A Statistical View of Some Aspects of Inverse Problems

Mark Westcott

Abstract

This paper illustrates, mainly by example, the links between some existing inverse problem formulations and methodologies, and statistical formulations of the same problems. The examples are from curve fitting, density estimation, image restoration and emission tomography.

1 Introduction

When Bob Anderssen asked me to speak on a statistical approach to inverse problems, my immediate instinct was to retort '*Statistics* is an inverse problem', and refer him to a worthy text on the subject. However, it is not prudent to treat one's scientific superior with such levity, so I sought a more modest interpretation of his request. I decided to use recent statistical interest in several areas of inverse problems as the framework for some more general observations.

This paper is the result. It does not say anything very original, and draws extensively on several other excellent contributions. I hope that it will illustrate two main themes: (i) how statistical ideas and methods link in with some of the existing formulations of inverse problems; (ii) if there is a natural setting for the problem, which will often include constraints on some elements, this should be exploited in formulation and analysis. The latter point is neatly summed up in the following quotation from Green [11], in a paper on tomographic reconstruction: 'We also take the view that more fundamental than the algorithm for reconstruction is the *principle* underlying that algorithm: what is being estimated, what is the basis for that estimation, and what is the performance of the algorithm with respect to that basis?'

The plan of this paper is to use the general topic of smoothing to introduce the two points mentioned above, and to illustrate each by discussion of a few applications, particularly image restoration and emission tomography.

2 The Smoothing Paradigm

Suppose there is a quantity $\mathbf{f}$ that we wish to estimate from data $\mathbf{x}$. Typically $\mathbf{f}$ will be a function of some sort, but it could also be a vector of constants. We would like the estimate to achieve the usually conflicting aims of:

- keeping faith with the data;
- reducing roughness due to noise.

That is, we want a result that is not manifestly perverse in comparison with the data, but whose features are not swamped by data variability.

Titterington [26] introduces a compromise criterion to quantify this idea. He defines the form

$$\Delta_1(\mathbf{f}, \hat{\mathbf{f}}_D) + h\Delta_2(\mathbf{f}, \hat{\mathbf{f}}_u), \tag{1}$$

where $\hat{\mathbf{f}}_D$ is a (rough) data-generated estimate and $\hat{\mathbf{f}}_u$ is an ultrasmooth 'benchmark' against which the roughness of an estimator can be judged. The $\Delta_i (i = 1, 2)$ are some measures of distance, and $h > 0$ expresses the relative weight we want to give to the two objectives. This leads to a *prescription* $\hat{\mathbf{f}}_{h,D}$ for estimation: choose $\mathbf{f}$ to minimise (1). This prescription is known as a *minimum penalized distance* (MPD) estimator, for obvious reasons.

The choice of h is usually made by invoking another optimality criterion such as minimizing the expected risk or predictive risk of the prescription $\hat{\mathbf{f}}_{h,D}$. The latter choice leads to the appealing idea of *cross-validation* (see the paper by Lukas in this volume). Occasionally there will be a well-defined prior choice for h; see [26], §3.1.

There is a full discussion in [26] of the various forms used for Δ_1 and Δ_2 in practice. We shall only look at a specific example.

Example: Nonparametric regression (Curve Fitting)

Here the model is

$$\mathbf{x} = \mathbf{f} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{f} = (f(t_i))^T$, $f(\cdot)$ the function of interest and $\boldsymbol{\epsilon}$ is a vector of noise components. Typically the choices of Δ_1, Δ_2 are (Silverman [23], §2),

$$\Delta_1 = \|\cdot\|_{L^2} \equiv \Sigma(\cdot)_i^2; \ \Delta_2 = \int \{f''(x)\}^2 dx, \tag{3}$$

so the benchmark of smoothness, $\hat{\mathbf{f}}_u$, is implicitly linear. In this case, it is well-known that $\hat{\mathbf{f}}_{h,D}$ is a cubic spline (Schoenberg [21]; see [23], §2).

