Methodologies for Statistical Modeling: A Selective Overview

A. Dąbrowski and J. Koronacki
Wroclaw Academy of Agriculture
and
Institute of Computer Science, Polish Academy of Sciences

Presented at the Workshop on Probabilistic Problems in Atmospheric and Water Sciences (Invited lecture)

Będlewo, December 16 – 18, 2002

1 Summary and Introduction

The aim of this paper is to provide an overview of statistical techniques of use in meteorological and hydrological sciences. While necessarily rather selective, the overview is at the same time intended to give a possibly systematic presentation of methodologies for statistical modeling.

It is well-known that regression analysis is a widely used tool in meteorological and hydrological studies, from direct prediction to donwnscaling to assessing modeling precision in weather and the like forecasting. Classical and modern approaches to regression will be recalled, from linear and generalized linear models to neural networks and regression and classification trees. Down the road, dimension reduction techniques, such as principal component and independent component analyses, will be hinted at. Concerning spatial prediction, kriging methodology will be extended to generalized linear mixed models to cover cases where Gaussian distributional assumptions are clearly inappropriate. Special attention will be given to estimating and modeling space-time correlation structures.

In a natural way, the earlier mentioned methodologies call for Bayesian inferential framework and, within that, Markov chain Monte Carlo (or MCMC) methods. Accordingly, these will be briefly outlined as well.

2 Global parametric models – from linear to generalized linear models, and further

Many reports on climate downscaling begin in the same vein as Sailor (2002):

Although General Circulation Models (GCMs) represent the main features of the global atmospheric circulation reasonably well, their performance in reproducing regional climatic details is rather poor. This is particularly true for variables such as precipitation and surface wind speed. [...] As a result there is a need to develop tools for downscaling GCM predictions of climate change to regional and local scales. [...] The approach involves relating these large scale parameters (e.g., upper level winds, geopotential heights, and sea level pressure) to historical observations of the surface parameter of interest (temperature, precipitation, wind speed, etc.). The required transfer function can be developed using a wide range of modeling tools such as linear regression, classification and regression tree analysis, or neural networks.

Regression tools can prove useful as formal models to capture uncertainty of and assess (or compare) models for weather forecasting. Recently, e.g., such an analysis was reported in Berk et al. (2002) in the context of mesoscale modeling of precipitation events (in the

Los Angeles basin, actually, of the event of February 7 and 8, 1993, when a particularly strong storm event occurred). Simulation results from two models (Mesoscale Model Version 5, or MM5, and another, based on four-dimensional data assimilation, or 4DVAR) were analyzed and compared. Each model was first assessed by conditioning on the data available and then the differences between the two model outputs were regressed on longitude, latitude and elevation (while simple linear regression proved inadequate, more involved regression model, allowing for nonlinearities and product variables, has led to better understanding of the virtual world that the models construct).

Thus motivated, let us turn to a brief discussion of *global parametric regression models*, however well-known they are[1].

Let

$$y(x) = f(x) + \varepsilon_x, \tag{1}$$

where $y(x) \in R$, $x = [x^{(1)}, \ldots, x^{(d)}]^T \in D \subset R^d$ (T denotes transposition) and the x's are assumed nonrandom, $f : D \to R$ is a real-valued function on D ε_x are random errors with distribution possibly depending on x, $E\varepsilon_x = 0$ and $\sup_x Var\varepsilon_x < \infty$ (at this moment, only a marginal distribution of ε_x for a fixed x is considered).

The aim is to estimate f (by an estimator $\hat{f}$), given a random sample of n pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$, where

$$y_i = f(x_i) + \varepsilon_{x_i}, \tag{2}$$

with $y_i \equiv y(x_i)$.

Unless explicitely stated otherwise, f will be assumed known up to some unknown parameters which have to be estimated from available data

$$f(x) = f(x, \beta),$$

where $[\beta_0, \beta_1, \ldots, \beta_d]^T = \beta$ are the parameters in question (in this section, our exposition follows that of Carroll and Ruppert (1988)). Let us rewrite model (2) in the following obvious form

$$y_i = f(x_i, \beta) + \varepsilon_{x_i}, \tag{3}$$

with $[\beta_0, \beta_1, \ldots, \beta_d]^T = \beta$ and β_i , $i = 0, 1, \ldots, d$, to be estimated. Model (2) can also be written as

$$E(y_i) = \mu_i(\beta) = f(x_i, \beta), \tag{4}$$

but a covariance structure of the ε_{x_i}'s still remains to be specified.

[1]The term *global* refers to the fact that an approximating function $\hat{f}$ is fitted to the data globally, not locally – i.e., the function parameters do not depend on the function's argument, but are constant over the whole domain.

