Constrained Maximum Entropy Methods in an Image Reconstruction Problem

Andrew Gelman*
Department of Statistics
Harvard University
Cambridge, Massachusetts 02138
U.S.A.

Abstract

Maximum entropy and maximum likelihood methods are compared for a simplified version of a medical imaging problem. Iterative reconstructions are tracked by plotting successive values of log-likelihood and entropy, and we find a tradeoff between these two measures of fit. Maximum likelihood is found to fit the data more closely, but maximum entropy creates more reasonable images. We conclude that the former uses the data efficiently, but the latter gives a better choice of image. This reasoning leads to a somewhat Bayesian version of the constrained maximum entropy method of Gull and Daniell (1978). The constraint of that method is interpreted from a Bayesian perspective.

1 Background and setting up the problem

This paper discusses image reconstruction from incomplete, noisy data. Our main example is a simplified model of positron emission tomography (Vardi et al. (1985)). We consider this problem on a theoretical level only, and the brief description which follows may be thought of as motivation for our study. In emission tomography, the image x of interest is an intensity function of radioactive emissions from a two-dimensional region in the human brain. We cannot directly observe x on a live person, but we can count the emissions that leave the brain, and observe their direction. These indirect observations come in the form of a finite set of counts, labeled $y = (y_1, \ldots, y_n)$, in n pairs of radiation detectors outside the brain. (Note: all vectors in this paper are column vectors.) The assumed probability model is:

$$\begin{aligned} y_i &\sim \text{independent Poisson } (M_i),\ i = 1, \ldots, n, \\ M &= (M_1, \ldots, M_n) \\ &= Ax. \end{aligned}$$

*Much of this paper derives from helpful comments by Donald Rubin and Stephen Ansolabehere. This research was supported by a U.S. National Science Foundation graduate fellowship.

J. Skilling (ed.), Maximum Entropy and Bayesian Methods, 429–435.

The expectations M_i (the 'mock data' of Skilling (1986)) are derived from x by a linear transformation A of conditional probabilities. To make the problem tractable, the image x is defined on the discrete space of a grid of N picture elements or 'pixels'. The image is then a vector $x = (x_1, \ldots, x_N)$ of nonnegative elements, and the linear transformation A can be identified with an matrix of rank n, each of whose columns sum to 1. (None of the entries of A will be negative.) The parameter N is chosen by the analyst; to avoid major discretization errors, we will typically assume $N > n$. Note, however, that care is then required in picking reasonable images from a large N-dimensional space.

For this problem, the likelihood is $f(y|M) \propto \prod_i M_i^{y_i} e^{-M_i}$, and we define

$$\begin{aligned} \text{-2LL}(M|y) &= -2\log f(y|M) + \text{ arbitrary constant} \\ &= -2\sum_i \left[y_i \log\left(\frac{M_i}{y_i}\right) + y_i - M_i \right]. \end{aligned}$$

(This corresponds to the 'chisquared' statistic of Skilling (1986).) In passing from the first line to the second, we have set the constant so that -2LL$(M|y) = 0$ at the maximum, when $M = y$.

In general, the entropy of a vector $a = (a_1, \ldots, a_K)$, relative to a measure $b = (b_1, \ldots, b_K)$, is defined as:

$$S(a|b) = -\sum_k \left(\frac{a_k}{\sum a_j}\right) \log\left(\frac{a_k/\sum a_j}{b_k/\sum b_j}\right).$$

2 Comparing maximum entropy and maximum likelihood estimators

We will consider two simple estimators $\hat{x}$ of x. In both cases, we define the estimated sampling expectations $\hat{M} = A\hat{x}$. First, the constrained maximum entropy estimate of Gull and Daniell (1978) and Skilling (1986) is the $\hat{x}$ that maximizes $S(\hat{x}|m)$, subject to the constraint: -2LL$(\hat{M}|y) \leq n$. (We will define the entropy relative to the uniform measure: $m_j = 1$, for all j.) If the constraint on -2LL cannot be satisfied, then the maximum likelihood estimate (defined below) will be labeled as 'constrained maximum entropy', too.

