Statistica Sinica **6**(1996), 215-232

BAYESIAN MODEL-BUILDING BY PURE THOUGHT: SOME PRINCIPLES AND EXAMPLES

Andrew Gelman

University of California, Berkeley

Abstract. In applications, statistical models are often restricted to what produces reasonable estimates based on the data at hand. In many cases, however, 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 illustrate this point with three well-known theoretical examples from spatial statistics and time series. First, we show that an autoregressive model for local averages violates a principle of invariance under scaling. Second, we show how the Bayesian estimate of a strictly-increasing time series, using a uniform prior distribution, depends on the scale of estimation. Third, we interpret local smoothing of spatial lattice data as Bayesian estimation and show why uniform local smoothing does not make sense. In various forms, the results presented here have been derived in previous 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.

Key words and phrases: ARMA, Bayesian statistics, conditional autoregression, image, scaling, sieve, spatial smoothing, spatial statistics, time series.

1. Introduction

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

In this article, we discuss other ways in which probability models can be evaluated, using a variety of invariance principles, before seeing any data. (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.) In Section 2, we consider the scale invariance of a family of time series models for pixels that are local averages of a continuous image, and Section 3 presents an unexpected example of how the estimate of a continuous function under constraints can depend critically on the scale of discretization. Section 4 gives an example of how a particular image smoothing estimate can be understood, and its parameters restricted, by exploiting the equivalence between probability and smoothing. We conclude in Section 5 with a discussion of the principles by which time series and spatial models can be criticized theoretically.

Our key intellectual lever is the Bayesian approach, which allows us to worry about probability models rather than techniques of estimation. Bayesian analysis is the simplest way for us to derive the results presented here, but is, of course, not the only approach. In various forms, 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.

2. Models for Averages

For our first example, we consider discrete models for continuous functions. Suppose $f(t)$ is to be approximated on a discrete set of intervals, with θ_i, the time series value at interval i, representing the average of the continuous function over the interval I_i: $\theta_i = \int_{I_i} f(t)dt / \int_{I_i} dt$. Consider the problem of estimating θ, and thus approximating f, from a fixed set of data (direct or indirect). Ideally, the discretization should not affect our estimate of the function. Of course, information will be lost if the intervals of estimation are too large, but it seems reasonable to demand that as the discretization intervals get smaller and smaller, the gross features of the estimated time series should depend less and less on the size of the intervals at which the process is modeled. If there is really an underlying continuous function, the discretization should be a convenience, not a fundamental part of the estimate. (In general, estimating finer intervals increases computation time, leading to a compromise between detail of modeling and feasibility of computation. If the computing power is available, however, we do not want fine-scale modeling to have adverse inferential effects.)

It turns out that standard time series models do not always satisfy the requirement that large-scale inference not depend on the scale of estimation. Here, we consider a familiar one-dimensional example and ask, does a given probability model for a discretized image make sense?

2.1. Autoregressive models for averages

Consider a continuous time series, $f(t)$, on the real line, divided into intervals of width Δ, parameterized by $\theta_i = (1/\Delta) \int_{i\Delta}^{(i+1)\Delta} f(t)dt$, for integer values of i. Suppose that the estimation scale Δ is arbitrary, chosen as fine as possible within the constraints of computation. The simplest conditional autoregressive model is symmetric in the two nearest neighbors, with $E(\theta_i|\theta_j, \text{ all } j \neq i) = (\rho/(1+\rho^2))\theta_{i-1} + \theta_{i+1})$ 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; thus, $\theta_i = (\delta/\Delta) \sum_{j=i\Delta/\delta+1}^{(i+1)\Delta/\delta} \phi_j$. If the AR(1) model is 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. (Given a probability distribution for $\phi_1, \phi_2, \ldots$, the distribution for $\theta_1, \theta_2, \ldots$ is defined uniquely and can be obtained by integrating out the "in-between" parameters. In contrast, a distribution for the θ_i's does not uniquely define a distribution for the more numerous ϕ_i's.)

