Spatial Process Estimates as Smoothers

Douglas W. Nychka
North Carolina State University , Raleigh, North Carolina, USA and
National Center for Atmospheric Research, Boulder, Colorado, USA

March 30, 1999

Contents

Chapter 1

Spatial Process Estimates as Smoothers

1.1 Introduction

Spatial statistics has a diverse range of applications and refers to the class of models and methods for data collected over a region. This region might be a mineral field for a geologist, a quadrat in a forest for an ecologist or a geographic region for an environmental scientist. A typical problem is to predict values of a measurement at places where it is not observed, or, if the measurements are observed with error, to estimate a smooth spatial process (i.e. a surface) from the data. A family of techniques, known loosely as kriging, were originally developed in geostatistics and are now being widely used for spatial prediction. In this chapter we will also use the term spatial process estimates to describe these techniques. The goals of kriging sound very much like nonparametric regression and the understanding of spatial estimates can be enriched through their interpretation as smoothing estimates. In the other direction, the random process model that is at the heart of spatial methods is also valuable in setting uncertainty estimates for function estimates especially in low noise situations. Lying between standard methods such as kernel smoothing and Kriging are smoothing splines. Because splines have a foot in both of these camps, they aid in understanding the common structure in these approaches. This chapter will present spatial process estimates and smoothing splines in a common form as least squares, ridge regression estimates. This leads to an interpretation for both of these estimates as a particular kind of orthogonal series estimate. The orthogonal series representation is a basis for efficient formulas for cross-validation and simulation and also suggests approximate estimators for large data sets.

There are good surveys of spatial estimates (e.g. Cressie 1991) and in fact the connection between spatial process models and splines has enjoyed much attention, most recently by Kent and Mardia (1994), and some lively debate (Cressie 1990 and Wahba 1990b). So why is another chapter on this subject needed? Most treatments of Kriging do not emphasize the fact that the form of the predictions are a linear combination of basis functions prescribed by the covariance function. The coefficients for this representation are linear functions of the data found by a specific ridge regression. This form makes it easy to see how the Kriging estimator behaves as a smoother and illustrates how important parameters of the covariance function can be found by generalized cross-validation or restricted maximum likelihood.

Part of the motivation for this research comes from an interest in the spatial distribution of ambient ozone pollution in the United States. Another less obvious connection comes from the analysis of experimental data for developing new materials and manufacturing processes (O'Connell and Wolfinger 1997). Here the "spatial" region consists of the variables that can be changed to modify the process. Driven by the applied nature of these problems, we have developed software, FUNFITS[1] (Nychka, et al. 1996) that simplifies the fitting and interpretation of these function estimates. This set object-oriented software benefits from the numerical algorithms based on the ridge regression/orthogonal series representation. The reader is referred to the examples and help files in this software for more supporting information on using these methods and details of computation. One contribution of this chapter is weaving the theoretical development of these estimates with the more pragmatic algorithms for their computation.

1.1.1 Basic model

The observational model considered in this work has the form

$$Y_i = f(\boldsymbol{x}_i) + \epsilon_i \quad \text{for } 1 \le i \le n \tag{1.1}$$

where Y_i is the observed response at the ith combination of design variables (or ith location), $\boldsymbol{x}_i \in \Re^d$, and f is the function of interest. The random components, $\{\epsilon_i\}$, usually associated with the measurement errors, are assumed be uncorrelated, zero-mean random variables with variances, $\{\sigma^2/W_i\}$.

One parsimonious strategy for representing f is as the sum of a low order polynomial and a smooth function:

$$f(\boldsymbol{x}) = P(\boldsymbol{x}) + h(\boldsymbol{x}) \tag{1.2}$$

Typically, in nonparametric regression and smoothing, the assumptions on h are based on higher order derivatives and in the case of thin-plate splines (Wahba 1990a, Green and Silverman 1994) the degree of P and the dimension of $\boldsymbol{x}$ imply a specific roughness penalty on h based on integrated squared (partial) derivatives. In spatial statistics models, the smoothness assumption is replaced by the assumption that h is a random field. Under the assumption that f is a realization of a spatial process,