Second, the maximum likelihood estimate is an nonnegative image $\hat{x}$ that minimizes -2LL$(\hat{M}|y)$. The estimate $\hat{x}$ will be unique, except when the absolute minimum, -2LL$(\hat{M}|y) = 0$ (that is, $A\hat{x} = y$) can be achieved. In this case, we choose, as a unique 'maximum likelihood estimate', the $\hat{x}$ that maximizes the entropy $S(\hat{x}|m)$, subject to the constraint: -2LL$(\hat{M}|y) = 0$.

Conditional on the true image x, an estimate $\hat{x}$ is a function of the random variable y. Rather than examine an $\hat{x}$ directly, we look at its fit to the prior measure m, observations y, true image x, and true sampling expectations M. These four summary comparisons are: $S(\hat{x}|m)$, -2LL$(\hat{M}|y)$, $S(\hat{x}|x)$, and $S(\hat{M}|M)$, respectively. We are interested in the expectations of these quantities, averaged over the sampling distribution of y. For fixed dimensions n and N, a fixed transition matrix A, and a fixed true image x, we can simulate independent data sets y (from the appropriate Poisson distributions). Given n, N, A, and y, a computer program finds the constrained maximum entropy and maximum likelihood estimates of x. For each estimator, the program then calculates the average values of the

Table 1: Approximate sampling expectations of various functions of two estimators $\hat{x}$ of the image x

True image	Dimension of data vector	Reconstruction grid	Estimator	Fit to prior measure: $-S(\hat{x}\|m)$	Fit to data: - 2LL$(\hat{M}\|y)$	Fit to true image: $-S(\hat{x}\|x)$	Fit to truth in data space: $-S(\hat{M}\|M)$
	n=6	4 × 4	max-ent	.13	6.0	.10	.0064
			m.l.e.	.38	0.0	.15	.0050
		8 × 8	max-ent	.10	6.0	.16	.0061
			m.l.e.	.29	0.0	.19	.0050
20 20 20 20	n=12	4 × 4	max-ent	.08	12.0	.12	.0050
20 100 100 20			m.l.e.	.89	4.1	.51	.0034
20 100 100 20		8 × 8	max-ent	.06	12.0	.17	.0053
20 20 20 20			m.l.e.	1.43	3.4	1.11	.0037
	n=24	4 × 4	max-ent	.16	24.0	.07	.0115
			m.l.e.	.57	9.3	.28	.0099
		8 × 8	max-ent	.13	24.0	.12	.0117
			m.l.e.	1.60	2.2	1.35	.0132
	n=6	4 × 4	max-ent	.07	6.0	.28	.0067
			m.l.e.	.30	0.0	.38	.0048
		8 × 8	max-ent	.06	6.0	.29	.0066
			m.l.e.	.23	0.0	.38	.0048
20 20 20 20	n=12	4 × 4	max-ent	.19	12.1	.36	.0064
20 200 20 20			m.l.e.	1.21	5.9	.89	.0055
20 20 20 20		8 × 8	max-ent	.27	12.1	.49	.0066
20 20 20 20			m.l.e.	2.37	5.3	2.18	.0057
	n=24	4 × 4	max-ent	.21	24.0	.19	.0299
			m.l.e.	.82	11.6	.69	.0411
		8 × 8	max-ent	.14	24.0	.21	.0133
			m.l.e.	1.69	2.2	1.24	.0152

four comparisons described above, over 20 simulations of y. For this paper, we did the above computation for 12 cases: 2 true images $\hat{x}$ (each defined on a 4 × 4 grid); 3 sets of n and A; and 2 reconstruction grids (4 × 4 and 8 × 8; that is, $N = 16$ and $N = 64$). The results are shown in Table 1.

3 Tradeoff between likelihood and entropy

Table 1 shows that maximum likelihood better fits the true M, as well as, of course, the data y. Constrained maximum entropy better fits the true x, as well as, of course, the prior measure m. These results imply a tradeoff between fit in data space and fit in image space, with constrained maximum entropy performing better in the key measure of fit to the true image. Looking at the results more closely, we also find that maximum likelihood does reasonably well when it fits the data exactly, and worse when it cannot.