Unfortunately, it is well known (see, e.g., Lutkepohl (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 $\delta \to 0$ with Δ fixed, 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.

If a family of models is not self-consistent, then the procedure of fitting the family at an *arbitrary* scale is, in general, flawed, because inferences depend on the scale of the model. This is quite different than modeling at a fixed scale (such as in the analysis of annual data), and using, say, an AR(1) model after checking its fit to the data, having considered models from a larger ARMA class.

2.2. A family of nested models

One solution to the problem of models depending on discretization scale is to just explicitly make the discretization interval (or pixel size) a parameter in a larger model, perhaps fixing the pixel size to a value that is consistent with the scientific purpose of the analysis (e.g., ths size of medical features of interest in a MRI scan). We do not like this strategy because it forces discretization to play a double role: both as a scale for computation and a parameter with substantive meaning. We do not want inference about real-world parameters to change every time we get a faster computer. (Or, conversely, we do not want to model at an unnecessarily coarse scale just because we do not have a rich enough family of models.) 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. In modeling a time series of averages, one can set up a family of stationary Gaussian models for which the time series is an ARMA(1,1) at any scale, but whose correlations depend on the discretization in such a way that the models at all scales are coherent. Two families of ARMA(1,1) models satisfy this criterion: the degenerate case of white noise (which is white noise at any scale), and a restricted ARMA(1,1) family that we describe here.

A stationary Gaussian model is determined by its correlations, which for an ARMA(1,1) model can be parameterized as $1, \rho, \eta\rho, \eta^2\rho, \ldots$. If a discretization at scale Δ is modeled as a Gaussian ARMA(1,1) process with parameters

$$\eta = e^{-\Delta/\Delta_0}, \quad \rho = \frac{(1-\eta)^2/2}{-\log\eta - (1-\eta)}, \tag{1}$$

then the family of models at different scales turn out to 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 parameter Δ_0, which is *not* supposed to vary with the scale of estimation, can be thought of as a "characteristic scale" of the family. The parameters η and ρ lie between 0 and 1 for any nonzero Δ, so the ARMA model is always stationary. The bound on η is obvious from its formula, and the bound on ρ is easily obtained from the Taylor expansion, $-\log(\eta) = (1-\eta) + (1-\eta)^2/2 + \cdots$.

The self-consistency property can be shown in two ways. The direct proof starts with the above correlation structure on $\phi_1, \phi_2, \ldots$, at scale δ, and then computes the variance and covariances of the the time series of averages, $\theta_1, \theta_2, \ldots$. After the algebra clears, the θ's have an ARMA(1,1) correlation structure with parameters η and ρ corresponding to the larger scale Δ. In the appendix a proof is presented that connects more clearly with the underlying continuous model and shows that the above models, at all scales Δ, are moving averages of a single stochastic process in continuous time. As $\Delta \to 0$, the correlations approach 1, and the model looks like an AR(1), with $\eta/\rho \to 1$, and the correlation has asymptotic form $\rho \sim 1 - \Delta/\Delta_0$, in the sense that $(1-\rho)/(\Delta/\Delta_0) \to 1$. As $\Delta \to \infty$, the correlations approach 0, and the model looks like an MA(1), with $\eta/\rho \to 0$, and the correlation has asymptotic form $\rho \sim \Delta_0/(2\Delta)$.

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 fit instead 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.

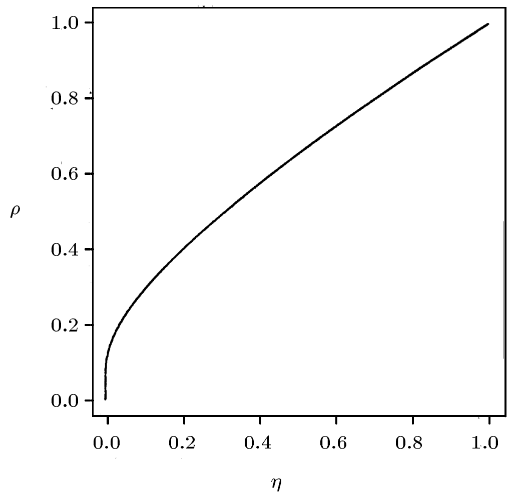

Figure 1. Correlation parameters of the restricted ARMA(1,1) family

Of course, in practice, one will often fit a time series model by first choosing a scale of discretization that fits the data and the scientific questions of interest and then fit the parameters of a particular model class (such as the ARMA family) based on the data. The point of the preceding analysis is not that there is a flaw in standard time series analysis, but rather that that some choices made in the model fitting stage can be determined by the structure of the problem rather than any details of the particular dataset.