[1]FUNFITS is a module of S functions for function fitting and statistical designs that supplement those in SPLUS. A manual and the software for UNIX and PC's can be located from the software subheading on Geophysical Statistics Project home page `http://www.cgd.ucar.edu/stats`

P can be identified as the "spatial trend" or "drift", the fixed part of the model, and h is modeled as a mean zero Gaussian process (Cressie 1991). Let

$$Cov(h(\boldsymbol{x}), h(\boldsymbol{x}')) = \rho k(\boldsymbol{x}, \boldsymbol{x}') \tag{1.3}$$

be a covariance function where k is a known function and ρ is an unknown multiplier. Conditions on the covariance function then control the smoothness properties of h.

To simplify the treatment in this chapter it will be assumed that the covariance function is continuous as $\boldsymbol{x}' \to \boldsymbol{x}$. Thus if the spatial process has a nugget effect it is due to the additive errors in 1.1 and not due to a discontinuity when $\boldsymbol{x} = \boldsymbol{x}'$ in the covariance function.

1.1.2 The unifying theme

Splines and spatial process estimates are complicated transformations of the data and a careful development of their properties will require some patience. However, it is important not to let the details obscure the basic form of these methods as orthogonal series estimators. We will construct the set of basis functions $\{g_\nu\}$ for $1 \leq \nu \leq n$ with the property that if $u_\nu = \sum_{k=1}^{n} g_\nu(\boldsymbol{x}_k) W_k Y_k$, then function $\sum_{\nu=1,n} u_\nu g_\nu(\boldsymbol{x})$ will interpolate the data, Y_k. This is possible because of an orthogonality property for the basis. Moreover, the orthogonality will ensure that these interpolation coefficients, u_k, are uncorrelated with each other, simplifying the simulation of these estimators.

In addition to the basis, there is a sequence of increasing nonnegative weights (eigenvalues) , $\{D_\nu\}$ so that

$$\hat{f}(\boldsymbol{x}) = \sum_{\nu=1}^{n} \frac{1}{1 + \lambda D_\nu} u_\nu g_\nu(\boldsymbol{x}) \tag{1.4}$$

where λ is the usual spline smoothing parameter or in terms of the spatial process estimate this is the "signal to noise" ratio $\lambda = \sigma^2/\rho$. Thus we see that both estimates for fixed λ are just tapered versions of the interpolating series. The role of the weighting function is to shrink the magnitude of coefficients where D_ν is large relative to λ.

Where does this magic orthogonal basis come from? For spatial process models the usual Kriging surface is a linear combination of low order polynomial functions and n functions, $\psi_j(\boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}_j)$, obtained by evaluating one argument of the covariance function at the observed locations. A standard matrix decomposition transforms this basis into the orthogonal one. The fact that the Kriging estimators start with an expansion using the covariance function has important consequences for their behavior away from the observed locations. For splines, one starts with polynomials and radial basis functions and then proceeds in the same way.

In this overview, the case of smoothing has been emphasized and is appropriate when the function is observed with error ($\sigma > 0$). When measurement error is not present this is the same as setting $\sigma = 0$ and thus $\lambda = 0$ in the tapering formula in 1.4. The result is a linear combination of the basis functions that interpolates the observed data and can be extrapolated to provide estimates of the surface at unobserved points. Throughout this chapter the interpolation results can be inferred from the smoothing estimators by considering the limiting case as $\lambda \to 0$.

1.1.3 Chapter outline

The next section presents the thin-plate spline estimate. Splines are of interest because they have both a spatial process interpretation and can also be justified as penalized least squares estimates independent of a stochastic model for the unknown function. For this reason it is appropriate that they are discussed first. Section 3 develops spatial process estimates for a random field as the best linear unbiased estimate. The derivation is different from standard treatments, using a QR matrix decomposition to enforce unbiasedness. This approach is included because it parallels the methods of computation.

Section 4 gives the common ridge regression form for Kriging and splines deriving equation 1.4. From this perspective, it is clear how to approximate a Kriging estimate for a large data set based on a reduced set of basis functions. Section 5 uses a simple response surface experiment to illustrate how the orthogonal series representation acts as a smoother and includes a multi-resolution decomposition of the data. Section 6 presents a spatial process example based on predicting surface ozone from a limited set of monitoring locations in the Great Lakes region of the US. The last section suggests some areas that need more research effort.

