Who Needs Data?

Restricting Image Models by Pure Thought

By

Andrew Gelman*

Technical Report No. 363
July 1992

*Presented at the Cincinnati meeting of the Institute of Mathematical Statistics, March, 1992. Thanks to Xiao-Li Meng, Julian Besag, and Terry Speed for helpful comments and the National Science Foundation for financial support

Department of Statistics
University of California
Berkeley, California 94720

Who Needs Data?
Restricting Image Models by Pure Thought

Andrew Gelman*
Department of Statistics
University of California
Berkeley, CA 94720

July 24, 1992

Abstract

In applications, statistical models for images are often restricted to what produces reasonable estimates for the data at hand. In many cases, the principles that allow a model to be restricted can be derived theoretically, in the absence of any data and with minimal applied context.

We present three theoretical examples. We interpret local smoothing of spatial lattice data as Bayesian estimation and show why uniform local smoothing does not make sense. In time series, we show that an autoregressive model for local averages violates a principle of invariance under scaling. Finally, we show how the Bayesian estimate of a strictly-increasing time series, using a uniform prior distribution, depends on the scale of estimation.

Concerns about the behavior of models and estimates under rescaling are especially important for image analysis. It is desirable for substantive inference about an image to not depend on the (often arbitrary) scale of "pixels," and it is important to know which families of models can be dismissed on theoretical grounds alone.

Keywords: ARMA model, Bayesian statistics, conditional autoregressive model, scaling, spatial statistics, spatial smoothing, time series.

1 Introduction

In the recent statistical literature, a wide variety of models and methods have appeared for the spatial analysis of images (notably Grenander, 1983, Geman and Geman, 1984, Besag, 1986, and Ripley, 1988). The true measure of any statistical method is how it fares in applications; however, we can gain quite a bit of understanding by theoretical analysis.

*Presented at the Cincinnati meeting of the Institute of Mathematical Statistics, March, 1992. Thanks to Xiao-Li Meng, Julian Besag, and Terry Speed for helpful comments and the National Science Foundation for financial support.

In many contexts, statistical models can be constrained by symmetry principles, the simplest being exchangeability among independent samples from a population. For many time series and spatial models, stationarity, or translation-invariance, is a useful assumption, at least before any specialized knowledge is added. To put it another way, a statistician is expected to provide a justification for a model that is *not* translation-invariant. Another useful default assumption in spatial modeling is isotropy, or rotational invariance.

In this article, we discuss ways in which image models can be evaluated, using statistical principles, without seeing any data.[1] Section 2 gives an example of how a particular image smoother can be understood, and its parameters restricted, by exploiting the equivalence between probability and smoothing. In Section 3, we consider the scale invariance of a family of probability models for pixels that are local averages of a continuous image, and Section 4 presents an unexpected example of how the estimate of a continuous function under constraints can depend on the scale of discretization. We conclude in Section 5 with a discussion of the principles by which statistical models can be criticized theoretically, and the implications for imaging.

Our key intellectual lever is the Bayesian approach, which allows us to worry about probability models rather than estimation—mathematics rather than statistics. Bayesian analysis is the simplest way for us to derive the results presented here, but is of course not necessary. Many of the results presented here have been derived in previous statistical work; our contribution is to draw out some principles that can be derived theoretically, even though in the past they may have been presented in detail in the context of specific examples.

[1] We do not go as far as some maximum entropy theorists (e.g., Skilling, 1988), who seek not just to restrict a model class, but actually to specify a model based on theoretical principles only.

2 Spatial smoothing

Our first example shows how one can restrict the parameters of a specific image model without seeing any data.

Consider a two-dimensional image θ, discretized by gray-level intensities in a grid of n square pixels, $\theta = (\theta_1, \ldots, \theta_n)$. Suppose a data vector, $y = (y_1, \ldots, y_n)$, has been observed; to keep things simple, assume independent normal data:

$$y_i \sim N(\theta_i, \sigma^2).$$

In general, one can imagine y observed directly or indirectly; in the latter case, the vector y could be augmented data that are imputed using the EM algorithm (Dempster, Laird, and Rubin, 1977) or the Gibbs sampler (Geman and Geman, 1984). Examples of data augmentation in imaging include Shepp and Vardi (1982) and Geman and McClure (1987).