3. Models with Constraints

Consider again the estimation of the discretized version of a continuous function. We would like our inferences about the large-scale features to stabilize as the size of the intervals of estimation approaches zero. In this section, we focus on the difficulties that arise due to constraints on the continuous model. Much statistical research has been devoted to the topic of inference about a restricted function (e.g., Barlow et al. (1972)); here, we do not directly consider practical issues of inference but, instead, focus on the mathematical artifacts that arise when applying a discrete model to a continuous function.

3.1. Modeling an increasing time series

Once again, we illustrate the principle involved with a mathematically tractable example. Suppose our unknown continuous time series, $f(t)$, is *known to be increasing* and is 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 $k-1$ equally-spaced points: $\theta_1, \ldots, \theta_{k-1}$, where $\theta_i = f(i/k)$. As Bayesians, we assign the seemingly innocuous uniform prior distribution on the vector $(\theta_1, \ldots, \theta_{k-1})$, 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: $0 < \theta_1 < \cdots < \theta_{k-1} < 1$. This in turn is equivalent to saying that $\theta_1, \ldots, \theta_{k-1}$ are the order statistics of a sample of size $k-1$ from the uniform distribution on $[0,1]$. In particular, each θ_i has a marginal beta distribution, with variance of order $1/k$.

3.2. Inference for a fixed data set

As $k \to \infty$, 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 k 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 $k \to \infty$. In this case, maximum likelihood is reasonable, but its obvious Bayesian extension is treacherous.

3.3. Example: fitting an increasing, convex mortality rate function

For a simple example, we reanalyze the data of Broffitt (1988), who presents a problem in the estimation of mortality rates. (Carlin (1993) provides another

Bayesian analysis of these data.) For each age, t, from 35 to 64 years, inclusive, Table 1 displays N_t, the number of people insured under a certain policy, and y_t, the number of insured who died. (People who joined or left the policy in the middle of the year are counted as half.) We wish to estimate the mortality rate (probability of death) at each age, under the assumption that the rate is increasing and convex over the observed range.

This is a nice example for illustrating the issues of scaling, because the data come binned by year (rather than continuously), but there is really no "natural" physical time scale for the analysis. We will show how a seemingly reasonable inference on the time scale of the data can be unacceptable in practice, a failure that could have been predicted by theoretical arguments alone.

Table 1. Mortality rate data from Broffitt (1988)

age, t	number insured, N_t	number of deaths, y_t
35	1771.5	3
36	2126.5	1
37	2743.5	3
38	2766.0	2
39	2463.0	2
40	2368.0	4
41	2310.0	4
42	2306.5	7
43	2059.5	5
44	1917.0	2
45	1931.0	8
46	1746.5	13
47	1580.0	8
48	1580.0	2
49	1467.5	7
50	1516.0	4
51	1371.5	7
52	1343.0	4
53	1304.0	4
54	1232.5	11
55	1204.5	11
56	1113.5	13
57	1048.0	12
58	1155.0	12
59	1018.5	19
60	945.0	12
61	853.0	16
62	750.0	12
63	693.0	6
64	594.0	10

The observed mortality rates are shown in Figure 2 as a solid line; due to random variation, they are not themselves increasing or convex, even if the true mortality rates are. The observed deaths at each age, y_t, are assumed to follow independent binomial distributions, with rates equal to the unknown mortality rates, θ_t, and known population sizes, N_t. Because the population for each age was in the hundreds, and the rates were so low, we use the Poisson approximation for mathematical convenience: $P(y|\theta) \propto \prod_t \theta_t^{y_t} e^{-N_t\theta_t}$. We used a computer optimization routine to maximize this likelihood under the constraint that the mortality rate be increasing and convex. The maximum likelihood fit is displayed as the dotted line in Figure 2.