Both methods of course fit the data or prior model better when they estimate over a finer grid; at the same time, they fit the truth less well. This makes sense in our example, because we defined the true image over the coarse grid. The constrained maximum entropy reconstructions are only slightly worse in the fine grid, however, while some maximum likelihood images fit far worse when allowed these extra degrees of freedom.

The maximum likelihood estimate (when there is no perfect fit) is found by EM (Vardi et al. (1985)). This is an iterative algorithm, each step of which increases the likelihood of the estimate (Dempster et al. (1977)). We can track the entropy and likelihood of the EM iterates, starting at a uniform image (thus moving from maximum entropy to maximum likelihood). We have examined two such plots: one that converges to an image for which $\hat{M} = y$ (and so -2LL($\hat{M}|y$) = 0) and one for which no such image exists. Interestingly, $S(\hat{x}|m)$ decreases in each step of the algorithm, in both cases. These plots imply a tradeoff, in models, between entropy and likelihood, especially in the region near maximum likelihood, where entropy shows a great decrease. For these same iterative estimates, we have also plotted their fit $S(\hat{x}|x)$ to the true image and the corresponding fit $S(\hat{M}|M)$ to the truth in data space. Here we find that in the first few iterations, both measures of fit improve. However, as the algorithm approaches the maximum likelihood estimate, the fit in data space gets slightly worse, and the fit in image space gets much worse. This is apparently due to the spiky character of the maximum likelihood estimates and holds even in a case of a very spiky true image.[1]

4 Rationale for constrained maximum entropy

To understand this apparent tradeoff, we must explore the link between a model in image space and the data in their space. We are interested in the image x, but the data tell us only about the sampling expectations M, and nothing about x, given M. To get an image, we must estimate M from the information provided by the data, and then choose an $\hat{x}$ consistent with our estimate $\hat{M}$. We need models on data space and on image space, given data. If we do not formalize our models, we are using implicit models. Perhaps these can explain the behavior of the methods presented above.

We will embed the parameter M in a Bayesian model, and hence determine its probable values, given the data. Then we will use maximum entropy to select one image among all those consistent with M, for each value in the posterior distribution of M. We do not extend our Bayesian model to image space because, given M, inference on x would depend solely on the prior distribution. It may be more desirable to choose our image-picking criterion as such, rather than to model in the vast space of images. This is the rationale of Skilling (1986).

As mentioned above, the fineness of the reconstruction grid is specified by the analyst; in fact, there is no logical upper bound for the number N of pixels. Aside from computational difficulties, a Bayesian modeler on x may wish to keep N low to moderate the task of specifying a plausible distribution over the space of all images x in N-dimensional space. Maximum entropy appears to solve this problem easily, however. Entropy is invariant under continuous reparameterization; thus, if an image is left unchanged but is pixellized more finely, its entropy (relative to a locally uniform measure) will not change. Furthermore, this identical image has the highest entropy of all images, on the fine grid, that are consistent with the original coarse image. The simulation results presented in Table 1 imply that this invariance works to our advantage, in that the maximum entropy solution performs relatively well over a too-fine grid.

[1] Graphs are available on request.

5 Bayesian maximum entropy methods

This section discusses a maximum entropy reconstruction method based on Bayesian estimation of parameters in data space, and connects it on a theoretical level to the original approach of Gull and Daniell (1978). Our goal is to suggest an improved method, and to clarify the hidden assumptions in the old method. Assume we have a posterior distribution on $(M|y)$. Assign, to every M, the maximum entropy image *max-ent* $[x(M)]$, satisfying $Ax = M$. This yields a probability distribution of images. If we want to pick just one image, we might take $\hat{M}$ to be the posterior mean $E(M|y)$, and pick the image *max-ent* $[x(\hat{M})]$.