The problem can be brought into the orbit of familiar statistical practice by parameterizing $f(\cdot)$. One way of doing this either exactly (Silverman [23], §6.1) or approximately (O'Sullivan [19], §3) is to write $f(\cdot)$ as a linear combination of basis splines. In either case, (1) becomes, from (3),

$$\sum_{i=1}^{n}\{x_i - (A\mathbf{b})_i\}^2 + h\mathbf{b}'C\mathbf{b}, \tag{4}$$

where $\mathbf{b}$ is a parameter vector and A, C are appropriate matrices. Calculation of $\hat{\mathbf{b}}$, and hence $\hat{\mathbf{f}}_{h,D}$, now looks like a least squares problem; an efficient, i.e. $O(n)$, algorithm is given in Hutchinson and de Hoog [14].

The point about (4) is that, to a statistician, it can arise as follows. There is a well-known procedure for combining information about a model from data and from prior knowledge, called *Bayes Theorem*. Informally, it can be expressed as

$$\text{prob(model} \mid \text{data)} \;\propto\; \text{prob(data} \mid \text{model)} \cdot \text{prob(model)}; \tag{5}$$

here prob$(A|B)$ means the probability of A given B. The first term on the RHS of (5) is usually called the *likelihood*, the second is the *prior*. Leaving aside the controversies associated with prior probabilities, we note that, on taking logs in (5), we obtain (1) with $\Delta_1 = \log(\text{ likelihood })$ and $\Delta_2 = \log(\text{ prior })$. The MPD prescription $\hat{\mathbf{f}}_{h,D}$ is now the model that is 'most likely, given the data', an intuitively reasonable choice. In this context, $\hat{\mathbf{f}}_{h,D}$ is often called a *maximum a posteriori* (MAP) estimator instead of MPD.

Once this link between (1) and (5) has been forged, it gives us a possible way of choosing appropriate Δ_1, Δ_2 in (1); if there is a natural *statistical* framework for the problem then expressing it in the form (5) provides the Δ_i rather straightforwardly. For instance, in the example above, (4) is statistically reasonable if:

- the errors ϵ_i are *independent normal* variables; *and*

- the parameters $\mathbf{b}$ have a prior *normal* distribution.

Should either of these assumptions be untenable, the standard method based on (4) might have to be rethought; an example is when the ϵ are not independent or cannot be assumed, or transformed, to be normally distributed.

In summary, Bayes Theorem provides a natural way of incorporating relevant statistical knowledge into the smoothing algorithm. It will be further illustrated in Section 4.

3 Natural Constraints

In many applications, there will be natural constraints on some elements of the problem which it is desirable to preserve in the estimation procedure. For example, we might have $f(\cdot) \geq 0$ in (2). This could be included *via* a constrained optimization (Wahba [29]), but it will often be better to build it explicitly into the estimation procedure. Examples can be found in Titterington [26]; see his §2.1 for probability estimates that are nonnegative and sum to 1, and §6.3 for nonnegativity imposed by choosing Δ_2 as an entropy function. We shall discuss a slightly different case.

Example. Nonparametric density estimation.

Here $f(\cdot)$ is a probability density, so $f(\cdot) \geq 0$ and $\int f(x)dx = 1$. The data $\mathbf{x}$ are independent observations from $f(\cdot)$. The most common form of estimator for $f(\cdot)$ is a kernel-based one, namely

$$\hat{f}_{h,D}(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - x_i)/h\}. \tag{6}$$

Although this cannot (as far as I know) be derived from any criterion like (1), there is still an $\hat{f}_D$ (when $h = 0$ i.e. $K(\cdot)$ is a δ-function) and an $\hat{f}_u$ (when $h \to \infty$, effectively a constant), and h is again a trade-off, between bias (fidelity to the data)

and variance (roughness). All this can be found in [26], §4, or in Silverman [24]. The point we wish to make here is that if $K(\cdot)$ is itself a probability density, then $\hat{f}_{h,D}$ will have the natural constraints built in from the start. For this reason, such a choice is nearly universal ([24], p.17).

