this approach can be seen in the right panel of Figure 2 that shows the deblurred image evaluated with $\lambda = 0.14$, using the Matlab

```
F = D.*conj(H)./(abs(H).^2+lambda.^2);
 f = real(ifft2(F));
```

3.1 *Stochastic bias in regularized inverses*

For non-linear problems, least-squares estimates and regularized estimates suffer from *stochastic bias*, i.e. a systematic offset in estimated values resulting from errors on measurements. Stochastic bias may be displayed in a simple form, as follows: If x is a (scalar) random variable from any distribution with mean $E(x) = \mu$ and variance σ^2, then $E(x^2) = \mu^2 + \sigma^2$. That is, in the presence of noise on variable x, the mean of the square equals the square of the mean *plus the variance of the noise*. The bigger the noise, the bigger the offset. Stochastic bias is a very real effect that, for example, is one of the mechanisms that allows profit to be made in volatile markets, whether increasing or decreasing [3]. The following example shows that stochastic bias occurs in (regularized) least-squares estimation.

3.2 Example 2: Estimating heat capacity from impulsive heating

Consider the problem of determining the heat capacity c in a 3-dimensional homogeneous medium of large extent for which the thermal conductivity k is known. We consider the stylized problem where the medium is subject to unit impulsive heating at $r = 0$ and time $t = 0$, and the temperature at the point of heating is subsequently measured. For this problem the forward map is available analytically, with the noise-free temperature at time t

$$T(t) = \frac{c^{1/2}}{(4\pi k t)^{3/2}}. \tag{1}$$

If measurement d_i is made at time $t = it_{\mathrm{s}}$, $i = 1, 2, \ldots K$, and is subject to additive noise with zero mean and variance σ_{n}^2, then the least-squares estimate of c is

$$\hat{c} = \arg\min_{c} \sum_{i=1}^{K} \left(d_i - b\frac{\sqrt{c}}{i^{3/2}} \right)$$

where $b = (4\pi k t_{\mathrm{s}})^{-3/2}$. The normal equations are solvable in this case, and give

$$\hat{c} = \left[\frac{\sum_{i=1}^{K} d_i / i^{3/2}}{b \sum_{i=1}^{K} 1/i^3} \right]^2 = w^2.$$

The term in the square brackets, denoted w, is the sum of many random variables and hence w is a random variable with mean $\sqrt{c}$ and variance $\sigma_{\mathrm{n}}^2 / \left(b^2 \sum_{i=1}^{K} 1/i^3 \right) \rightarrow \sigma_{\mathrm{n}}^2 / \left(b^2 \zeta(3) \right)$ as $K \rightarrow \infty$. Here $\zeta(3) \approx 1.2021$ is the Riemann zeta function evaluated at 3. Using the result quoted above, for a large (infinite) number of measurements, the least-squares estimate of c has expected value $\hat{c} = c + \sigma_{\mathrm{n}}^2 / \left(1.2021 \times b^2\right)$. That is, in the presence of measurement error the least squares estimate is systematically biased, for all K. Note that reducing the measurement error gives less bias. However, increasing the number of measurements makes the estimate worse (not better as is often claimed) by increasing bias.

In this case the bias may be easily removed (by subtracting $\sigma_{\mathrm{n}}^2 / \left(1.2021 \times b^2\right)$) since we were able to calculate the bias analytically. The result is a better estimator, having the same variance but lower bias, and is explicitly not the least-squares estimator. Unfortunately the obvious conclusion, that best-fit to data is not the same as best-fit to parameters, is not commonly observed in the imaging literature. In most imaging problems we are not able to determine the bias analytically, making the least squares estimate both biased and difficult to fix. The application of regularization actually compounds this problem. For example, for this stylised example the Tikhonov-regularised estimate may be calculated analytically, and has bias that is dependent on the *unknown* value of c, leaving an implicit problem to remove bias.