Given the observations y (directly or indirectly), Silverman et al. (1990) propose a local linear smoothed estimate, in which the estimate $\hat{\theta}$ is the convolution of y with a specified kernel:

$$\hat{\theta} = Sy.$$

Using the notation (s_{ij}) for the elements of the matrix S, the smoother is required to be a weighted average: $\sum_j s_{ij} = 1$ for each i.

Silverman et al. suggest smoothing over a 3×3 grid, with a weighted average of the center and the eight neighbors:

$$s_{ij} = \begin{cases} \frac{W}{W+8} & \text{if } i = j \\ \frac{1}{W+8} & \text{if } i \text{ and } j \text{ are neighbors.} \end{cases}$$

To bend notation slightly, we can write the smoothing kernel S in spatial form as,

$$S = \frac{1}{W+8} \begin{array}{|c|c|c|} \hline 1 & 1 & 1 \\ \hline 1 & W & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array}.$$

Silverman et al. (1990) report success with values of W from 25 to 100. Is advice useful in general, or only for their particular example? If the former, is there a principle that could tell us not to use $W = 1$, say, or $W = 200$? It turns out that, yes, theoretical reasoning alone can make us distrust the smoother with $W = 1$, or, for that matter, $W = 5$.

2.1 Bayesian interpretation

To undertand the smoothing estimate better, we will interpret it as a posterior mean from a Bayesian model. Since $\hat{\theta} = Sy$ is a linear estimate, it corresponds to a Gaussian prior distribution. As discussed by Besag (1974), any multivariate normal distribution can be expressed as a *conditional autoregression*, in which the distribution of each component of θ is expressed conditionally on all the others:

$$\begin{aligned} E(\theta_i|\theta_j, \text{ all } j \neq i) &= E(\theta_i) + \sum_{j \neq i} c_{ij}(\theta_j - E(\theta_j)), \\ \text{var}(\theta_i|\theta_j, \text{ all } j \neq i) &= \tau_i^2, \end{aligned}$$

The joint density of θ under this model is,

$$P(\theta) \propto \exp(-\frac{1}{2}\theta^t \text{diag}(\tau^2)(I - C)\theta),$$

where C is the matrix of coefficients (c_{ij}), with the diagonal elements, c_{ii}, understood to be zero. In addition, the precision matrix, $\text{diag}(\tau^2)(I - C)$, must be symmetric. For simplicity, we will assume that the prior variances are equal: $\tau_i^2 = \tau^2$, for all i. (Besag, York, and Mollie, 1991, discuss the conditional autoregressive model in more detail in an applied context.)

Combining the prior distribution with the likelihood, $(y|\theta) \sim N(\theta, \sigma^2 I)$, yields the following posterior mean:

$$\hat{\theta} = \frac{\tau^2}{\sigma^2 + \tau^2}\left(I - \frac{\sigma^2}{\sigma^2 + \tau^2}C\right)^{-1} y,$$

which corresponds to the linearly smoothed estimate, with a smoother

$$S \propto (I - \lambda C)^{-1}, \tag{1}$$

where $\lambda = \frac{\sigma^2}{\sigma^2+\tau^2}$. Given that the smoothing operator is a weighted average (i.e., the smoothing coefficients sum to 1), the following statements are well known to be true (see Kimmeldorf and Wahba, 1970, and Wahba, 1987, for a general discussion and Besag, 1986, for the image smoothing interpretation):

- The conditional autoregression is intrinsic of order 1, in the sense of Matheron (1973) and Kunsch (1987); that is, $\sum_j c_{ij} = 1$ for all i.
- The model for θ is nonstationary.
- The prior distribution for θ is improper.
- The matrix $(I - C)$ is noninvertable.

Incidentally, the posterior distribution for θ, being a multivariate normal distribution, can itself described as a conditional autoregression, but with new coefficients that do not sum to 1; conditional on data, the distribution is proper.