Gull and Daniell perform the more (computationally) difficult task of maximizing $S(x|m)$ subject to the nonlinear constraint: -2LL$(M|y) \leq C$. If we wish to follow this route, we might set C to the posterior mean of -2LL, given y. Asymptotically (that is, with A and n fixed, but with more Poisson data), -2LL$(M|y) \sim \chi^2_n$, with mean n. This gives some justification for the usual constraint value $C = n$. From the Bayesian perspective, however, we should consider the posterior distribution of -2LL, conditional on the data y. In a small sample, we would certainly prefer to set $C = E$(-2LL$(M|y)$), rather than $C = n$, for the constraint: -2LL$(M|y) \leq C$. In fact, one may observe data y such that -2LL$(M|y) > n$ for all positive images x.

6 Illustrative examples

This section shows the use of the methods described above as applied to three situations. We start with a simple, straightforward example and move to an approximation of the main example of this paper. The examples in this section will be based on the Normal model:

$$y_i \sim N(M_i, \sigma^2),\ i = 1, \ldots, n.$$

The sampling expectations M will again be expressed as an all-positive linear transformation of an all-positive image:

$$M = Ax,\ \text{with } N \text{ pixels in } x,\ N \geq n.\ A \text{ has rank } n.$$

The fit to the data is then measured by a sum of squares:

$$\text{-2LL}(M|y) = \sum_i (M_i - y_i)^2.$$

The range of the transformation A, applied to the set of nonnegative images x, is a convex region in data space that we will call P. If $y \in P$, then there is an image (in general, an $(N - n)$-dimensional space of images) that fits the data perfectly. We put a uniform prior distribution on M, for all $M \in P$.

In our first example, we set $N = n$ and A to the identity, so $M = x$. The posterior distribution of M is truncated Normal:

$$(M_i|y_i) \sim N(y_i, \sigma^2),\ \text{constrained to } M_i > 0,\ \text{for } i = 1, \ldots, n.$$

If all the observations y_i are appreciably greater than σ, the truncation will be unimportant. In this case,

$$\text{-2LL}(M|y) \sim \chi^2_n.$$

For any specific $\hat{M}$, the only possible image is $\hat{x} = \hat{M}$, and the posterior distribution of maximum entropy estimates is the truncated n-dimensional Normal, centered at y.

Our second example is the same, but with the additional prior restriction that all the M_i's be equal. Thus restricted, the n-dimensional Normal posterior distribution becomes a univariate Normal on the common parameter M_i:

$$\begin{aligned}(M_i|y) &\sim N(\bar{y}, \frac{\sigma^2}{n}), \text{ constrained to } M_i > 0,\\ \text{-2LL}(M|y) &= \frac{n(M_i-\bar{y})^2}{\sigma^2} + \sum_i \frac{(y_i-\bar{y})^2}{\sigma^2}.\end{aligned}$$

Assuming $\frac{y}{\sigma}$ is sufficiently far from 0, the conditional distribution of $(\frac{n(M_i-\bar{y})^2}{\sigma^2}|y)$ is χ^2_1, and -2LL$(M|y)$ is just that random variable plus a constant that is known, given y. Unconditional on the data, this constant has expectation $E(\frac{\sum_i (y_i-\bar{y})^2}{\sigma^2}) = n-1$.

In our third example, $N > n$ and A is a complicated matrix. P is now a convex region in data space bounded by m hyperplanes that intersect the origin. The posterior distribution of M, given y, is truncated Normal once again, but this time the truncation matters. The data y might not lie within P. Also, -2LL$(M|y)$ will no longer be approximately distributed as χ^2_n, and its expectation, given y, will most likely not be close to n. As σ^2 decreases, however, y becomes closer to the true M and less likely to be near the boundary of P. Thus, the truncation becomes less important, and as $\sigma^2 \to 0$, we return to the geometry and distribution of $(M|y)$ of the first example of this section. Of course, for any $\hat{M}$, we must still choose a maximum-entropy image $\hat{x}$ from an $(N-n)$-dimensional space satisfying $A\hat{x} = \hat{M}$. This third example is very similar to the main example of this paper, inasmuch as the Normal distribution approximates the Poisson. The asymptotic case of infinite data corresponds to $\sigma^2 \to 0$.