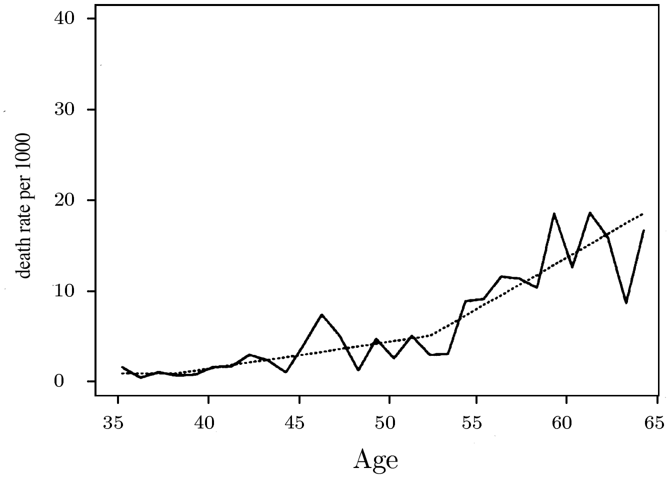

Figure 2. Observed mortality frequencies and the maximum likelihood estimate of the mortality rate function, under the constraint that it be increasing and convex

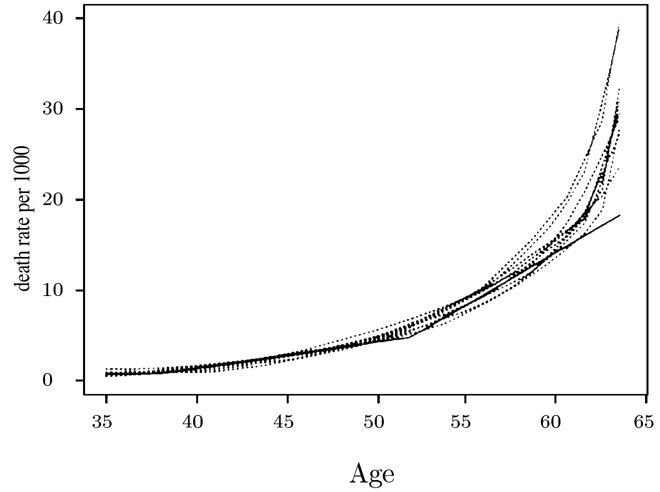

Figure 3. Increasing, convex mortality rates: nine draws from the posterior distribution, with the maximum likelihood estimate (solid line) as a comparison

To perform Bayesian inference, we need to define a prior distribution for θ. Since we were willing to use the maximum likelihood estimate, it seems reasonable to use a uniform prior distribution on θ, with the constraint of increasing convexity. (The uniform distribution is also chosen here for simplicity; Broffitt (1988) and Carlin (1993) apply various forms of the gamma prior distribution. It is an interesting question as to whether their analyses based on informative prior distributions are subject to the problems we discuss here.) Samples from the posterior distribution are generated by the algorithm of Metropolis et al. (1953), as detailed in Gelman, Meng and Stern (1996). Nine samples from the posterior distribution for θ are plotted as dotted lines in Figure 3, with the maximum likelihood estimate from Figure 2 displayed as a solid line for comparison.

The uniform prior distribution for θ seems natural, and the maximum likelihood estimate displayed in Figure 2 seems quite reasonable. However, the prior distribution, as evidenced by the simulations displayed in Figure 3, seems to have been too strong. Rather than "letting the data speak", the posterior simulations of the mortality rate curve seem to curve upwards too strongly, as if they are being forced into the shape of a parabola. Rather than being "noninformative," the multivariate uniform prior distribution has the effect of a strong prior belief that the true function is a quadratic.