The covariance matrix of $[y_1, y_2, \ldots, y_n]^T$ (and thus, equivalently, of $[\varepsilon_{x_1}, \ldots, \varepsilon_{x_n}]^T$) will in general be written as the $n \times n$ matrix $\sigma^2\Lambda = \sigma^2\Lambda(\mu, z, \theta)$, to emphasize possible dependence of the covariance matrix on the mean vector μ, the structural variance parameter θ (possibly a vector) and a matrix of known variables $Z = [z_1, \ldots, z_n]^T$ whose individual component vectors z_i may or may not include some or all of the predictors x_i (note that this is a fairly general dependence model as the x_i's form a set of observations of predictor values). Throughout this section, Λ will be assumed diagonal and the variance structure will be written as

$$Var(y_i) = \sigma^2 g^2(\mu(\beta), z_i, \theta) = \sigma^2/w_i, \tag{5}$$

where function g, the so-called variance function, is either assumed fully known or known up to an unknown parameter θ or unknown. In the last case, weights w_i have to be estimated directly from data (to estimate the weight for particular i, more than one observation at x_i is needed; this case will not be considered here. In the second case, it is parameter θ which has to be estimated; we shall refer to this case later in this section. For the time being, we shall stay with the first case.

Of course, with f of the form

$$f(x) = \beta_0 + \beta^T x,$$

i.e., with f assumed to be a linear (or, more precisely, affine) function in x with unknown parameters

$$[\beta_0, \beta_1, \ldots, \beta_d]^T = \beta$$

and with $g \equiv 1$ in (5), we have a classical *linear regression model* (which, by the way, includes such models as analysis of variance, or ANOVA, and analysis of covariance, or ANCOVA, models as special cases). The estimates for β are then obtained via the *least squares* (LS) method, i.e., by minimizing with respect to β the exspression

$$\sum_{i=1}^{n} [y_i - f(x_i, \beta)]^2.$$

No distributional assumptions for the y_i's are needed, except that they should be continuous random variables. If, however, the y_i's are normally distributed,

$$y_i \sim N(\mu_i, \sigma^2),$$

the estimates $\hat{\beta}_i$, $i = 1, 2, \ldots, n$, obtained via LS are at the same time the maximum likelihood (ML) estimates.

Remark: Note that the linear regression model is in fact fairly general as it, e.g., includes polynomials of fixed order of original predictors. Indeed, e.g., for a quadratic polynomial in two variables, $v^{(1)}$ and $v^{(2)}$, with unknown parameters,

$$\beta_0 + \beta_1 v^{(1)} + \beta_1 v^{(1)} + \beta_2 v^{(2)} + \beta_3 v^{(1)} v^{(2)} + \beta_4 (v^{(1)})^2 + \beta_5 (v^{(2)})^2,$$

simple substitution like $v^{(1)} = x^{(1)}$, $v^{(2)} = x^{(2)}$, $v^{(1)}v^{(2)} = x^{(3)}$, etc., makes the problem fit the general framework for linear models (in this case, in R^5). To put it otherwise, what makes a regression model "linear", is its linearity with respect to the parameters, not to the original predictors.

Note that, in general, finding LS estimates for a nonlinear model (3) with constant variance σ^2 requires either linearization of the model or using some quasi-Newton algorithm. Usually, the latter approach is followed (special algorithms for the problem mentioned are known as *nonlinear least squares methods*).

For a general model (4)-(5), linear or not, with nonconstant variances, it is immediately seen that multiplying both sides of (3) by $w_i^{1/2}$ leads to a new model with constant variance. The LS expression for the new model, when translated back to the original model becomes the *weighted least squares* (WLS) expression,

$$\sum_{i=1}^{n} w_i[y_i - f(x_i, \beta)]^2,$$

to be minimized with respect to β.

Minimizing WLS amounts to solving for β the equation

$$\sum_{i=1}^{n} f_\beta(x_i, \beta)[y_i - f(x_i, \beta)]/g^2(\mu_i(\beta), z_i, \theta) = 0. \tag{6}$$

Clearly, even if the variance function in (5) is known in principle, solving (6) requires using some estimate of w_i since β, and hence true value of w_i is in fact unknown. Methods which use such estimates for solving (6) are known as *generalized least squares* (GLS). In particular, *iteratively reweighted least squares* algorithms proceed essentially by starting with some initial estimate of β, substituting it into the WLS exspression and minimizing the expression by nonlinear least squares, then using the solution obtained as a new estimate of β and so iterating the steps mentioned until convergence. In practice, the Gauss-Newton form of iteratively reweighted least squares is used, which avoids repeated nonlinear least squares minimizations (see, e.g., McCullagh and Nelder (1989)).

Generalized least squares is a powerful technique with easily computable estimates and nice asymptotic properties which make statistical inference about the regression parameters feasible. For fixed sample size, n, no distributional assumptions about the (continuous) y_i's are required to make the approach workable. At the same time, it is obvious that if such distribution were known, it could be used, at least in principle, to advantage, e.g., to reduce estimates' variance. Indeed, when the distribution in question is, or can be assumed, known, one should switch to ML estimation since, provided suitable regularity conditions are fulfilled, the estimates thus obtaind are not only asymptotically unbiased but efficient (i.e., have minimum asymptotic variance). Furthermore, whatever the asymptotic properties of ML estimators, one can be interested in obtaining "most likely" values

of estimates, not those which minimize the (weighted) residual sum of squares. And, last but not least, **regression model is clearly inappropriate for the y_i's being, e.g., counts or proportions** (of successes in binomial experiments).