Finally it should be acknowledged that there is much missing from this treatment. Deeper connections between these two methods based on Bayesian formalism has been largely omitted. Also from a practical viewpoint the craft of identifying and estimating covariance functions has been alluded to in sections 3.4 and 6 but the technology based on variograms has been largely left out.

1.2 Thin-plate splines

A spline is defined implicitly as the solution to a variational (minimization) problem. The form of the minimization criterion determines the kind of spline. A thin-plate spline estimator of f is the minimizer of the penalized sum of squares

$$S_\lambda(f) = \frac{1}{n}\sum_{i=1}^{n} W_i(y_i - f(\boldsymbol{x}_i))^2 + \lambda J_m(f) \tag{1.5}$$

for $\lambda > 0$ and where

$$J_m(f) = \int_{\Re^d} \sum \frac{m!}{\alpha_1!...\alpha_d!}\left(\frac{\partial^m f}{\partial x_1^{\alpha_1}...\partial x_d^{\alpha_d}}\right)^2 d\boldsymbol{x}$$

The sum in the integrand is taken over all non-negative integer vectors, α, such that $\sum \alpha_1 + \cdots + \alpha_d = m$, and $2m > d$. For one dimension and $m = 2$, $J_2(f) = \int \left(f^{(\prime\prime)}(x) \right)^2 dx$, giving the standard cubic smoothing spline roughness penalty. The thin-plate spline is a generalization of the usual cubic smoothing spline with a "roughness" penalty function, $J_m(f)$, that is rotationally invariant and may involve higher partial derivatives than second order. An important feature of a spline estimate is the part that is unaffected by the roughness penalty. The set of functions where the roughness penalty is zero is termed the null space, and for J_m, consists of all polynomials with degree less than or equal to $m - 1$. There is a rich theory that blends the construction of different roughness penalties with statistical models and the reader is referred to Chapter 11 in this volume for more details on different kinds of spline models.

Minimization of S_λ results in a function that tracks the data but is also constrained to be smooth. A Bayesian viewpoint is to interpret $-S_\lambda$, to within a constant, as the log of the joint density function for f and Y. Let p be the conditional density of Y given f and π the prior density for f.

$$-S_\lambda = ln(p(Y|f)\pi(f)) = ln(p(Y|f)) + ln(\pi(f))$$

If one adopts equation 1.5 then the roughness penalty is associated with a log prior distribution for f. Furthermore, this prior can be interpreted as a distribution such that f is a realization from a smooth (almost m differentiable) Gaussian process. From Bayes Theorem we know that for fixed Y the posterior is proportional to the joint density. Thus maximizing the joint density or *minimizing* S_λ results in a mode of the posterior distribution. The reader is referred to Wahba (1990a) and Omre and Halvorsen (1989) for additional discussion of this correspondence. Although this brief discussion sketches out a connection between splines and spatial process estimates, the following sections will pursue a more direct connection. It will be shown that the spline estimate is in fact a minimum variance unbiased estimator under a specific stochastic model. In this way, the interpretation does not depend explicitly on a Bayesian formulation and circumvents the measure theoretic problems of putting prior distributions on function spaces. The next section will set the stage for this (frequentist) interpretation.

1.2.1 Form of the spline estimate

The most important step in deriving the spline estimate is in identifying the solution as a finite linear combination of basis functions. In this way, the abstract minimization over a function space collapses into minimizing over the coefficients associated with the basis functions.