It is instructive to note that the quantity w is an unbiased estimator for $\sqrt{c}$, since the data is a linear function of $\sqrt{c}$. Hence the least squares estimate of $\sqrt{c}$ makes a good estimate, but its square is not a good estimate of c. This apparently paradoxical behaviour is an example of how the algebra of (random) variables with uncertainty is quite different to the algebra of deterministic variables (see e.g. [4]). For this reason it is necessary to track the *distribution* of possible values that a variable can take, not just the single 'best' estimate. Maintaining and summarizing distributions over variables is the basis of Bayesian inference.

Anticipating section 4., we briefly describe how a Bayesian approach could solve this example. The likelihood function combines the forward map in equation 1 and the distribution over measurement noise. As is typical in inverse problems, the range of the forward map is a small fraction of data space, and so the noise statistics may be determined from the measurements. In this case, at large times when $T(t) \approx 0$ the data solely consists of measurements of the noise. Hence the noise distribution can be determined. The form of the likelihood function depends on the noise statistics, and in this way the Bayesian approach can deal with *any* noise distribution. For this simple single parameter estimation problem, an 'objective Bayesian' analysis is feasible, by choosing the Jeffreys type [2] prior distribution that is invariant to choice of units for heat capacity, giving $\pi_{\mathrm{p}}(c) \propto c^{-1/2}$. This is an 'improper' prior distribution and would require modification if only a few measurements were available, though then the data so poorly constrains c that no approach would give good results. When the data is adequate, the posterior mean gives a suitable point estimate for c. In the ideal case where the noise is independent identically distributed (iid) zero-mean Gaussian, and measurements are very accurate, there is little to choose between the least squares estimate and the posterior mean, except in computational cost. In most other circumstances the posterior mean does a much better job of estimating the unknown true heat capacity.

3.3 Example 3: Uncertainty in estimates with uniform noise: a counter example to estimators

The ability to calculate data-dependent or posterior variance is a distinct advantage of a Bayesian approach to inverse problems. In contrast, methods such as least squares are justified on the basis of the variance of the *estimator*, i.e. the average variance over all possible measurements. Yet, in many practical inverse problems the variance of the estimator has nothing to do with the uncertainty in parameters estimated from the data set at hand. This example demonstrates that issue in a very simple setting, where the forward map is the identity function with uniform measurement errors.

Consider the simple case where a scalar quantity μ is measured directly, subject to uniform noise, with mean zero and width 2. Then the i^{th} measurement is

$$d_i = \mu + n_i$$

where each $n_i \sim U(-1, 1)$, i.e. is uniformly distributed over the interval $[-1, 1]$. Since $\mu - 1 \leq d_i \leq \mu + 1$ for all i, it follows that $\max\{d_i\} - 1 \leq \mu \leq \min\{d_i\} + 1$. In fact these bounds are exactly what the data tell us about μ. We note that the likelihood function precisely expresses these bounds, and so they are automatically included in a Bayesian analysis.

The least squares estimate of μ from K measurements is easily seen to be

$$\hat{\mu}_{\text{ls}} = \frac{1}{K}\sum_{i=1}^{K} d_i.$$

It is instructive to note that this estimate can lie *outside* the interval $[\max\{d_i\} - 1, \min\{d_i\} + 1]$, in which case it is not even consistent with the measured data. For $K = 10$, this happens a little more that 30% of the time, so in 1 out every 3 experiments the least squares estimate is not even a possible value. More troublesome, in practice, is that the error often quoted for the least-squares estimate has little to do with the actual uncertainty in the value of μ as determined by the data. From the considerations above, we see that $\mu \in \frac{1}{2}(\max\{d_i\} + \min\{d_i\}) \pm \frac{1}{2}(2 + \min\{d_i\} - \max\{d_i\})$ so the uncertainty is (certainly) $1 + \frac{1}{2}(\min\{d_i\} - \max\{d_i\})$. The mean-square-error for the least squares estimator of $1/\sqrt{3K}$ is often quoted as the error in the least-squares estimate. Note that it is independent of the data. Three simulations for the case $\mu = 0$ and $K = 2$, returned the values $(d_1, d_2) = (-0.7477, 0.6688)$, $(-0.6112, -0.6136)$, and $(0.6278, -0.0376)$ giving estimates with posterior error ± 0.2918, ± 0.9988, and ± 0.6673. This is sometimes larger, and sometimes smaller than the least-squares error of ± 0.4082.