Within the framework of linear models, all these issues have been resolved by introducing generalized linear theory and models (for a book length exposition see McCullagh and Nelder (1989)). The standard Gaussian linear model can be described by the following assumptions about the three thus introduced components of the model:

(i) Predictors may only influence the distribution of response y through a *systematic component*

$$\eta = x^T \beta.$$

(ii) A *random component* of the model specifies a probability distribution for response y:

$$y \sim N(\mu, \sigma^2).$$

(iii) *Link function* gives the functional relationship between the systematic component, η, and the expected value of the random component, μ:

$$\eta = m(\mu),$$

where $m(\cdot)$ is a suitable function, in the standard Gaussian linear case the identity function, $\eta = \mu$.

The *generalized linear model* (or GLM) is obtained via replacing density $N(\mu, \sigma^2)$ in assumption (ii) by a density or probability mass function from the exponential family (with parameters ν and ϕ):

$$h(y; \nu, \phi) = \exp\left[\frac{y\nu - b(\nu)}{a(\phi)} + c(y, \phi)\right]$$

for some functions $b(\cdot)$ and $c(\cdot)$. Unless h is a Gaussian density (which is a special case from the exponential family of distributions), the link function is not equal to identity function. Other special cases from the exponential family are, e.g., Binomial, Poisson and Gamma distributions.

If $m(\cdot)$ is such that $\eta = m(\mu) = \nu$, then it is called the canonical link. Thus, for normal distribution the canonical link is the identity function, $\eta = \mu$. It can be easily shown that for Poisson distribution the canonical link is equal to natural logarighm. For Binomial distribution and the response defined as the proportion rather than the count (and hence $\mu = p$, the probability of success), the canonical link is given by the logit or logistic transformation, $\eta = \log(p/(1-p)$.

It is easy to see that if in (3) and (4)

$$f(x_i, \beta) = f(x_i^T \beta), \tag{7}$$

then one gets a generalized linear model. More generally, if (7) does not hold, model (4)-(5) can be considered a generalized nonlinear one (see Carroll and Ruppert (1988) for a detailed exposition). Interestingly, it can be shown that, in the class of generalized nonlinear models, ML estimation amounts to solving for β equation (6). Thus, in this class, generalized least squares is equivalent to ML estimation.

Until now we have been assuming that θ in (5) is known. If it is not, GLS methodology is still used, but combined with ML estimation of θ under the assumptions that the true value of β is provided by its current estimate and that the residuals are Gaussian; the procedure is iterated. In general, the likelihood function used is not true likelihood and is therefore called *pseudo-likelihood*. Again, for details see Carroll and Ruppert (1988; see there also for regression diagnostics within the framework discussed as well as for a thorough discussion of useful transformations of nonlinear models).

In the next section we show that GLM's have their place in geostatistics.

3 From kriging to generalized linear models

Model (1) may be extended to the case of spatial data, where, in its standard formulation, one assumes that the data are generated by the model

$$y(x) = \mu + S(x) + \varepsilon_x, \tag{8}$$

where $S(x)$ is a stationary Gaussian process with

$$E[S(x)] = 0$$

and

$$\gamma(x, x') \equiv Cov\{S(x), S(x')\} = \sigma^2 \rho(x - x').$$

Random variables ε_x are, for different x, mutually independent $N(0, \tau^2)$, μ is a constant mean effect. The parameter $x \in R^d$ is spatial location of the process $y(x)$.

The aim is to estimate unknown parameters and find a predictor for underlying spatial surface $S(x)$ given a random sample of n pairs of observations (x_i, y_i), $i = 1, 2, \ldots, n$, where

$$y_i \equiv y(x_i) = \mu + S(x_i) + \varepsilon_{x_i}, \tag{9}$$

An equivalent formulation is that, conditional on $S(\cdot)$, the y_i are mutually independent, with

$$y_i | S(x_i) \sim N\left(\mu + S(x_i), \tau^2\right). \tag{10}$$

A natural, linear predictor $\widehat{S}(x)$, minimizing

$$E\left[\left\{\widehat{S}(x) - S(x)\right\}^2\right]$$

is the *kriging predictor* (named after D.G. Krige, a South African mining engineer, who, in the 1950s, developed empirical methods for determining true ore-grade distributions from distributions based on sampled ore grades; see Cressie (1993) for a detailed treatment of kriging methodolody)

$$\widehat{S}(x) = \sum_{i=1}^{n} w_i(x)\, y_i, \tag{11}$$

where the kriging weights $w_i(x)$ are derived from the data.

Using properties of the multivariate normal distribution, the kriging predictor for $S(x)$ is

$$\widehat{S}(x) = g(x)^T \left(\tau^2 I + G\right)^{-1} (Y - \mu), \tag{12}$$

with prediction variance $V(x) \equiv Var(S(x)|Y))$

$$V(x) = \sigma^2 - g(x)^T \left(\tau^2 I + G\right)^{-1} g(x); \tag{13}$$

here

$$\begin{aligned} \mu &= (\mu(x_1), \mu(x_2), ..., \mu(x_n))^T, \\ Y &= (y_1, y_2, ..., y_n)^T \\ g(x) &= (\gamma(x_1, x), \gamma(x_2, x), ..., \gamma(x_n, x))^T \\ G &= [G_{ij}] = [\gamma(x_i, x_j)]. \end{aligned}$$

In some applications, there may be grounds for linking spatial variation in the regression function to a vector of observable spatial explanatory variables, say $d(x)$. This leads to a more general formulation

$$E[y_i | S(x_i)] = M(x_i), \tag{14}$$

where

$$h[M(x)] = d(x)^T \beta + S(x) \tag{15}$$

and y_i are, conditionally on S, mutually independent, with densities $f_i(u | S(x_i))$.