Thin-plate splines are expressed in terms of radial basis functions. Let $\psi_i(\boldsymbol{x}) =$

$E_{md}(\boldsymbol{x} - \boldsymbol{x}_i)$ where E_{md} are the radial basis functions

$$E_{md}(\boldsymbol{r}) \begin{array}{ll} = a_{md} \parallel \boldsymbol{r} \parallel^{(2m-d)} log(\parallel \boldsymbol{r} \parallel) & d \text{ even} \\ = a_{md} \parallel \boldsymbol{r} \parallel^{(2m-d)} & d \text{ odd} \end{array}$$

and a_{md} is a constant. Also, let $\{\phi_j\}$ be a set of t polynomial functions that span the space of all d dimensional polynomials with degree $m-1$ or less. Now consider a function of the form:

$$f(\boldsymbol{x}) = \sum_{j=1}^{t} \phi_j(\boldsymbol{x})\beta_j + \sum_{i=1}^{n} \psi_i(\boldsymbol{x})\delta_i$$

clearly parameterized by $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$.

An elegant development of the spline solution can be obtained from Minguet (1979) and Duchon (1977) with an overview provided by Wahba (1990a). See also Kent and Mardia (1994) for a discussion of generalized functions. For the purposes of this chapter we will abstract three key results that yield the computational form for the spline. Define T, a $n \times t$ matrix with $T_{k,j} = \phi_j(\boldsymbol{x}_k)$. Let

$$\mathcal{V} = \{f : J_m(f) < \infty\} \text{ and } \mathcal{M} = \{f : f = \sum_{j=1}^{t} \phi_j\beta_j + \sum_{i=1}^{n} \psi_i\delta_i \text{ and } T^T\boldsymbol{\delta} = 0\}.$$

Then

1. $\mathcal{M} \subset \mathcal{V}$
2. $min_{f\in\mathcal{V}} S_\lambda(f) = min_{f\in\mathcal{M}} S_\lambda(f)$
3. If T has full rank then a unique solution exists.

A key point is that $\mathcal{M}$ is spanned by a finite number of basis functions. Moreover, from these three basic properties it follows that the solution to the spline minimization problem can be reduced to a (quadratic) optimization problem over a finite dimensional space. With this result it is useful to rephrase the minimization problem in matrix/vector form, substituting the finite dimension form into $S_\lambda(f)$. Although the spline estimator has its roots in the abstract minimization over a function space, it order to derive computational algorithms we will be using this matrix form for most of the remainder of this chapter.

An important detail in the theory of thin-plate splines is a remarkable simplification for the roughness penalty when the solution is a linear combination of radial basis functions. Let $K_{i,k} = \psi_i(\boldsymbol{x}_k)$, then if $T^T\boldsymbol{\delta} = 0$

$$J_m(\sum_{j=1}^{t} \phi_j\beta_j + \sum_{i=1}^{n} \psi_i\delta_i) = \boldsymbol{\delta}^T K \boldsymbol{\delta}$$

This expression is very useful for obtaining a simple form for S_λ.

Assembling all these pieces,

$$S_\lambda(\sum_{j=1}^{t}\phi_j\beta_j + \sum_{i=1}^{n}\psi_i\delta_i) = (Y - T\boldsymbol{\beta} - K\boldsymbol{\delta})^T W(Y - T\boldsymbol{\beta} - K\boldsymbol{\delta}) + \lambda\boldsymbol{\delta}^T K\boldsymbol{\delta}. \qquad (1.6)$$

Now taking partials with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ the solution satisfies the system of equations

$$\begin{aligned} -2T^T W(Y - T\boldsymbol{\beta} - K\boldsymbol{\delta}) &= 0 \\ -2KW(Y - T\boldsymbol{\beta} - K\boldsymbol{\delta}) + 2\lambda K\boldsymbol{\delta} &= 0 \end{aligned}$$

If $\{\boldsymbol{x}_i\}$ are unique locations then K will have full rank. Multiply the second equation by K^{-1} and substitute this new equation into the first. This yields the system of equations

$$T^T\boldsymbol{\delta} = 0 \qquad (1.7)$$

$$T\boldsymbol{\beta} + K\boldsymbol{\delta} + \lambda W^{-1}\boldsymbol{\delta} = \boldsymbol{Y} \qquad (1.8)$$

It is easy to verify that the solution is given by

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (T^T(K + \lambda W^{-1})^{-1}T)^{-1}T^T(K + \lambda W^{-1})^{-1}\boldsymbol{Y} \\ \hat{\boldsymbol{\delta}} &= (K + \lambda W^{-1})^{-1}(Y - T\hat{\boldsymbol{\beta}}) \end{aligned} \qquad (1.9)$$

As a link to later discussion it should be noted that formally $\hat{\beta}$ is a generalized least squares regression estimate if we identify $(K + \lambda W^{-1})$ as a covariance matrix for Y. This linear system also gives a simple form for the residuals from the splines fit. The predicted values are just $T\hat{\boldsymbol{\beta}} + K\hat{\boldsymbol{\delta}}$, so based on 1.8, the residual vector must be equal to $\lambda W^{-1}\hat{\boldsymbol{\delta}}$.