Posterior error estimates are particularly informative when recovering spatially-varying parameters such as the conductivity of an inhomogeneous material. It is clear on physical grounds that the spatial dependence of uncertainty in a reconstruction must be data dependent. For example, when a region of low conductivity is surrounded by a region of high conductivity, the outer region shields the inner region from external sources. Hence the conductivity of the inner region cannot be accurately determined from measurements based on external sources. However, if the whole medium has similar conductivity, then energy can flow through all regions, resulting in more accurate estimation. Since the data reflects the distribution of conductivity, the spatially-varying error must be dependent on the data. As mentioned above, the mean square error of the least-squares estimate, or of any other fixed estimator, gives no clue to this effect.

4. BAYESIAN INFERENCE

4.1 Example 4: Aquifer Parameters from Pump Test Data

In this example, we consider the analysis of 'pump test' data used to determine the groundwater availability at a site, and to predict the effects of drawing water from a bore at that site. A pump test consists of pumping groundwater from a borehole and monitoring the groundwater level, or drawdown, at a second borehole at some distance. Estimation of aquifer parameters from pump test data is traditionally based on a graphical approach: pump test data is plotted and parameters from a hydraulic model are varied to achieve a 'best' fit to the data. With the advent of the personal computer this process is commonly automated using a least-squares fit.

In practice pump test data is corrupted by noise which means that even under ideal conditions two pump tests from the same site will result in the different drawdown traces, and hence different least square parameter estimates. In this example we concentrate on measurement error, in contrast to the model error that dominated in the Example 1. We assume then that we observe noise perturbed measurements d of the 'true' drawdown s in a monitoring bore:

$$d = s + \text{noise}. \tag{2}$$

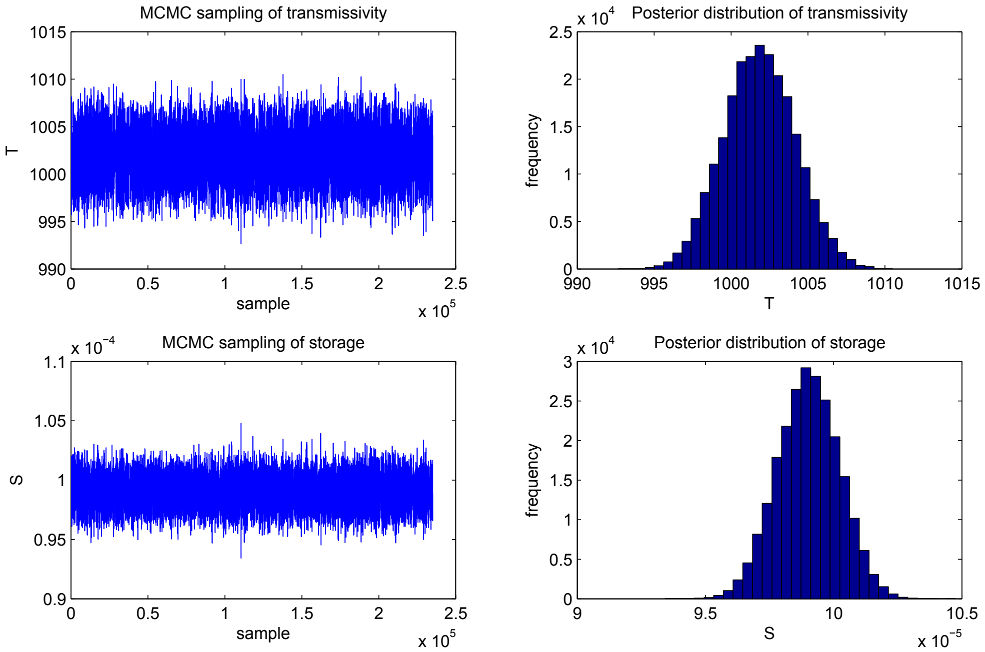

Figure 5: T-walk sampling for the Theis model, after burn-in. Left: output traces of transmissivity T (top) and storage S (bottom). Right: posterior histograms for transmissivity T (top) and storage S (bottom).

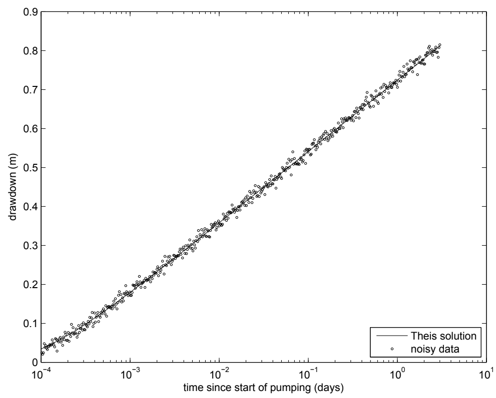

Figure 4: Synthetic pump test data for the Theis model

For simplicity we take the noise to be iid Gaussian distributed with mean 0 and standard deviation σ.

We simulated data using the classical Theis hydraulic model [5], parametrized by transmissivity and storage parameters. Let $s = s(r,t)$ be drawdown at a distance r from the pumping well at time t. Then s solves the Theis equation

$$S\frac{\partial s}{\partial t} = T\nabla^2 s + Q\delta(x)\delta(y),$$

where ∇^2 is the Laplacian, S is storage, T is transmissivity, Q is a constant pumping rate and $\delta(\cdot)$ is the Dirac δ-function. The Theis model has the analytic solution

$$s(r,t)|(T,S) = -\frac{Q}{4\pi T}\,\mathrm{Ei}\left(-\frac{Sr^2}{4Tt}\right) \qquad (3)$$

where $\mathrm{Ei}(\cdot)$ is the exponential integral. The notation $s(r,t)|(T,S)$ is read 's of r and t *given* T and S' and indicates that the true drawdown is parametrized by transmissivity and storage parameters.

Equation 3 defines the *forward map* from the parameters (T,S) to noise-free measurements of drawdown. We assume that measurements are made at times $t_1, t_2, \ldots, t_N$ that we denote $d_1, d_2, \ldots, d_N$. Under the assumption of Gaussian noise, above, the probability density over measurements has the form

$$\pi\left(\{d_i\}\,|T,S\right) = \frac{1}{(2\pi\sigma^2)}\exp\left\{-\frac{\sum_{i=1}^N \left(d_i - s(r,t_i)|(T,S)\right)^2}{2\sigma^2}\right\} \qquad (4)$$

For a numerical example we suppose that we measure noise perturbed measurements of drawdown in an observation well 50 m from a well pumped over three days with a pump rate of $Q = 1{,}000\ \mathrm{m}^3/\mathrm{day}$. The true transmissivity is $T = 1{,}000\mathrm{m}^2/$ day and storage is $S = 1 \times 10^{-4}$. Figure 4 shows one set of measurements simulated