Here $h()$ is a known link function, β are unknown parameters. So, the data $y_i, i = 1, 2, ..., n$ follow generalized linear (mixed) model[2]. The role of the Gaussian process S is to explain the residual spatial variation after accounting for all known explanatory variables.

We thus have the following generalized linear model:

[2]In statistical terminology, $d(x)^T\beta$ is a fixed effect, $S(x)$ is a random effect and a model whose systematic component contains both effects is called *mixed*. See Breslow and Clayton (1993) for an account on generalized linear mixed models.

1. S is Gaussian, with mean $E[S(x)] = 0$ and covariance function

$$\gamma(x, x') \equiv Cov\{S(x), S(x')\} = \sigma^2 \rho(x - x');$$

2. conditional on $S(\cdot)$, the y_i, $i = 1, 2, \ldots, n$, are mutually independent, with distributions

$$f_i(y \,|S(x_i)) = f(y; M_i) \tag{16}$$

specified by the values of the conditional expectations $M_i = E[y_i \,|S(x_i)]$;

3. $h(M_i) = S(x_i) + d(x_i)^T \beta$, for some known link function h and parameters β.

Write $S = (S(x_1), S(x_2), ..., S(x_n))^T$ for the set of values of $S(x)$ at the sampling locations $x_1, x_2, ..., x_n$ and $S^* = (S(x_1^*), S(x_2^*), ..., S(x_n^*))^T$ for the set of values of $S(x)$ at the locations $x_1^*, x_2^*, ..., x_m^*$ for which predictions are required. Also, let $g_k(s)$ denote the multivariate normal probability density of the first k elements of $\left(S^T, S^{*T}\right)^T$, with elements $s_1, s_2, ..., s_n, s_{n+1}, s_{n+2}, ..., s_{n+m}$.

For general GLM, the generalized linear predictor for location x_j^* is given by

$$\widehat{S}\left(x_j^*\right) = \frac{\int s_{n+j} \left\{\prod_{i=1}^n f_i(y_i \,|s_i)\right\} g_{n+m}(s)\, ds ds_{n+j}}{\int \left\{\prod_{i=1}^n f_i(y_i \,|s_i)\right\} g_{n+m}(s)\, ds}, \tag{17}$$

with prediction variance

$$V\left(x_j^*\right) = \frac{\int s_{n+j}^2 \left\{\prod_{i=1}^n f_i(y_i \,|s_i)\right\} g_{n+m}(s)\, ds ds_{n+j}}{\int \left\{\prod_{i=1}^n f_i(y_i \,|s_i)\right\} g_{n+m}(s)\, ds} - \left\{\widehat{S}\left(x_j^*\right)\right\}^2, \tag{18}$$

For details and examples see Diggle et al. (1998).

4 From local smoothing to adaptive modeling

Much of contemporary statistics is a march away from causal modeling, statistical inference and confirmatory analysis towards exploratory data analysis and prediction models. The fad which goes like "Let the data speak for themselves" is now prevalent – the emphasis is shifted from explaining and understanding an observed phenomenon to fitting a model to data and predicting (for a fascinating discussion of the subject see Thompson (2001)). No doubt, this strand has two fathers: the late John Tukey and increasing at an unbelievable pace computational power of contemporary computers. Like it or not, one is now capable of describing or predicting behavior of truly complex phenomena, not long ago too complex to be subjected to any systematic study.

Regarding the above statement, regression analysis and classification are no exception to the rule. When an apriori model is hardly available, one can now turn to one of (perhaps too) many methods capable of fitting to data in a very flexible way, in particular via *nonparametric approach* (i.e., the approach where a regression function to be estimated is assumed, say, continuous but otherwise completely unknown).

When discussing such flexible approaches, one should first distinguish between low dimensional problems ($d \leq 2$ or, at most, $d \leq 3$) and others (in fact, other than low dimensional problems could be further split, but we skip this detail as too technical to be included in a broad overview).

Concerning low dimensional problems, we shall confine ourselves to merely listing polynomial (most often cubic) splines, the so-called kernel (Nadaraya-Watson or the like) estimators and radial basis functions estimators, and to a brief discussion of *local smoothing methods*.

Let a (local) smoother be defined as:

$$\hat{f}(x) = \tilde{f}(x; \{\hat{a}_j(x)\}_1^p), \tag{19}$$

where $\tilde{f}$ is a simple parametric function (say, a polynomial of order 1 or 2) and

$$\{\hat{a}_j(x)\}_1^p = \text{argmin}_{\{a_j\}_1^p} \sum_{i=1}^{n} w(x, x_i) \times (y_i - \tilde{f}(x_i; \{a_j\}))^2; \tag{20}$$

$$w(x, x_i) = K(||x - x_i||/s(x))$$

with $K : R^+ \to R^+$ a symmetric kernel function with maximum at zero and compact support, and s being a positive smoothing factor depending on x (e.g., such that a given proportion of data around x gives $K > 0$, this proportion being either fixed in advance or chosen adaptively, so as to minimize the so-called cross-validated error). For details on locally weighted polynomial (or LOESS) smoothers and other local regression models see chapter 8 by Cleveland et al. in Chambers and Hastie (1992), Fan and Gijbels (1996); for other nonparametric approaches in low dimensions see Green and Silverman (1994), Wand and Jones (1995); see also Ripley (1996), Hastie, Tibshirani and Friedman (2001).