1.2.2 The QR parameterization

The previous derivation produces a solution that is readily interpreted but it is not a form that is suitable for computation. Also in terms of making a comparison with Kriging estimates, a different parameterization is required. This computational approach is due to Bates et al. (1987) and the reader is referred to this source and Gu and Wahba (1991) for further details and software.

To enforce the constraint on δ implied by 1.7, let $T = F_1R$ denote the QR decomposition of T, $F = [F_1 \mid F_2]$ is an orthogonal matrix. F_1 has columns that span the column space of T, F_2 has columns that span the space orthogonal to T and R is an upper triangular matrix. Reparameterizing by $\delta = F_2\omega_2$ for $\omega_2 \in \Re^{n-t}$ now enforces the necessary number of linear constraints on δ and makes the parameters identifiable. Now, setting $\boldsymbol{\beta} \equiv \boldsymbol{\omega}_1$ and $\boldsymbol{\omega}^T = (\boldsymbol{\omega}_1^T, \boldsymbol{\omega}_2^T)$, the system of equations reduces to the form:

$$T\boldsymbol{\omega}_1 + KF_2\boldsymbol{\omega}_2 + \lambda W^{-1}\boldsymbol{F}_2\boldsymbol{\omega}_2 = \boldsymbol{Y}$$

Multiplying through by F_2^T gives

$$\begin{aligned} \hat{\boldsymbol{\omega}}_2 &= (F_2^T(K + \lambda W^{-1})F_2)^{-1}F_2^T\boldsymbol{Y} \\ \hat{\boldsymbol{\omega}}_1 &= R^{-1}F_1^T(\boldsymbol{Y} - (K + \lambda W^{-1})F_2\hat{\boldsymbol{\omega}}_2) \quad . \end{aligned} \tag{1.10}$$

and of course we can relate this solution back to the original parameterization by $\hat{\delta} = F_2\hat{\omega}_2$.

1.2.3 The hat matrix, kernels and cross-validation

Examining the equations leading to $\hat{\omega}$, we see that all matrices involved are independent of $\boldsymbol{Y}$. Thus, for fixed λ and fixed locations, splines are linear functions of the data and there is an $n \times n$ hat matrix, $A(\lambda)$, such that

$$\begin{pmatrix} \hat{f}(\boldsymbol{x}_1) \\ \vdots \\ \hat{f}(\boldsymbol{x}_n) \end{pmatrix} = A(\lambda) \begin{pmatrix} \boldsymbol{Y}_1 \\ \vdots \\ \boldsymbol{Y}_n \end{pmatrix}$$

In analogy with linear regression, the trace of A will be used as a measure of the effective number of parameters being subsumed by the spline. More will be said about this in Section 5. The ith row of the A matrix gives the weights that are applied to the data to give the function estimate at $\boldsymbol{x}_i$. When one plots these weights they look like higher order kernels, although an explicit formula between λ and a bandwidth is not obvious. In one dimension the kernel approximation to a smoothing spline has been studied (Nychka 1995 and Messer 1991) and for uniformly distributed locations the bandwidth is proportional to $\lambda^{1/2m}$. Based on this work, it is not difficult to conjecture in the multivariate case that the approximating kernel is a Greens function for a particular 2mth order partial differential operator. An explicit form for the kernel has been derived by Zhang and Stein (1993) but a mathematically rigorous extension to the multivariate case remains to be accomplished.