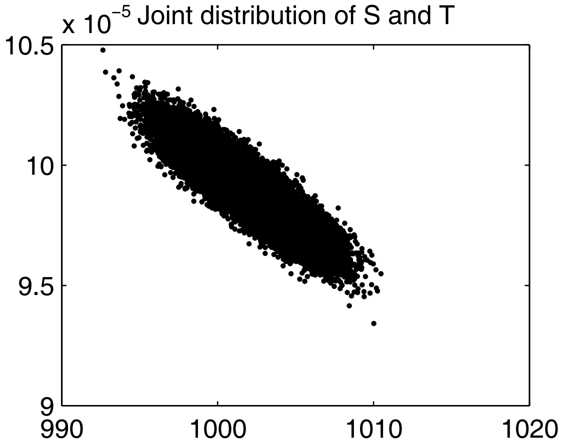

Figure 6: Scatter plot showing the joint distribution of parameters

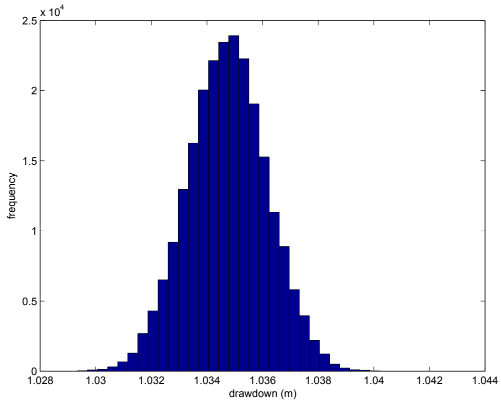

Figure 7: Histogram over drawdown predictions in the monitoring well after 50 days of pumping

by evaluating the forward map and then adding zero mean Gaussian noise with a standard deviation of 0.005 m.

The posterior distribution over unknown parameters is easily explored using Markov chain Monte Carlo (MCMC) algorithms [6]. Even as recently as five years ago, implementing an MCMC for any problem, including this simple two parameter case, would require learning how to write an MCMC algorithm, and then *tuning* the proposal densities to give reasonable performance. Fortunately now there are a few black-box sampling algorithms available where the user simple provides the (log) target density and the black-box does the rest. Here we use the t-walk [7], which is not the fastest algorithm in terms of computation, but is easily the fastest in terms of time between forming the posterior distribution and having results that solve the problem. Figure 5 shows output traces from the t-walk sampling, and posterior histograms for T and S.

More importantly than finding 'best' estimates for the unknown parameters, the sampler output can be used to determine joint distribution over parameters that is implied by the data and model. Figure 6 shows a scatter plot of the joint distribution over S and T, showing that the data asserts a negative correlation.

The purpose of performing pump tests is, ultimately, to predict the effect on the water table of drawing water from the borehole. Since the data and model define a (posterior) distribution over parameters, accordingly the data implies a distribution over possible drawdowns. Figure 7 shows the drawdown prediction in the monitoring well after 50 days of pumping. The variability of prediction could be quantified in terms of a mean and standard deviation. This will give the actual degree of certainty in drawdown. In contrast, the drawdown predicted by the least-squares estimate, as noted earlier, will exhibit a stochastic bias with no indication of accuracy.

5. CONCLUSION

The frameworks compared for imaging of *direct inversion*, *optimizing an objective function*, and *model-based inference* each have advantages and drawbacks. At present the gold standard are the Bayesian inferential methods. These methods are now straightforward to apply, but remain costly in terms of computer time.

6. REFERENCES

[1] Millman J and Grabel A 1987 *Microelectronics*, McGraw-Hill

[2] Jeffreys H 1931 *Scientific Inference*, Cambridge University Press

[3] Sinclair E 2008 *Volatility Trading*, Wiley

[4] Springer M D 1979 *The Algebra Of Random Variables*, John Wiley & Sons

[5] Theis C V 1941 The effect of a well on the flow of a nearby stream, *American Geophysical Union Transactions* **22** (3) 734-738

[6] Liu J S 2001 *Monte Carlo Strategies in Scientific Computing* New York: Springer

[7] Christen J A and Fox C 2009 A General Purpose Sampling Algorithem for Continuous Distributions (the t-walk), submitted to *Bayesian Analysis*