Estimate $\hat{f}$ for unknown function f, given by (19), is provided by solving (20) for a sufficiently dense grid of x's from the function's domain. Nothing, however, can be gained for free. Flexibility in the way the function's estimate is constructed requires that in some (small) neighborhood of any of the x's in the grid there were observations x_i; indeed, if no pairs (x_i, y_i) have x_i in a small neighborhood of some x, then there are no y_i's which could be used to provide a reasonable smoothed estimate of $f(x)$ for this particular x. Put otherwise, it is here where the so-called curse of dimensionality enters – huge data sets are needed to reasonably fill a high-dimensional domain (say, 100 equally spaced data points form a dense grid in interval $[0, 1]$, but hardly so in square $[0, 1] \times [0, 1]$, let alone

in a unite cube in R^3, not to mention higher dimensions). In fact, local modeling not only requires huge sets of data in higher dimensions but the methods themselves prove computationally demanding in those dimensions.

These are the reasons that, when turning to higher-dimensional problems, it is wise to project original problems into lower dimensions in the first place. Let us begin with additive models which are given as:

$$\hat{f}(x) = \beta_0 + \sum_{j=1}^{d} \tilde{f}_j(x^{(j)}), \tag{21}$$

where

$$\{\tilde{f}_j(x^{(j)})\}_{j=1}^{d} = \text{ argmin}_{\{\tilde{f}_j\}} \sum_{i=1}^{n} (y_i - \sum_{j=1}^{d} \tilde{f}(x_i^{(j)}))^2.$$

Functions $\tilde{f}_j$ are unknown and are estimated by, e.g., (local) univariate smoothers (or other nonparametric estimators). Finding the $\tilde{f}_j$'s proceeds iteratively, via the so-called *back-fitting algorithm* (see, e.g., Fan and Gijbels (1996) for the algorithm and more on local modeling).

One way of generalizing additive models is by extending generalized (non)linear models to *generalized additive models* (see Hastie and Tibshirani (1990) for a book length treatment).

Another approach consists in adaptively looking for low-dimensional projections which reveal interesting patterns in the data. One such method is *projection pursuit regression* or PPR, as proposed by Friedman and Stuetzle (1981). It can be considered a strict sense adaptive method of estimation. The estimator has the form:

$$\hat{f}(x) = \beta_0 + \sum_{j=1}^{r} \tilde{f}_j(\alpha_j + \beta_j^T x) \tag{22}$$

where the $\tilde{f}_j$, $j = 1, \ldots, r$, are unknown and estimated nonparametrically, and r is chosen adaptively. To put a long story short: both the estimator's parameters and the f_j's are chosen by minimizing LS; given the current estimate, say for $j = 1$, α_2, β_2 and $\tilde{f}_2$ are found which minimize the current residual sum of squares; the procedure is continued until the residual sum of squares becomes suffciently small. In practice, PPR is found by a version of the back-fitting algorithm. Note that, in the version of PPR as described by (22), interesting one-dimensional projections are looked for. Extension to two-dimensional projections is straightforward (cf. Friedman and Stuetzle (1981)).

It is worthwhile to note that interactions between independent variables are allowed within the PPR model; e.g.,

$$x^{(1)}x^{(2)} = \frac{1}{4}[(x^{(1)} + x^{(2)})^2 - (x^{(1)} - x^{(2)})^2].$$

We have the following theorem which gives a strong rationale for using the PPR approach:

Theorem (Diaconis and Shahshahani (1984)). Let m be a positive integer. Any homogeneous polynomial f of order m can be written as

$$f(x) = \sum_{j=1}^{\binom{m+d-1}{m}} c_j (a_j^T x)^m$$

for some real c_j and vectors a_j.

Hence, any contiuous function on $[a,b]^d$ can be approximated by a function of the form

$$\sum_{j=1}^{r} \psi_j (a_j^T x),$$

which is equivalent of (22). PPR is therefore often called a "universal approximator".

Interestingly, with $\tilde{f}_j$'s fixed in advance, say, to be sigmoidal functions, PPR reduces to a *feedforward neural network* (with one hidden layer), which can also be shown to be a "universal approximator" (see Cybenko (1989), Hornik, Stinchcombe and White (1989)[3]; see also Ripley (1996)).

Another approach to adaptive estimation is by the so-called *recursive partitioning regression*. Methods of this type are of wider applicability than PPR. Although Friedman and Stuetzle (1981) reported applications of PPR to problems with x in R^{10}, in general the method may fail even for problems of lower dimension, say, 4. Recursive partitioning regression seems a much more promising approach in higher dimensions. The early invention of this latter type was *classification and regression tree*, or CART, of Breiman et al. (1984). A more recent one is *multivariate adaptive regression spline*, or MARS, of Friedman (1991). We shall confine ourselves to briefly describing only MARS (for details see Friedman (1991) and Hastie, Tibshirani and Friedman (2001); see Hastie et al., p. 290, for a brief account of the differences between CART and MARS; in essence, the latter is a refinement over the former).