7 Discussion

This last example allows us to understand the problems of the maximum likelihood reconstruction. If $y \notin P$, this reconstruction will lie on the boundary of P, yielding an image with zeroes in many cells. The chance of this happening depends on how close A is to being singular, as well as on the amount of Poisson data and on the true image; it can happen even with a smooth true image. On the other hand, if $y \in P$, then the maximum likelihood estimate gives $\hat{M} = y$. This may overfit the data in those dimensions in which M is constrained to a narrow region. In such directions, the posterior density of M may be nearly constant. An estimator that fits too closely to the position of y will be subject to the latter's great sampling variability. These problems disappear asymptotically, of course. As the number of counts increases, maximum likelihood becomes as precise as any estimator of M.

Our theoretical study also explains the entropy-likelihood tradeoff in two ways. First, as -2LL is allowed to increase, a larger region in data-space, and thus in image-space,

becomes available in which to search for high-entropy images. Second, if the data y are a small sample, parameters M with likelihoods near the maximum will generally be peculiar points, probably nearer to the boundary of P than any reasonably smooth true image. This is apparently common in practice, to judge from reports of unrealistic maximum likelihood reconstructions for hypothetical and real tomography data (Vardi et al. (1984), Fox (1987)). Such behavior can make the tradeoff more extreme near the maximum of likelihood.

In conclusion, maximum entropy and maximum likelihood estimates for our problem differ; the former better fits the true image and the latter better fits the data. When fit on an overly fine grid of pixels, maximum entropy produces reasonable images; maximum likelihood does not. In light of these results, we suggest attacking our image reconstruction problem with separate analyses on data space and image space. We can first estimate our knowledge of the sampling expectations M, from the noisy data y. For any point in the posterior distribution of M, we can then choose the maximum entropy image x consistent with this incomplete information. A simple example shows the connection between this method and that of Gull and Daniell (1978) and Skilling (1986). We interpret the 'hard constraints' of the latter methods as an approximation to our Bayesian approach.

Bibliography

Dempster, A. P., Laird, N. M., and Rubin, D. B. 'Maximum Likelihood from Incomplete Data via the EM Algorithm'. *J. Royal Stat. Soc.* **B 39**, 1 (1977).

Fox, P. Private communication (1987).

Good, I. J. *The Estimation of Probabilities.* M.I.T. Press (1965).

Gull, S. F., and Daniell, G. J. 'Image Reconstruction from Incomplete and Noisy Data'. *Nature* **272**, 686 (1978).

Skilling, J. 'Theory of Maximum Entropy Image Reconstruction'. In *Maximum Entropy and Bayesian Methods in Applied Statistics*, J. H. Justice, ed. Cambridge U. P. (1986).

Vardi, Y., Shepp, L. A., and Kaufman, L. 'A Statistical Model for Positron Emission Tomography'. *J. Amer. Stat. Assoc.* **80**, 8 (1985).