This model is, of course, just a slight variant of the example of Sections 3.1–3.2; in fact, the two models are identical if we replace the convex, increasing series $\theta_1, \theta_2, \ldots$, by the positive, increasing series $(\theta_2 - \theta_1), (\theta_3 - \theta_2), \ldots$. (A uniform distribution on the original scale corresponds to a uniform on the differences, since the latter are a linear transformation of the former.) The differences are constrained to be positive and thus, as in the previous example, their prior distribution becomes ever-stronger as the scale of the time intervals becomes smaller. The prior distribution on the series $(\theta_t - \theta_{t-1})$ becomes concentrated on a linear function of t, and so the distribution on the series θ_t becomes concentrated on a quadratic. In our example, if the data were analyzed by age in months, rather than years, the posterior distribution would be focused even more on quadratic curves, virtually ignoring the data. For our purposes, the most important thing about this example is that we should have known not to fit this model, even before seeing any data at all!

The pathological performance of this model is related to the well known result of Stein (1955) on inadmissibility of Bayesian estimates for many parameters based on a joint uniform prior distribution. A general remedy for this problem in the Bayesian context is to replace the improper uniform prior distribution by a hierarchical family of proper prior distributions, with the information content of the prior distribution determined by hyperparameters to be estimated from the data (see, e.g., Morris (1983)). The desired goal is that as the scale of discretization changes with fixed data, the strength of the prior distribution in the Bayesian analysis would remain roughly constant, thus avoiding the problem discussed in Section 3.2. For the mortality rate example, with its nonnormal data and constraint of increasing convexity, it is an open problem whether a suitable hierarchical model will solve the difficulties of scale dependence.

4. Spatial Smoothing

Our final example shows how one can restrict the parameters of a specific image model without seeing any data. As in the previous examples, it is not our purpose to present new results as much as to show that previous results, while derived from data-analytic considerations, can be seen to have an underlying theoretical justification.

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:

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

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

$$S = \frac{1}{W+8} \begin{array}{|c|c|c|} \hline 1 & 1 & 1 \\ \hline 1 & W & 1 \\ \hline 1 & 1 & 1 \\ \hline \end{array} \, . \tag{2}$$

Silverman et al. (1990) report success with values of W from 25 to 100. Is this 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, $W = 5$.

In this example, we consider the discretization as fixed and consider some theoretical considerations that affect the choice of smoothing estimator. The dependence of inferences on the size and orientation of pixels is an important related problem that we do not discuss here.

4.1. Bayesian interpretation of the smoothing parameter

As a way of better understanding the smoothing estimate, we 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 prior density of θ under this model is $P(\theta) \propto \exp[-(\theta - E(\theta))^t \operatorname{diag}(\tau^{-2}) (I - C)(\theta - E(\theta))/2]$, where C is the matrix of conditional autoregression coefficients (c_{ij}), with the diagonal elements, c_{ii}, understood to be zero. In addition, the precision matrix, $\operatorname{diag}(\tau^{-2})(I - C)$, must be symmetric. For simplicity, we 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

$$S \propto (I - \lambda C)^{-1}, \tag{3}$$

where $\lambda = \sigma^2/(\sigma^2 + \tau^2)$. We show that a local conditional autoregression approximately yields the smoother (2), with weight

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

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 $\lambda \approx 1/6$, 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.

There is thus 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 small values of W corresponds to ugly conditional autoregression models, with alternately positive and negative coefficients c_{ij} extending far beyond the local neighborhood. For example, for $W = 4$, the conditional autoregressive coefficients c_{ij} are 0.08, 0.64, -0.53, -0.06, and -0.41, for neighbors of distance 1, $\sqrt{2}$, 2, $\sqrt{5}$, and $\sqrt{8}$, respectively, and coefficients as high as 0.10 appear for neighbors as far apart as $6\sqrt{2}$.