A more extensive example of incorporating constraints is given in Section 5.

4 Image Restoration

In a number of applications in remote sensing, data are typically collected on a two-dimensional rectangular array of sites, called *pixels*, one or more observations being made on each pixel. A well-known example is LANDSAT satellite data, consisting of readings in four spectral bands from pixels of about 80m^2. The data $\mathbf{x}$ are a distorted form of a true but unknown image, or scene, $\mathbf{f}$, which we wish to reconstruct. Typically, the amount of data is very large.

There is a variety of possible models for $\mathbf{f}$ and for the distortion mechanism. One common assumption is that imperfect resolution has blurred $\mathbf{f}$, so that x_i, for pixel i, is some combination of f_i and neighbouring f_j's; usually the combination is weighted linear. On top of this, there is *additive* noise. The case when $\mathbf{f}$ is also modelled stochastically is discussed in detail by Hall and Titterington [13] as a smoothing problem. We shall use a different example, based on work of Geman and Geman [7] and Besag [2]. For a review, see Ripley [20].

The simplest version of their model has the following assumptions:

(a) Given $\mathbf{f}$, the x_i are independent and depend only on f_i;

(b) There is some prior probability model for $\mathbf{f}$.

Assumption (a) leads to a product form for the likelihood of Π_i prob$(x_i|f_i)$. The usual model in (b) is that $\mathbf{f}$ is a realisation of a *Markov random field* (MRF). This

means that the f_i are assumed to have only local dependence *via* some neighbour structure (see [2], §2.2). A (nonobvious) consequence of this assumption (see, for instance, Isham [15]) is that the prior probability takes the form

$$\text{prob}(\mathbf{f}) = Z^{-1} \exp\{-\beta \sum_{i \sim j} \phi(f_i - f_j)\}, \tag{7}$$

a so-called *nearest neighbour Gibbs distribution.* Here, $\phi(\cdot)$ is a so–called *potential function*, β is a positive constant, and Z is a (typically revolting) normalizing constant. The sum in (7) is taken only over *neighbours* i, j, so the j might be the indices of the four nearest pixels to i.

Referring back to (5) and (1), we can see clearly how the statistical assumptions have determined appropriate forms for Δ_1 and Δ_2. In the regularization context, $\phi(\cdot)$ will be some measure of how smooth we believe $\mathbf{f}$ is initially, while $h = \beta$, so β implies how much weight we give to the prior information relative to the data. This duality of roles is rather satisfying.

The main purpose of the example, to demonstrate a statistical basis for (1) in an application, has now been achieved. However, a few further comments are in order.

In typical applications, neighbouring pixels are likely to have similar values *a priori*; think of land types in the satellite imaging context. In such cases, we would like any prior distribution to incorporate this property. It is intuitively clear that the choice of a MRF prior with $\phi(\cdot)$ in (7) being negative for small arguments (an *attractive potential*) makes realizations $\mathbf{f}$ with like–valued neighbours more probable. Nonetheless, the choice of a MRF for the prior has a considerable element of mathematical convenience, especially as the qualitative properties of a MRF are not usually at all obvious on a larger scale. So it will usually not be easy to say whether a particular choice in (7) will produce realizations that 'look like' any preconceived notions we may have for $\mathbf{f}$. Besag ([2], §2.5) gives a restoration procedure which relies only on the local behaviour of the prior, which is algorithmically, if perhaps not conceptually, satisfactory.

One feature which a MRF like (7) would not include is edges (e.g. roadways in a satellite picture). There are generalizations of the MRF idea that try to incorporate such features (Geman and Geman [7], Brown & Chui [3]). A related approach is to build a very detailed probability model of the class of possible $\mathbf{f}$, assuming of course that the class is susceptible of such precise definition. A famous example is the HANDS project at Brown University (Chow et al [5]).