Fig. 1a: Iterates of the EM algorithm:
fit to prior measure and to data

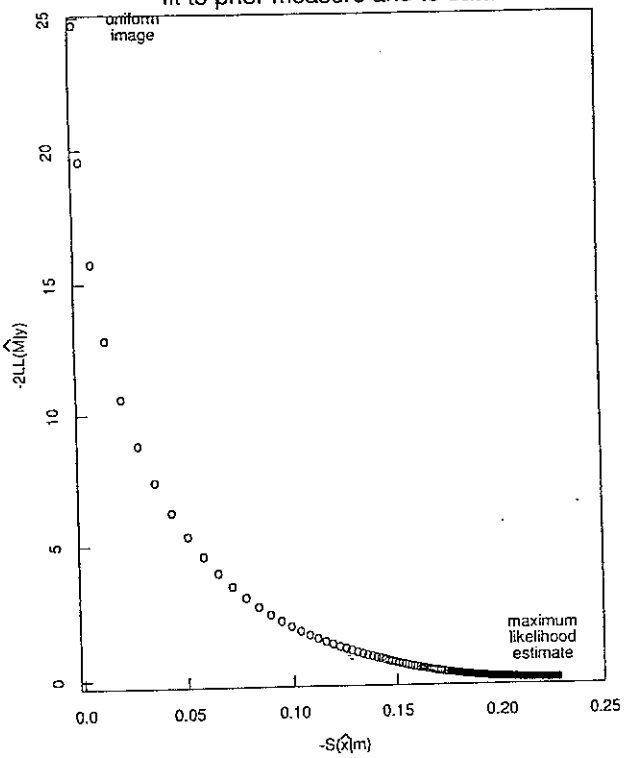

Fig. 1b: Iterates of the EM algorithm:
fit to truth in image and data space

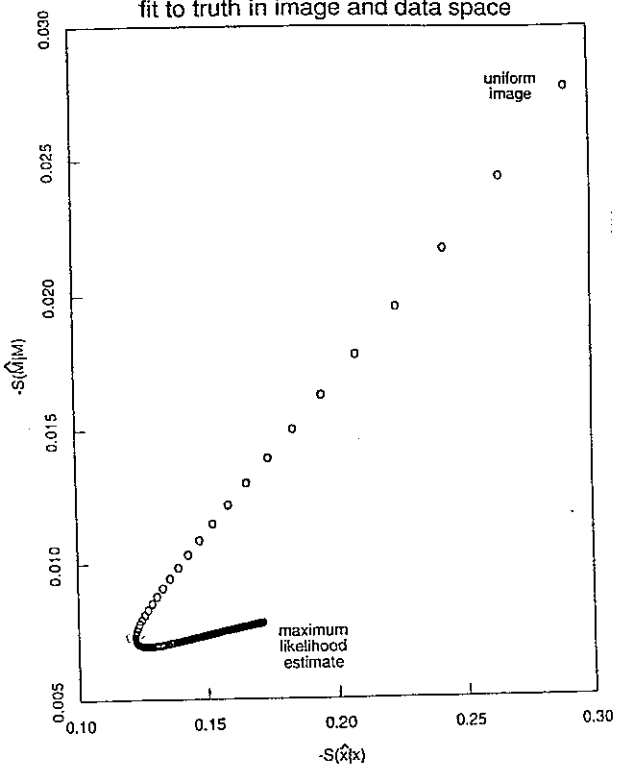

Fig. 2a: Iterates of the EM algorithm:
fit to prior measure and to data

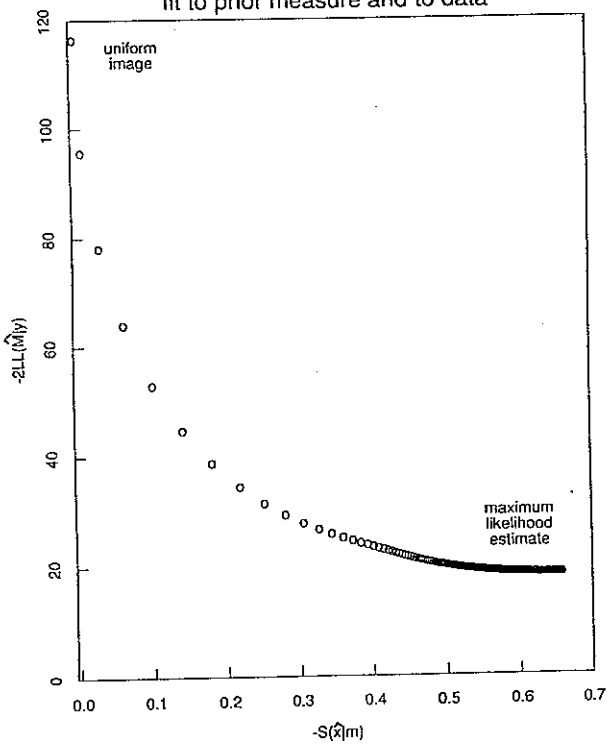

Fig. 2b: Iterates of the EM algorithm:
fit to truth in image and data space