4.2. The local neighborhood smoother

We now derive the results just presented. Given that the smoothing operator is a weighted average (i.e., the smoothing coefficients sum to 1), the following implications are well known (see Kimeldorf and Wahba (1970) and Wahba (1978) for a general discussion and Besag (1986) for the image smoothing interpretation): (a) 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—(b) the model for θ is nonstationary; (c) the prior distribution for θ is improper; (d) the matrix $(I - C)$ is noninvertible.

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

As discussed in Gelman (1990a, b), the smoother can be approximated using the Taylor expansion of (3):

$$S \propto I + \lambda C + \lambda^2 C^2 + \cdots . \qquad (5)$$

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

Consider the following matrix of autoregression coefficients:

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

This is a translation-invariant kernel and is thus characterized by its values c_{ij} for any fixed i, which can be written spatially as

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

with the coefficients centered at the pixel $j = i$.

To get the second order approximation, we compute C^2, which again is characterized by its neighborhood structure for any i:

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

again centered at $j = i$. Then an approximate fit to C^2 that is linear in C is

$$C^2 \approx \frac{1}{8} I + \frac{3}{8} C, \tag{6}$$

and (5) 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 is given by (4). The linear approximation (6) supplies an approximate smoother S with the same neighborhood structure as C or, to look at it another way, an approximate conditional autoregressive model with neighborhood structure C the same as the given smoother S.

In addition, we can reduce the error in the approximation in equation (6) by further theoretical argument. Using only the eight nearest neighbors, we can make C^2 look most like a linear combination of I and C by setting

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

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.

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. One might also consider estimating 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.

5. Definitions of Scale Invariance

The above examples show how consideration of scaling and implicit prior distributions can help one understand and criticize probability models and statistical procedures. It seems desirable to avoid statistical methods corresponding to unnatural models, especially when such methods can be identified before any data have been observed; this is related to the "device of imaginary results" of Good (1950). A model that violates an invariance principle can still be of practical use, but as in goodness-of-fit testing, it is important to understand its flaws. In this discussion, we consider how the principles of scaling that arose in Sections 2 and 3 can be formalized in the context of time series and image models.

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. This is somewhat similar to sieve methods for estimation (see, e.g., Grenander (1981)). It is the family, not any individual model, that should be closed under the scaling operation.

We propose several different definitions of scale invariance. The strongest condition of consistency under scaling is that all discrete models should be derived from a single underlying distribution for the continuous variable, 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)). Our practical goal, however, is to ensure that our inferences under any given class of models are not affected by an arbitrary scale of analysis; construction of an underlying continuous model can be a useful mathematical way of demonstrating scale invariance but is not generally an end in itself.

Many useful families of time series and image models, such as intrinsic autoregressions (Kunsch (1987)), 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 that are nested within the larger pixels, and consider the probability distribution obtained by aggregating to the larger grid. The weaker, limiting scale invariance principle states that for any Δ, the distribution obtained by aggregating smaller pixels should approach a non-degenerate limiting distribution (that will be a function of Δ) as $\delta \to 0$; thus, in that limit, the large-scale distribution is invariant to δ. Another approach that has been suggested is the hierarchical, or multi-grid, model, with components at an infinite series of finer scales. Of course, such models should also satisfy other invariance properties that are applicable in a given problem, such as translational invariance (homogeneity) and rotational invariance (isotropy).

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 2 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. Thus, the model class is inconsistent in the prior but not in the posterior distribution. In contrast, the uniform models in Section 3 violate posterior as well as prior scale invariance. For an example in spatial statistics, the long-range dependence of the Ising model disappears in the posterior distribution that is obtained by conditioning on data observed on the lattice, as is noted by Besag (1991). 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.

Acknowledgements

I thank Xiao-Li Meng for helpful discussions and two referees for helpful comments. This work was supported in part by the National Science Foundation grant DMS-9457824 and a Mathematical Sciences Postdoctoral Fellowship.