Estimation of $\mathbf{f}$ by maximising (5), i.e. MAP, is available in principal but very difficult to achieve in practice. A technique called *simulated annealing* has been suggested for this purpose: see [2], §2.3; [7]. The technique is in fact inspired precisely by the probability interpretation of (1) in this context. However, some evidence from the binary case, where the MAP estimate can be calculated by other means, suggests that in practice it is not likely to work very well (Greig et al. [12]). Very recently, Green [10] has provided a new approach based on the EM algorithm (see Section 5). We have already mentioned Besag's alternative procedure, originally proposed as an approximation to MAP.

Finally, assumption (a) can be generalized to include x_i being determined by neighbours of f_i also, and the x_i being dependent ([2], §5.2, 3; Campbell and Kiiveri [4]).

5 Emission Tomography

Emission tomography is one of a range of medical imaging techniques that attempt to reconstruct some feature inside the body from observations outside, a classic indirect problem. Its characteristic target is metabolic activity, which is studied by introducing a radioactive source, detecting emissions and hence building up a picture of the emission intensities inside the body. By contrast, transmission methods such as CAT scans send particles through the body.

There are two principal forms of emission tomography, SPECT (single photon)

and PET (positron). We refer to Shepp and Vardi [22], or the papers by Natterer and Monk in this volume, for background information. Their formal structure is very similar. Particles are emitted in a Poisson process at rate $f(\mathbf{y})$ from a point $\mathbf{y}$ (inside the body). There is a probability $K(\mathbf{x}, \mathbf{y})$, independently for each particle, of being picked up by a detector at $\mathbf{x}$ (outside the body). The rate of detection at $\mathbf{x}$ is therefore

$$g(\mathbf{x}) = \int K(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}. \tag{8}$$

This relation would be fairly generally true for any emission process. The crucial extra property induced by the very plausible Poisson assumption for emissions is that the counts at each detector $\mathbf{x}$ are *independent Poisson* variables, means $g(\mathbf{x})$. Thus we have a problem with much more structure than just the solution of a Fredholm integral equation (8) using data on $g(\cdot)$; obviously $f \geq 0$ and the statistical properties of the data are clear-cut. It seems persuasive that any solution should contain and exploit this structure.

For future discussion, we shall assume that K is known. This ignores the very real practical difficulties caused by attenuation and scattering of emissions, among other things; see Vardi et al. [28] and Kak and Slaney [16], Ch 4.

One common method of estimating f from (8) is to discretize the problem, and then treat (8) as a parameterized form of (2) (cf. (4)). The discussion in Section 2 shows that this approach is not a natural consequence of the above structure. For further comments in the tomography context see Titterington [27] and the enlightening dialogue between Herman et al. (discussion of [28]) and the authors in their reply.

Another method is to adapt the convolution backprojection procedure of transmission tomography ([16], Ch 3), which reduces (8) to an integral of f along a line through detector $\mathbf{x}$. This takes no account of the different physics and detection apparatus; also, the counts in emission tomography are typically orders of magnitude smaller so that proper treatment of random fluctuations becomes more

important. For examples, see [22] (Figs 5, 7, 8, 9) and [28], §2.3.1.

Because the detector counts are independent Poisson variables, the statistical technique known as *maximum likelihood* has obvious appeal. This means maximizing the likelihood part of (5) with respect to f. In practice, the model will be discretized, so we have counts n_t in detector t ($t = 1, \ldots, T$), emission rate f_s in body pixel s ($s = 1, \ldots, S$), and the n_t are independent Poisson variables with means $\Sigma_s p_{s,t} f_s$, $p_{s,t}$ being a (known) discretized version of K such that $\Sigma_t p_{s,t} = 1$. We wish to determine the f_s from the n_t by maximum likelihood (ML).