Let, for $v, t \in R$,

$$(v-t)_+ = \begin{cases} v-t & \text{if } v > t \\ 0, & \text{otherwise.} \end{cases}$$

Let

[3]Essentially, Kolmogorov's theorem on the 13th problem of Hilbert was thus rediscovered.

$$\mathcal{C} = \{(x^{(j)} - t)_+, (t - x^{(j)})_+\}_{t \in \{x_1^{(j)}, \ldots, x_n^{(j)}\},\ j=1,\ldots,d}$$

be the collection of basis functions and let

$$\hat{f}(x) = \beta_0 + \sum_{m=1}^{M} \beta_m h_m(x),$$

where each h_m is a function in $\mathcal{C}$, or a product of two or more such functions.

The model for f is constructed stagewise. We start with only the constant function $h_0(x) = 1$ in the model, and all functions in $\mathcal{C}$ are candidate functions to be added to the model. Given a model with $\mathcal{M}$ functions h_m, a new function is added which has the form

$$\beta_{\mathcal{M}+1} h_m(x)(x^{(j)} - t)_+ + \beta_{\mathcal{M}+2} h_m(x)(t - x^{(j)})_+.$$

From all possible functions of the given form this function is chosen which produces the largest decrease of the residual sum of squares. The two new product terms in the formula above are then added to the set of functions h_m and the process is continued until the model contains some preset maximum number of terms. So obtained "rich" model is then pruned by a cross-validation-like procedure.

See Hastie, Tibshirani and Friedman (2001) for a comprehensive exposition and discussion of modern approaches to regression and classification. Concluding this section, it is worthwhile to note that in the earlier mentioned Sailor (2002), CART was used for downscaling predictions of precipitation. Equally interestingly, linear regression was there applied in the CART's leaves to improve results obtained from CART alone.

5 From time series to space-time correlation modeling

In a significant number of applications, special models and methodology are required to cater for a lack of independence of response variables. A common characteristic of these applications is the dependence of responses $y(x)$ on time. Model (1), with $x \in R$ playing the role of time may be used. The linear order of time introduces special structure on distribution of random errors ε_x, while $f(x)$ is a trend function. Generally, trend function is composed from cyclic component and nonperiodic trend. If all the covariances

of random errors ε_x depend only on time distance and trend function is constant, the stochastic process $y(x)$ is said to be *stationary*.

Statistical methods are applicable if we record observations at equally spaced moments $t_1, t_2, ..., t_n$. The data set which records values of a response variable at the moments t_i is a *time series* $y_1, y_2, ..., y_n$. A stationary time series with trend level μ is an ARMA(p,q) time series if the stochastic difference equation

$$y_i - \mu = \phi_1 (y_{i-1} - \mu) + ...\phi_p (y_{i-p} - \mu) + a_i - (\theta_1 a_{i-1} + ...\theta_q a_{i-q}) \tag{23}$$

holds for $i > \max(p,q)$ and i.i.d., zero mean random variables $a_1, a_2, ...$

A modification of methods described in (1)-(7) is used to fit the parameters in ARMA models and to forecast the values beyond the last observed value t_n.

A special procedures like filtering, periodogram analysis, seasonal decomposition, X-11 census method are used to estimate trend function.

Non-stationary time series may be described as ARCH and GARCH models. The models may be useful in modeling rainfall data, because of periods of "quiet" and activity of the process.These models are especially useful when the goal of the study is to analyze and forecast volatility.

The ARCH model is a linear model

$$\begin{aligned} y_t &= X^T \beta + \epsilon_t, \\ \epsilon_t &= a_t \sqrt{\alpha_0 + \alpha_1 \epsilon_{t-1}^2 + ... + \alpha_p \epsilon_{t-p}^2} \end{aligned}$$

where a_t is $N(0,1)$.

The so-called GARCH models are generalizations of the ARCH models.

A large number of meteorological and climatological models may be regarded as realizations of space-time random fields (De Cesare et al. 2001). The obvious solution to this problem is to consider the spatiotemporal phenomenon as a realization of a random field defined on R^{d+1} where d is physical space dimension. However, there are a number of constraints on such models e.g. positive definiteness of covariance function and because of different topology of time and physical space.

So a restricted class of covariance/variogram models should be introduced to estimate and model of space-time random fields. One of the simplest ways is to consider additive or multiplicative separable models. The additive covariance models

$$\gamma_{st}(x, x'; t, t') = a^2 \gamma_s(x, x') + b^2 \gamma_t(t, t') \tag{24}$$

where $\gamma_s(x, x')$ and $\gamma_t(t, t')$ are covariance functions in space and time domain respectively, are positively semidefinite for certain configurations of spatiotemporal data. The

kriging systems for such data may be non-invertible and are unsatisfactory for optimal prediction.