Appendix: Derivation of the Restricted ARMA (1,1) Model Using a Continuous Underlying Model

Start by modeling $f(t)$ as an Ornstein-Uhlenbeck process, a Gaussian process with spectral density function, $f_{\text{continuous}}(\omega) \propto (1 + (\omega\Delta_0)^2)^{-1}$, where Δ_0 is a characteristic scale of the continuous process, without reference yet to any discretization. To obtain the spectrum of averages of width Δ, the spectral density must be multiplied by the spectrum of the moving average operator of width Δ:

$$f_{\text{averaged}}(\omega) \propto \frac{\sin^2(\omega\Delta/2)}{(\omega\Delta/2)^2} f_{\text{continuous}}(\omega).$$

Finally, the spectrum of the desired time series, $\theta_1, \theta_2, \ldots$, which are averages in intervals $[0, \Delta]$, $[\Delta, 2\Delta], \ldots$, is obtained by aliasing out wavelengths lower than δ:

$$\begin{aligned} f_{\text{discrete}}(\omega) &= \sum_{k=-\infty}^{\infty} f_{\text{averaged}}\,(\omega + 2\pi k/\Delta) \\ &\propto \sin^2\left(\frac{1}{2}\omega\Delta\right) \sum_{k=-\infty}^{\infty} \frac{1}{(\omega + 2\pi k/\Delta)^2\,(1 + (\omega + \frac{2\pi}{\Delta}k)^2\Delta_0^2)}. \end{aligned}$$

The infinite series can be evaluated using partial fractions and complex integration. Pulling out constant factors that do not depend on ω and simplifying yields,

$$f_{\text{discrete}}(\omega) \propto 1 - \frac{\sinh(\delta/\Delta_0)}{\Delta/\Delta_0} \frac{1 - \cos(\omega\Delta)}{\cosh(\Delta/\Delta_0) - \cos(\omega\Delta)}.$$

When considered as a function of ω, this expression is proportional to the ARMA(1,1) spectrum,

$$f_{\text{ARMA(1,1)}}(\omega) \propto \frac{1 + \eta^2 - 2\rho\eta - 2(\eta - \rho)\cos(\omega\Delta)}{1 + \eta^2 - \eta\cos(\omega\Delta)},$$

with η and ρ defined as in (1) above.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972). Statistical Inference Under Order Restrictions. John Wiley, New York

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). J. Roy. Statist. Soc. Ser.B **36**, 192-236.

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

Besag, J. (1991). Rejoinder to discussion of "Bayesian image restoration, with two applications in spatial statistics," by Besag, York and Mollie. Ann. Inst. Statist. Math. **43**, 45-59.

Besag, J., York, J. and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). Ann. Inst. Statist. Math. **43**, 1-59.

Broffitt, J. D. (1988). Increasing and increasing convex Bayesian graduation. Trans. Soc. Actuaries **40**, 115-148.

Carlin, B. P. (1993). A simple Monte Carlo approach to Bayesian graduation. Trans. Soc. Actuaries **44**, 55-76.

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

Gelman, A. (1990a). Topics in image reconstruction for emission tomography. 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. J. Roy. Statist. Soc. Ser.B **52**, 314-315.

Gelman, A., Meng, X. L. and Stern, H. S. (1996). Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Statist. Sinica, to appear.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. 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.

Good, I. J. (1950). Probability and the Weighing of Evidence. Hafner, New York.

Grenander, U. (1981). Abstract Inference. John Wiley, New York.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Statist. **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. Econom. Lett. **14**, 345-350.

Matheron, G. (1973) The intrinsic random functions and their applications. Adv. in Appl. Probab. **5**, 439-468.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953). Equation of state calculations by fast computing machines. J. Chem. Phys. **21**, 1087-1092.

Morris, C. N. (1983). Parametric empirical Bayes inference: theory and applications (with discussion). J. Amer. Statist. Assoc. **78**, 47-65.

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

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

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

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. J. Roy. Statist. Soc. Ser.B **52**, 271-324.

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Proc. 3rd Berkeley Symp. **1**, 197-206. University of California Press, Berkeley, CA.

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. J. Roy. Statist. Soc. Ser.B **40**, 364-372.

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

Department of Statistics, University of California, Berkeley, CA 94720, U.S.A.

(Received August 1993; accepted January 1995)