2.2 The local neighborhood smoother

As discussed by Gelman (1990a, b), the smoother can be approximated using the Taylor expansion of (1):

$$S \propto I + \lambda C + \lambda^2 C^2 + \ldots. \tag{2}$$

To first order, C should have the same neighborhood of S.

Consider the following matrix of autoregression coefficients:

$$c_{ij} = \begin{cases} \frac{1}{8} & \text{if } i \text{ and } j \text{ are orthogonal or diagonal neighbors} \\ 0 & \text{otherwise,} \end{cases}$$

which can be written graphically as,

$$C = \frac{1}{8}\begin{array}{|c|c|c|}\hline 1 & 1 & 1 \\ \hline 1 & 0 & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array}.$$

To get the second order approximation, we compute

$$C^2 = \frac{1}{64}\begin{array}{|c|c|c|c|c|}\hline 1 & 2 & 3 & 2 & 1 \\ \hline 2 & 2 & 4 & 2 & 2 \\ \hline 3 & 4 & 8 & 4 & 3 \\ \hline 2 & 2 & 4 & 2 & 2 \\ \hline 1 & 2 & 3 & 2 & 1 \\ \hline \end{array}.$$

Then,

$$C^2 \approx \frac{1}{8}I + \frac{3}{8}C, \tag{3}$$

and (2) can be approximated as,

$$S \propto \begin{array}{|c|c|c|}\hline 1 & 1 & 1 \\ \hline 1 & W & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array},$$

where

$$W = \frac{8}{\lambda}\left(\frac{1 + \frac{1}{8}\lambda^2}{1 + \frac{3}{8}\lambda}\right). \tag{4}$$

2.3 Interpretation of the smoothing parameter

Different values of the smoothing parameter, W, can be obtained by allowing λ to range from 0 to 1 in equation (4). As $\lambda \to 0$, $W \to \infty$. This makes sense, because if λ is low, the prior variance is high, and the Bayesian estimate will weight the data more highly.

A smoothing parameter of $W = 50$, recommended by Silverman et al. (1990), corresponds to $\lambda \approx \frac{1}{6}$, a fairly weak prior distribution with variance five times the data variance.

The lowest value of W possible in equation (4) is $W = 7$, corresponding to $\lambda = 1$. However, if λ is even close to 1, the second order Taylor expansion for $(I - \lambda C)^{-1}$ will not be accurate, and many more terms will be required. The terms C^3, C^4, and so on, will bleed far beyond the original 3×3 grid, and so the corresponding smoother, S, will no longer be based on the eight nearest neighbors.

Thus, there is a logical basis for considering restricting the parameter W to exceed 10 for the local neighborhood smoother. If one is fitting such a model and a lower smoothing parameter seems warranted, it would probably be better to smooth over a larger neighborhood. Conversely, applying the eight-neighbor smoother with $W = 1$ (a local moving average smoother) corresponds to an ugly conditional autoregression model, with alternately positive and negative coefficients c_{ij} extending far beyond the local neighborhood.

Finally, we can reduce the error in the approximation in equation (3). Using only the eight nearest neighbors, we can make C^2 look most like a linear combination of I and C by setting

$$C = \frac{1}{4+4\sqrt{2}} \begin{array}{|c|c|c|} \hline 1 & \sqrt{2} & 1 \\ \hline \sqrt{2} & 0 & \sqrt{2} \\ \hline 1 & \sqrt{2} & 1 \\ \hline \end{array},$$

which leads to a smoother of the approximate form,

$$S \propto \begin{array}{|c|c|c|} \hline 1 & \sqrt{2} & 1 \\ \hline \sqrt{2} & \mathrm{W} & \sqrt{2} \\ \hline 1 & \sqrt{2} & 1 \\ \hline \end{array}.$$

Such a smoother is a better approximation to the Bayes estimate corresponding to a model with all-positive conditional autoregression coefficients.

2.4 Byproducts of the Bayesian approach

In addition to defining a posterior mean, the Bayesian model also supplies a posterior variance,

$$\begin{aligned} \mathrm{var}(\theta|y) &= \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}\left(I - \frac{\sigma^2}{\sigma^2+\tau^2}C\right) \\ &= \sigma^2 S. \end{aligned}$$

Using the original smoother, the variance for each pixel estimate is,

$$\mathrm{var}(\theta_i|y) \approx \frac{W}{W+8}\sigma^2.$$

Since, as we have already established, W should be larger than 8, the local smoother with eight neighbors will reduce the variance of estimation of the pixel gray levels by less than half.