The hat matrix also yields a simple formula for leave-one-out prediction. Let $\hat{Y}_k$ denote the predicted value for the kth data point based on the spline estimate and let $\hat{Y}_k^*$ be the prediction for this point based on omitting the kth data point and refitting the spline to the remaining $n-1$ data values. The difference $Y_k - \hat{Y}_k^*$ will be referred to as the cross-validated (CV) residual. Using the properties of linearity and the fact that the spline solves a minimization problem, we have the following compact formula for the CV residuals based on leave-one-out prediction.

$$Y_k - \hat{Y}_k^* = \frac{(Y_k - \hat{Y}_k)}{1 - A_{kk}(\lambda)}$$

Note that the right side only depends on the full data spline and the diagonal elements of the A matrix. Thus, the sum of squared CV residuals is a weighted sum using the full data residuals and can be computed without the necessity of refitting the spline to n different subsets. A further simplification is to substitute the average of the diagonal elements for each of the individual terms in the denominator. These remarks serve as motivation for the generalized cross-validation (GCV) function

$$V(\lambda) = \frac{(1/n)\sum_{k=1}^{n} W_k (Y_k - \hat{Y}_k)^2}{(1 - trA(\lambda)/n)^2} = \frac{(1/n)||(I - A(\lambda))\boldsymbol{Y}||^2}{(1 - trA(\lambda)/n)^2}$$

The principle is that good choices of λ will give small CV residuals. So one might choose λ by minimizing $V(\lambda)$. There is some debate as to how much GCV really has to do with cross-validation. However, it can be justified independently on theoretical grounds and typically is nearly unbiased for estimating the average prediction error of the spline estimate. One pitfall is the variability associated with the estimated smoothing parameter from minimizing the GCV criterion and a small fraction of estimates may severely undersmooth the data. These cases can often be spotted in practice by unreasonably small estimates of σ.

1.3 Spatial process estimates

The derivation of a spline estimate in the previous section assumed that the unknown function is fixed. In contrast, a spatial process estimate is based on a model for f as a random function. Accordingly, we now assume f has the form from 1.2 and h is a mean zero Gaussian process. We will see that the covariance function for h now plays the same role for spatial process estimates as the radial basis functions do for splines.

1.3.1 The universal kriging estimator

Given observed data, $\{Y_k\}$, an estimate for $f(x)$ can be found as a linear combination of the observations that minimizes the variance and is unbiased. The standard derivation usually enforces the unbiasedness condition (a linear constraint) using Lagrange multipliers. A more direct approach is to use the QR decomposition to restrict the linear combination to a subspace where unbiasedness holds. In this way we convert the constrained minimization problem to an unconstrained one that is simple to solve.

For clarity, let $\boldsymbol{x}_0$ be the new point for prediction and so the goal is an optimal estimate for $f(\boldsymbol{x}_0) = \boldsymbol{T}_0^T\boldsymbol{\beta} + h(\boldsymbol{x}_0)$ where $T_{0,j} = \phi_j(\boldsymbol{x}_0)$.

Any linear estimate has the form $\hat{Y}_0 = \boldsymbol{l}^T\boldsymbol{Y}$ with $\boldsymbol{l} \in \Re^n$ so our goal is to identify $\boldsymbol{l}$. Now let F and R be the QR decomposition of the T matrix, $(T = F_1 R)$, described in the previous section. Equivalently we can express $\boldsymbol{l}$ as $\boldsymbol{l} = (F\boldsymbol{a})$ for some $\boldsymbol{a} \in \Re^n$. Thus, the estimate is $\hat{Y}_0 = (F\boldsymbol{a})^T\boldsymbol{Y}$. Furthermore, $F\boldsymbol{a}$ can be partitioned as $F_1\boldsymbol{a}_1 + F_2\boldsymbol{a}_2$ where we have subdivided the vector $\boldsymbol{a}^T = (\boldsymbol{a}_1|\boldsymbol{a}_2)^T$. Although this form may seem awkward at first sight, it is very easy to enforce the unbiasedness using this decomposition. Determining $\boldsymbol{a}_1$ is easy and most of our work will focus on finding $\boldsymbol{a}_2$. From the orthogonality of F we have

$$E[(F\boldsymbol{a})^T\boldsymbol{Y}] = (F\boldsymbol{a})^T\boldsymbol{T}\boldsymbol{\beta} = \boldsymbol{a}_1^T R\boldsymbol{\beta}$$

Note that $E(\hat{Y}_0) = \boldsymbol{T}_0^T\boldsymbol{\beta}$ and so in order for the estimate to be unbiased for all $\boldsymbol{\beta}$ it must follow that $\boldsymbol{a}_1 = (R^T)^{-1}\boldsymbol{T}_0$.