One possible route to the solution is the EM algorithm (Dempster et al. [6]), usually thought of as a method for 'maximum likelihood with incomplete data'. *If* we had $\nu_{s,t}$, the count from source s that finished up in detector t, the problem would be trivial; the $\nu_{s,t}$ are independent Poisson variables with means $p_{s,t} f_s$, so $\hat{f}_s = \Sigma_t \nu_{s,t}$. We actually have $n_t = \sum_s \nu_{s,t}$, so the ν's are the missing data. The EM algorithm proceeds iteratively:

- *given* $\hat{f}_s^{(i-1)}$ (current estimate of f_s) and n_t, *estimate* $\nu_{s,t}$ (E step)
- use the *estimated* complete data to get a new estimate $\hat{f}_s^{(i)}$ by maximization (M step).

As already indicated, the M step here is trivial. The E step is also easy, thanks to another useful property of Poisson variables; given the value of the *sum* of independent Poisson (λ_s) variables, N say, the *expected* contribution of component s to the sum is $N\lambda_s / \Sigma_s \lambda_s$. That is, we 'backproject' the total data in a weighted fashion! Consequently, ([25], [28]),

$$\hat{\nu}_{s,t} = n_t p_{s,t} \hat{f}_s^{(i-1)} / \Sigma_s p_{s,t} \hat{f}_s^{(i-1)}, \tag{9}$$

and

$$\hat{f}_s^{(i)} = \Sigma_t \hat{\nu}_{s,t} = \hat{f}_s^{(i-1)} \cdot \Sigma_t p_{s,t} n_t / \Sigma_s p_{s,t} \hat{f}_s^{(i-1)}. \tag{10}$$

The EM method is not the only way to ML estimates, but it has the attractive features of ensuring the $\hat{f}_s$ are nonnegative at each step (provided the initial values are) and of always increasing the likelihood at each step ([6], Theorem 1). So, as the problem is concave ([22]), convergence is guaranteed.

This is by no means the end of the story, though we shall only sketch the rest. If $S > T$, the problem is clearly ill-posed; there is no *unique* ML esimate. In any case, ML solutions are very rough. In Silverman et al. [25], an EMS algorithm is proposed, which includes a simple local *smoothing* of the $\hat{f}_s^{(i)}$ at each step, and this appears to converge. Nychka [17] relates this to penalized likelihood estimation (this is (5) with perhaps a heuristic penalty function replacing the prior; see [25], §2.3; [26], §2.2). Green [10] has suggested how to make this computationally feasible, and applied it [11] specifically to emission tomography with a Markov random field as a prior distribution for the $\mathbf{f}_s$, an idea already suggested by Geman and McClure [8], [9]. This links in with the previous section, and reinforces the relation between prior information and regularization.

6 Conclusion

It must be stressed that this review is very fragmentary. It has attempted to illustrate the two themes mentioned in the Introduction, but has inevitably omitted much of the detail of each example. I hope that there are enough references for the interested reader to track down these details, and to appreciate how much extra care and thought is needed to implement the methods in any particular case.

I conclude with two final applications. One is to stereology, specifically the problem of obtaining information from planar cross-sections; this may be a more familiar problem to readers, and is worth a mention because of Bob Anderssen's association with it! See [1]. The basic problem is similar to recovering f from (8). It is another good instance of the alternative approaches *via* regularization (Nychka

et al.[18]; [19]) and probability modelling ([25], §3). The second is more speculative, but I believe that the diffusion tomography problem described by Latham in this volume can also be usefully treated by explicitly exploiting its probability structure.

References

1. Anderssen, R.S. and de Hoog, F.R. (1990) Abel integral equations. In Numerical Solution of Integral Equations, Ed. M.A. Goldberg. Plenum, New York.

2. Besag, J.E. (1986) On the statistical analysis of dirty pictures (with discussion) J.R. Statist. Soc. B, **48**, 259-302.

3. Brown, T.C. and Chui, J. (1990) Image restoration and edge processes. Paper to 10th Aust. Statist. Conference, Sydney.

4. Campbell, N.A. and Kiiveri, H.T. (1990) Neighbour relations and remotely sensed data. Preprint.

5. Chow, Y, Grenander, U and Keenan, D.M. (1990) Hands: A Pattern Theoretic Study of Biological Shapes. Brown University Report.

6. Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977) Maximum likelihood from incomplete data via the EM algorithm (with discussion). J.R. Statist. Soc. B, **39**, 1-38.

7. Geman, S and Geman, D (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Pattn. Anal. Mach. Intell. **6**, 721-741.

8. Geman, S and McClure, D.E. (1985) Bayesian image analysis: an application to single photon emission tomography. Proc. Am. Statist. Ass. Statist. Comput. Sect. 12-18.

9. Geman, S. and McClure, D.E. (1987) Statistical methods for tomographic image reconstruction. Bull. Int. Statist. Inst. LII-4, 5-21.

10. Green, P.J. (1990) On use of the EM algorithm for penalized likelihood estimation. Preprint.

11. Green, P.J. (1990) Bayesian reconstruction from emission tomography data using a modified EM algorithm. Preprint.

12. Greig, D.M., Porteous, B.T. and Seheult, A.M. (1989) Exact maximum *a posteriori* estimation for binary images. J.R. Statist. Soc. B, **51**, 271-280.

13. Hall, P. and Titterington, D.M. (1986) On some smoothing techniques used in image restoration. J.R. Statist. Soc B, **48**, 330-342.

14. Hutchinson, M.F. and de Hoog, F.R. (1985) Smoothing noisy data with spline functions. Numer. Math. **47**, 99-106.

15. Isham, V. (1981). An introduction to spatial point processes and Markov random fields. Int. Statist. Review **49**, 21-43.

16. Kak, A.C. and Slaney, M. (1988) Principles of Computerized Tomographic Imaging. IEEE Press, New York.

17. Nychka, D. (1990) Some properties of adding a smoothing step to the EM algorithm. Statist. Prob. Letter. **9**, 187-193.

18. Nychka, D., Wahba, G., Goldfarb, S. and Pugh, T. (1984) Cross-validated spline methods for the estimation of three-dimensional tumour size distributions from observations on two-dimensional cross sections. J. Amer. Statist. Assoc. **79**, 832-846.

19. O'Sullivan, F. (1986) A statistical perspective on ill-posed inverse problems (with comments). Statist. Sci. **1**, 502-527.

20. Ripley, B.D. (1986) Statistics, images and pattern recognition. Canad. J. Statist **14**, 83-111.

21. Schoenberg, I.J. (1964) Spline functions and the problem of graduation. Proc. Nat. Acad. Sci. USA **52**, 947-950.

22. Shepp, L.A. and Vardi, Y. (1982) Maximum likelihood reconstruction for emission tomography. IEEE Trans. Med. Imaging **1**, 113-122.

23. Silverman, B.W. (1985) Some aspects of spline smoothing to non-parametric regression curve fitting (with discussion). J.R. Statist. Soc. B, **47**, 1-52.

24. Silverman, B.W. (1986) Density Estimation for Statistics and Data Analysis. Chapman and Hall, London.

25. Silverman, B.W., Jones, M.C., Wilson, J.D. and Nychka, D.W. (1990) A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography (with discussion). J.R. Statist. Soc. B, **52**, to appear.

26. Titterington, D.M. (1985) Common structure of smoothing techniques in statistics. Int. Statist. Review. **53**, 141-170.

27. Titterington, D.M. (1987) On the iterative image space reconstruction algorithm for ECT. IEEE Trans. Med. Imaging **6**, 52-56.

28. Vardi, Y., Shepp, L.A., and Kaufman, L. (1985) A statistical model for position emission tomography (with comments). J. Amer. Statist. Assoc. **80**, 8-37.

29. Wahba, G. (1982). Constrained regularization for ill posed linear operator equations with applications in meteorology and medicine. In Statistical Theory and Related Topics III 2, Ed. S.S. Gupta and J.O. Berger, pp 383-418. Academic Press, New York.

CSIRO Division of Mathematics and Statistics
GPO Box 1965, Canberra
ACT–2601, Australia