The Bayesian formulation is also useful in practice, because it treats σ^2 and τ^2 as hyperparameters that can be estimated from data. Thus, W does not need to be pre-specified.

These variance results can also be derived non-Bayesianly (e.g., Henderson, 1950, as discussed by Robinson, 1991), and are presented here just to show that they follow directly from the earlier model specification.

One might also try to estimate W using a non-modeling approach, such as cross-validation: estimate the image using several values of W, then choose the W that minimizes some error measure. Suppose this is done, and the best estimate is a value such as $W = 1$ that is too low (based on the above Bayesian reasoning). In this case, the correct response is probably *not* to use $W = 1$, and not to "artificially" set W to a higher value, but to expand the class of estimators to allow smoothing over a larger neighborhood.

3 Models for averages

Our second example focuses on discrete models for a continuous image, $f(x, y)$, that is approximated by on a discrete set of pixels, with θ_i, the gray level for pixel i, representing the average of the continuous image over the region, R_i, occupied by the pixel:

$$\theta_i = \frac{\int_{R_i} f(x,y)dx\,dy}{\int_{R_i} dx\,dy}.$$

As in the previous example, imagine a fixed set of data (direct or indirect). Ideally, the discretization should not affect our image estimate. Of course, information will be lost if the pixels are too large, but it seems reasonable to demand that after a point, the gross features of the estimated image should not depend on the size of the pixels as they get

smaller. If there is really an underlying continuous image, the discretization should be a convenience, not a fundamental part of the estimate.

It turns out that this requirement is not always satisfied by image models. Here, we will avoid the mathematical difficulty of two-dimensional models and consider a familiar one-dimensional example and ask, does a given probability model for a discretized image make sense?

3.1 Autoregresive models for averages

Consider a continuous time series, $f(t)$, on the real line, divided into intervals of width Δ, parameterized by

$$\theta_i = \frac{1}{\Delta}\int_{i\Delta}^{(i+1)\Delta} f(t)dt,$$

for integer values of i. The simplest conditional autoregressive model is symmetric in the two nearest neighbors, with

$$E(\theta_i|\theta_j, \text{ all } j \neq i) = \frac{\rho}{1+\rho^2}(\theta_{i-1} + \theta_{i+1}), \quad \text{for each } i,$$

where $|\rho| \leq 1$. In one dimension, this is equivalent to the unidirectional AR(1) model with correlation ρ.

Now suppose that $f(t)$ is modeled on a finer scale, with local averages $\phi_1, \phi_2, \ldots$, defined on intervals of width δ; for simplicity, assume that Δ/δ is an integer. If the AR(1) model was reasonable for θ, we should also be willing to apply it to the local averages, ϕ; after all, the original spacing Δ is arbitrary. It would be desirable if aggregating up an AR(1) model on ϕ were to yield an AR(1) model on θ; then we could consider the model on ϕ to be a refinement of the original model on the coarser grid.

Unfortunately, it is well known (see, e.g., Lutkephohl, 1984) that the aggregation of an AR(1) model is not an AR(1) but an ARMA(1,1). In fact, in the limit as $\delta/\Delta \to 0$, an AR(1) model on ϕ implies an MA(1) model on θ. Thus, when fitting an AR(1) model to a

discretized one-dimensional "image," the discretization scale is itself a key parameter and can affect inference about real-world parameters.

3.2 A family of nested models

One solution to the problem of models depending on discretization scale is to just explicitly make the pixel size a parameter in a larger model. We do not like this strategy because it forces discretization to play a double role; we would like to be able to create ever-finer local models without disturbing the large-scale structure.