A very general model combining products and sums can be obtained in the following way:

$$\gamma_{st}(x, x'; t, t') = a^2\gamma_s(x, x') + b^2\gamma_t(t, t') + c^2\gamma_s(x, x')\gamma_t(t, t') \tag{25}$$

A sufficient condition for positive definiteness of the above covariance function is that inequalities

$$\begin{aligned} \max\{\gamma_s(x, x), \gamma_t(t, t)\} &\leq \gamma_{st}(x, x; t, t) < \gamma_s(x, x) + \gamma_t(t, t), \\ \gamma_s(x, x) &> 0, \gamma_t(t, t) > 0 \end{aligned}$$

hold for any x and t

The class of nonseparable, spatiotemporal stationary covariance functions has been derived by Cressie and Huang (1999). Let

- for each $\omega \in R^d$, $\rho(\omega, .)$ be a continuous autocorrelation function
- the positive function $K(\omega)$ satisfy

$$\int K(\omega)\, d\omega < \infty$$

- $H(\omega, t) \equiv \rho(\omega, t) K(\omega)$,
  $H(\omega, t) = (2\pi)^{-d} \int \exp\left(-ih_s^T\omega\right) \gamma_{st}(h_s, t)\, dh_s$.

Then $\gamma_{st}(h_s, t)$ is a spatiotemporal stationary covariance function.

6 Bayesian inference and MCMC

Let us return to the generalized linear mixed model of Diggle et al. (1998). Let θ denote the set of parameters comprising the signal variance σ^2 and any further parameters in the specification of the correlation structure of S, and let β consists of all the regression parameters.

Our objective should now be to use Markov chain Monte Carlo (or MCMC) methods to estimate the model parameters, θ and β, and to generate samples from the conditional distribution of (S, S^*) given Y, where $S = [S(x_1), \ldots, S(x_n)]^T$ is the set of values of $S(x)$

at the sampling locations x_i, and $S^* = [S(x_1^*), \ldots, S(x_n^*)]^T$ are the corresponding values of $S(x)$ at the locations x_i^* for which predictions are required.

Under standard MCMC scheme one should generate random samples from the posterior distribution of (θ, S, β) given Y for inference and from the posterior distribution of S^* given (Y, θ, S, β) for prediction.

The implementation of Diggle et al.'s MCMC scheme requires sampling from the conditional distributions $\pi(\theta|Y, S, \beta)$, $\pi(\beta|Y, S, \theta)$ and $\pi(S(x_i)|S_{-i}, Y, \theta, \beta)$, where S_{-i} denotes the vector S with its ith element, $S(x_i)$, removed.

The general idea behind MCMC is simple (in our exposition we borrow from Smith and Roberts (1993); see also, e.g., Besag et al. (1995)). Suppose that we wish to generate a sample from a distribution $\pi(x)$ for $x \in \mathcal{X} \subset R^p$ but cannot do this directly. However, suppose that we can construct a Markov chain with state space $\mathcal{X}$, which is straightforward to simulate from and whose equlibrium distribution is $\pi(x)$. Now, essentially, what remains is to run the chain for a long time, until the equilibrium is attained.

Gibbs sampler often helps:

Let $\pi(x) = \pi(x_1, \ldots, x_k)$, $x \in R^p$, denote a joint density, and let $\pi(x_i|x_{-i})$ denote the induced full conditional densities for each of the components x_i, given values of the other components $x_{-i} = (x_j; j \neq i)$, $i = 1, \ldots, k$, $1 < k \leq p$.

Now, pick arbitrary starting values $x^0 = (x_1^0, \ldots, x_k^0)$. Then successively make random drawings from $\pi(x_i|x_{-i})$, $i = 1, \ldots, k$, as follows:

$$
\begin{aligned}
x_1^1 \quad &\text{from} \quad \pi(x_1|x_{-1}^0); \\
x_2^1 \quad &\text{from} \quad \pi(x_2|x_1^1, x_3^0, \ldots, x_k^0); \\
x_3^1 \quad &\text{from} \quad \pi(x_3|x_1^1, x_2^1, x_4^0, \ldots, x_k^0); \\
&\ldots \\
x_k^1 \quad &\text{from} \quad \pi(x_k|x_{-k}^1).
\end{aligned}
$$

This completes transition from x^0 to x^1. Iteration of this cycle produces a sequence which is a realization of a Markov chain with transition probability from x^t to x^{t+1}

$$\prod_{l=1}^{k} \pi(x_l^{t+1}|x_j^t,\ j > l,\ x_j^{t+1},\ j < l).$$

The chain so obtained, known as the Gibbs sampler, can be shown to have π as its invariant distribution (see Appendix in Smith and Roberts (1993) for a brief account of general convergence theory for MCMC).

One other algorithm, usually faster than the Gibbs sampler, and again having π as its invariant distribution (again, see Smith and Roberts (1993)), is the Metropolis-Hastings algorithm. In Diggle et al. (1998), that other algorithm is used.

Let us conclude this section by mentioning that in the discussion to Gustafsson (2002), it is noted that real-time weather forecasting problem, which in fact concerns nonlinear filtering, can be stated as follows (see a comment by Høst):

We have a hidden Markov process of interest $X(t)$, described by a transition equation for its probability distribution. Related to the hidden process is a set of observations y_t; these are described by an observation equation. The problem is to estimate the hidden process given the observations.

Høst believes that MCMC can hopefully be used to advantage within this context too.