An alternative approach is to expand the model class so that the models are nested. One solution is a family of AR(1,1) models, which we will parameterize by labeling the correlations as $1, \rho, \eta\rho, \eta^2\rho, \ldots$. If a discretization at scale Δ is modeled as a Gaussian AR(1,1) process with parameters

$$\begin{aligned} \eta &= e^{\Delta/\Delta_0} \\ \rho &= \frac{1}{2}\left(\frac{(1-\eta)^2}{-\log\eta - (1-\eta)}\right), \end{aligned}$$

then the family of models at different scales will be nested, or self-consistent, or closed under scaling: the model on ϕ, based on a scale of δ, averages up to the appropriate model on θ.

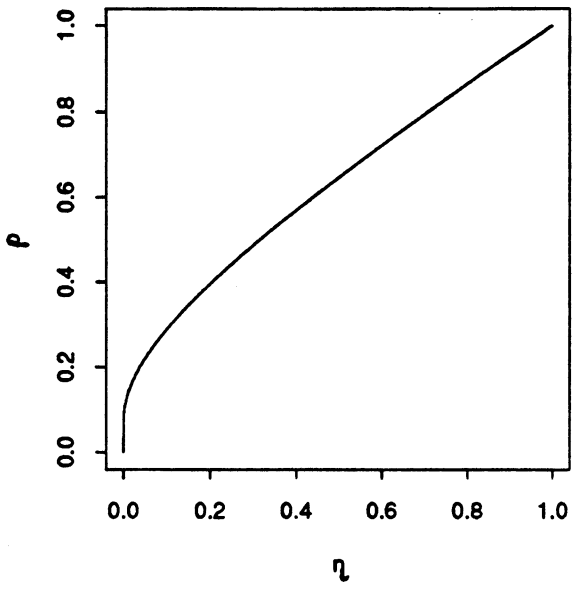
The scale parameter Δ_0 characterizes this family of nested models, which can be derived from a Gaussian stochastic process for the underlying continuous "image," f, with spectral density function

$$f(\omega) \propto \frac{1}{1 + (\omega\Delta_0)^2}.$$

The ARMA(1,1) correlations can be obtained by multiplying the above spectrum with the spectrum of a moving average operator of width Δ, aliasing out wavelengths longer than Δ, and performing the Fourier transform.

As $\Delta \to 0$, the model looks like an AR(1), with $\eta/\rho \to 1$, and the correlation has asymptotic form $\rho \to 1 - \Delta/\Delta_0$. As $\Delta \to \infty$, the model looks like an MA(1), with $\eta/\rho \to 0$, and the correlation has asymptotic form $\rho \to \Delta_0/2\Delta$.

Figure 1: Correlation parameters of the restricted ARMA(1,1) family



There is only one independent correlation parameter for any ARMA(1,1) model in this family—η determines ρ, and vice-versa, as shown in Figure 1. All the parameter values off the curve in the figure—such as AR(1) models with low correlation or MA(1) models with high correlation—are "illegal" under this class of models.

If we were originally planning to fit an AR(1) model to the parameters θ, it seems reasonable now to instead fit a model from the restricted ARMA(1,1) class pictured in Figure 1. What if, however, we were to fit an unrestricted ARMA(1,1) model to a set of data, and found that the data supported parameter values off the "legal" line; e.g., $(\rho, \eta) = (0.9, 0.2)$? Since the unrestricted ARMA(1,1) family is not closed under averaging and rescaling, it would make sense to expand the model class, perhaps to a restricted ARMA(2,2) family, so that the data can be fit without introducing scaling artifacts.

Thus, by allowing an extra parameter, the goodness-of-fit of the whole model class can be tested.

4 Models with constraints

Consider again the estimation of the discretized version of a continuous image. We would like our inferences about the large-scale features of the image to stabilize as the size of the pixels approaches zero. In this section, we focus on the difficulties that arise due to constraints on the continuous model; e.g., the requirement that a gray-level image be all-nonnegative.

4.1 Modeling an increasing time series

Once again, we illustrate the principle involved with a mathematically tractable one-dimensional example. Suppose our unknown "image" is a continuous time series, $f(t)$, *known to be increasing*, and defined on the range $t \in [0,1]$. For simplicity, we assume that $f(0)$ and $f(1)$ are known to be 0 and 1, respectively.

Now suppose we estimate the time series at $n-1$ equally-spaced points: $\theta_1, \ldots, \theta_{n-1}$, where $\theta_i = f(i/n)$. As Bayesians, we will assign the seemingly innocuous uniform prior distribution on the vector $(\theta_1, \ldots, \theta_{n-1})$, so that the mode of the posterior distribution equals the maximum likelihood estimate.

A uniform distribution on the values θ_i, along with the constraint that they are increasing and the known values of $f(0)$ and $f(1)$, is equivalent to a uniform distribution on the simplex: $0 < \theta_1 < \ldots < \theta_{n-1} < 1$. This in turn is equivalent to saying that $\theta_1, \ldots, \theta_{n-1}$ are the order statistics of a sample of size $n-1$ from the uniform distribution on $[0,1]$. In particular, each θ_i has a marginal beta distribution, with variance of order $\frac{1}{n}$.

4.2 Inference for a fixed data set

As $n \to \infty$, the prior distribution becomes ever more concentrated about the straight line $f(t) = t$, the uniform cumulative distribution function. The strength of the prior distribution thus depends on the discretization, with potentially grave consequences.

For example, consider inference from a fixed set of data; e.g., measurements of $f(t)$, observed with error, for several values of t. As n increases, the prior precision increases while the data, of course, stay the same. If we are unfortunate enough to choose an extremely fine scale of estimation, the mass of the posterior distribution will virtually ignore the data. In the limit, all the posterior mass lies on the line, $f(t) = t$. (Interestingly, though, the posterior mode respects the data even as $n \to \infty$. In this case, maximum likelihood is reasonable, but its obvious Bayesian extension is treacherous.)

5 Discussion

The above examples show how a consideration of scaling can help one understand and criticize probability models and statistical procedures. In this discussion, we first consider how the principles of scaling can be formalized, and then comment on the implications for image models.

The purpose of our discussion is not to prove any new results, but to point out some of the subtleties in the theoretical analysis of statistical models—most notably, that a model class may be inconsistent in the prior but not the posterior distribution.

5.1 Definitions of scale invariance

Scale invariance, as applied here, is a subtler principle than translational or rotational invariance. Instead of requiring that a single probability model be invariant under scaling (i.e., self-similarity or fractal behavior), we demand a *family* of models, indexed by scale, that are mutually consistent. It is the family, not any individual model, that should be closed under the scaling operation.

The strongest condition of consistency under scaling is that all discrete models should be derived from a single underlying distribution for the continuous image, with any parameters in the model present in the underlying continuous distribution. In spatial statistics, it is

sometimes easier, and more physically plausible, to construct an underlying continuous model in space-time, with the spatial distribution obtained by averaging over the time parameter (see Whittle, 1962).

Many useful families of image models, such as intrinsic autoregressions, cannot be derived from a continuous spatial model, but can still satisfy the following weaker condition of scaling invariance. Consider an image divided into square pixels of linear dimension Δ. Now model the image using pixels of size Δ/n, and aggregate to obtain a probability distribution on larger grid. The weaker, limiting scale invariance principle states that for any Δ, the distribution obtained by aggregating smaller pixels should approach a limiting distribution (that will be a function of Δ) as $n \to \infty$.

The definitions of scale invariance can be further weakened by considering posterior distributions rather than prior distributions. For example, if the restricted class of ARMA(1,1) models in Section 3 were actually true, then it would be acceptable in practice to estimate the parameters (ρ, η) under the unrestricted model, because with enough data, the parameter estimates would almost certainly fall along the line of "legal" parameter values pictured in Figure 1. In contrast, the uniform model in Section 4 violates posterior as well as prior scale invariance.

For another example, Besag (1991) points out that the long-range dependence of the Ising model disappears in the posterior distribution that is obtained by conditioning on data observed on the lattice.

5.2 Remarks on image models

As mentioned above, a Gaussian model is specified by its spectrum, and a non-Gaussian model can sometimes be expressed as a hidden Markov model determined by an underlying Gaussian image. In either case, discrete models can be created at any scale to approximate the continuous spectrum.