7 In lieu of conclusions

Of the many omissions made in this overview, however unavoidable, at least some should be named. In addition to having skipped diagnostics for regression models already built, also the preliminary step of analysis, which precedes building a model, was left almost unmentioned. Here we mean transformation of variables, their selection and extraction (by variable extraction we mean constructing new explanatory variables as combinations of the original ones). For such methods, see, e.g., Hastie et al. (2001), Ripley (1996), Carroll and Ruppert (1988) and Miller (1990). From among the methods of variable extraction, let us barely mention two methods of much wider applicability, in fact borrowed from statistical multivariate analysis, namely principal component analysis and much newer independent component analysis.

A major omission was that of statistical (and other) methods for classification: discriminant analysis or supervised classification and clustering or unsupervised classification. In this regard, let us refer the reader to Hastie et al. (2001), Ripley (1996) as well as to such fundamental texts on statistical multivariate analysis as Mardia et. al (1979), Krzanowski (1988), Krzanowski and Marriott (1994 and 1995)[4].

[4]Statistical monographs mentioned include of course exhaustive accounts of principal component analysis, linear or classical in the case of older monographs, both linear and nonlinear in the case of Krzanowski and Marriott (1994 and 1995)

Finally, within both modern regression and discriminant analysis, let us mention such recent and very promising inventions as boosting and additive trees (see, e.g., Hastie et al. (2001)), and random forests of Breiman (2001). For a new clustering paradigm, which allows for clustering on different subsets of variables for different clusters, see Friedman and Meulman (2002).

References

Berk, R.A. et al. (2002), Workshop on Statistical Approaches for the Evaluation of Complex Computer Models, *Statistical Science* **17**, 173-192.

Besag, J. et al. (1995), *Statistical Science* **10**, 3-66.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984), *Classification and Regression Trees*, Chapman & Hall.

Breiman, L. (2001), Random Forests,
http://stat-www.berkeley.edu/users/breiman/rf.html.

Breslow, N.E. and Clayton, D.G. (1993), *J. American Statist. Assoc.* **88**, 9-25.

Carroll, R.J. and Ruppert, D. (1988), *Transformation and Weighting in regression*, Chapman & Hall.

Chambers, J.M. and Hastie, T.J. [eds.] (1992), *Statistical Models in S*, Wadsworth and Brooks/Cole.

Cressie, N.A.C. (1993), *Statistics for Spatial Data*, Wiley.

Cressie, N.A.C and Huang, H. (1999), *J. American Statist. Assoc.* **94**, 1330-1340.

Cybenko, G. (1989), *Math. of Control Signals and Systems* **2**, 303-314.

De Cesare, L, Myers, D.E. and Posa, D (2001), *Statist. & Probab. Letters* **51**, 9-14.

Diaconis, P. and Shahshahani, M. (1984), *SIAM J. on Scientific and Statistical Computing* **5**, 175-191.

Diggle, P.J., Moyeed, R.A. and Tawn, J.A. (1998), Model-based Geostatistics (with discussion), *J. R. Statist. Soc.* C, **47**, 299-350.

Fan, J. and Gijbels, I. (1996), *Local Polynomial Modelling and Its Applications*, Chapman & Hall.

Friedman., J.H. (1991), *Ann. Statist.* **19**, 1-141.

Friedman, J.H. and Stuetzle, W. (1981), *J. American Statist. Assoc.* **76**, 817-823.

Friedman, J.H. and Meulman, J.J. (2002), Clustering Objects on Subsets of Attributes, http://www-stat.stanford.edu/ jhf/#reports.

Green, P.J. and Silverman, B.W. (1994), *Nonparametric Regression and Generalized Linear Models*, Chapman & Hall.

Gustafsson, N. (2002), Statistical Issues in Weather Forecasting, *Scand. J. Statist.* **29**, 219-243.

Hastie, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, Chapman & Hall.

Hastie, T., Tibshirani, R. and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.

Hornik, K., Stinchcombe, M. and White, H. (1989), *Neural Networks* **2**, 359-366.

Krzanowski, W.J. (1988), *Principles of Multivariate Analysis: A User's Perspective*, Clarendon Press.

Krzanowski, W.J. and Marriott, F.H.C. (1994), Multivariate Analysis, Part 1, Arnold.

Krzanowski, W.J. and Marriott, F.H.C. (1995), Multivariate Analysis, Part 2, Arnold.

Mardia, K.V., Kent, J.T. and Bibby, J.M. (1979), *Mulitivariate Analysis*, Academic Press.

McCullagh, P. and Nelder, J.A. (1989), *Gemeralized Linear Models*, Chapman & Hall.

Miller, A.J. (1990), *Subset Selection in Regression*, Chapman & Hall.

Ripley, B.D (1996), *Pattern Recognition and Neural Networks*, Cambridge Univ. Press.

Sailor, D.J. (2002), Climate Downscaling,
http://www.me.tulane.edu/Faculty/Sailor/homepage/SailorGroup/downscaling.htm.

Smith, A.F.M. and Roberts, G.O. (1993), *J. R. Statist. Soc.* B, **55**, 3-23.

Thompson, J.R. (2001), The Age of Tukey, *Technometrics* **43**, 256-265.

Wand, M.P. and Jones, M.C. (1995), *Kernel Smoothing*, Chapman & Hall.