For example, the conditional autoregressive model of Section 2 can be understood through its correlation structure as well as its conditional form. Consider a translation-invariant conditional autoregression on the infinite lattice, with common residual variance $\tau^2 = 1$ and coefficients that can be expressed as $c_{ij} = c_k$, where k is the (two-component vector) difference in the positions of pixels i and j. This model has spectral density

$$f(\omega) \propto \frac{1}{1 - \sum_k c_k \cos(\omega \cdot k)},$$

for $\omega \in (-\pi, \pi] \times (-\pi, \pi]$ (Rosanov, 1968). Calculating correlations (or semivariograms in the nonstationary case) for this model is tricky, and is discussed in detail by Besag (1981) and Kunsch (1987). It is not clear if it is even possible to derive the two-dimensional intrinsic autoregression model of Section 2 as a coarsening of a continuous model.

Another approach that has been suggested is the explicitly hierarchical model, with components at an infintite series of finer scales.

Ultimately, we wish to use procedures which will not self-destruct in practice. An important area for further research is to understand which classes of models and procedures are consistent under scaling when applied to a fixed set of data.

References

Besag, J. (1974). Spatial interactions and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society B* **36**, 192–235.

Besag, J. (1981). On a system of two-dimensional recurrence equations. *Journal of the Royal Statistical Society B* **43**, 302–309.

Besag, J. (1986). On the statistical analysis of dirty pictures (with discussion). *Journal of the Royal Statistical Society B* **48**, 259–302.

Besag, J. (1991). Rejoinder to discussion of Besag, York, and Mollie (1991). *Annals of the Institute of Statistical Mathematics* **43**, 45–59.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society B* **39**, 1–38.

Gelman, A. (1990a). Topics in image reconstruction for emission tomograpy. Ph.D. thesis, Harvard University.

Gelman, A. (1990b). Comment on "A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography," by Silverman et al. *Journal of the Royal Statistical Society B* **52**, 314–315.

Geman, S., and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **6**, 721–741.

Geman, S., and McClure, D. E. (1987). Statistical methods for tomographic image reconstruction. *Proceedings of the ISI Meetings.* Tokyo.

Grenander, U. (1983). Tutorial in pattern theory. Technical Report, Division of Applied Mathematics, Brown University.

Henderson, C. R. (1950). Estimation of genetic parameters (abstract). *Annals of Mathematical Statistics* **21**, 309–310.

Kimeldorf, G. S., and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Annals of Mathematical Statistics* **41**, 495–502.

Kunsch, H. R. (1987). Intrinsic autoregressions and related models on the two-dimensional lattice. *Biometrika* **74**, 517–524.

Lutkepohl, H. (1984). Linear aggregation of vector autoregressive moving average processes. *Economic Letters* **14**, 345–350.

Matheron, G. (1973) The intrinsic random functions and their applications." *Advances in Applied Probability* **5**, 439–468.

Ripley, B. D. (1988). *Statistical Inference for Spatial Processes.* Cambridge University Press.

Robinson, G. K. (1991). That BLUP is a good thing: the estimation of random effects (with discussion). *Statistical Science* **6**, 15–51.

Rosanov, Y. A. (1967). On Gaussian fields with given conditional distributions. *Theory of Probability and its Applications* **12**, 381–391.

Shepp, L. A., and Vardi, Y. (1982). Maximum likelihood reconstruction for emission tomography. *IEEE Transactions on Medical Imaging* **MI-1**, 113–122.

Skilling, J. (1988). The axioms of maximum entropy. In *Maximum-Entropy and Bayesian Methods in Science and Engineering*, ed. G. J. Erickson and C. R. Smith, 173–187. Dordrecht: Kluwer Academic Publishers.

Silverman, B. W., Jones, M. C., Wilson, J. D., and Nychka, D. (1990). A smoothed EM approach to indirect estimation problems, with particular reference to stereology and emission tomography. *Journal of the Royal Statistical Society B* **52**, 271–324.

Spitzer, F. (1964). *Principles of Random Walk.* New York: Springer-Verlag.

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society B* **40**, 364–372.

Whittle, P. (1962). Topographic correlation, power-law covariance functions, and diffusion. *Biometrika 49*, 305–314.