It remains to determine the second component of $\boldsymbol{a}$. Here we minimize the mean squared error of the estimate. To fix ideas we start with the elementary expression for the mean squared error

$$E\left[(f(x_0) - \hat{Y}_0)^2\right] = Var(f(x_0) - \hat{Y}_0) \tag{1.11}$$

because the estimate is unbiased.

$$\begin{aligned} Var(f(x_0) - \hat{Y}_0) &= Var(f(x_0) - \boldsymbol{l}^T\boldsymbol{Y}) & (1.12)\\ &= Var(f(x_0)) - 2\rho\boldsymbol{l}^T\boldsymbol{k}_0 + \boldsymbol{l}^T(\rho K + W^{-1})\boldsymbol{l} & (1.13)\end{aligned}$$

where $\rho K_{ij} = Cov(h(\boldsymbol{x}_i), h(\boldsymbol{x}_j)) = \rho k(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $\rho\boldsymbol{k}_{0,j} = Cov(h(\boldsymbol{x}_0), h(\boldsymbol{x}_j)) = \rho k(\boldsymbol{x}_0, \boldsymbol{x}_j)$. The covariance matrix of the observations is $\rho K + \sigma^2 W^{-1}$.

We now switch to the notation where $\boldsymbol{l}$ has been partitioned.

$$\begin{aligned} E\left[(f(x_0) - \hat{Y}_0)^2\right] &= E\left[(Z_1 + Z_2)^2\right] & (1.14)\\ &= E(Z_1^2) + 2E(Z_1Z_2) + E(Z_2^2) & (1.15)\\ &= Var(Z_1) + 2Cov(Z_1, Z_2) + Var(Z_2) & (1.16)\end{aligned}$$

where $Z_1 = \boldsymbol{T}_0^T\boldsymbol{\beta} - \boldsymbol{a}_1^T F_1^T\boldsymbol{Y}$, $Z_2 = h(x_0) - \boldsymbol{a}_2^T F_2^T\boldsymbol{Y}$ and using the fact that $E(Z_1) = E(Z_2) = 0$.

The remaining part of the derivation is easy. We will just identify 1.16 as a quadratic function of $\boldsymbol{a}_2$, and set the partial derivatives equal to zero. The solution to this linear system defines the minimizer. Note that in this process it is not necessary to work out the term for $Var(Z_1)$ because it does not depend on $\boldsymbol{a}_2$.

$$Cov(Z_1, Z_2) = -\boldsymbol{a}_1^T F_1^T\rho\boldsymbol{k}_0 + \boldsymbol{a}_1^T F_1^T(\rho K + \sigma^2 W^{-1})F_2\boldsymbol{a}_2.$$

and

$$Var(Z_2) = Var(h(x_0)) - 2\boldsymbol{a}_2^T F_2^T\rho\boldsymbol{k}_0 + \boldsymbol{a}_2^T F_2^T(\rho K + \sigma^2 W^{-1})F_2\boldsymbol{a}_2$$

Taking (vector) partial derivatives yields

$$\frac{\partial}{\partial\boldsymbol{a}_2}E[(Z_1+Z_2)^2] = 2F_2^T(\rho K + \sigma^2 W^{-1})F_1\boldsymbol{a}_1 - 2F_2^T\rho\boldsymbol{k}_0 + 2F_2^T(\rho K + \sigma^2 W^{-1})F_2\boldsymbol{a}_2 \quad .$$

Now divide this equation by ρ and set $\lambda = \sigma^2/\rho$. Solving for the parameters when the partials are zero yields the solution

$$\boldsymbol{a}_2 = \left(F_2^T(K+\lambda W^{-1})F_2\right)^{-1}\left(F_2^T\rho\boldsymbol{k}_0 - F_2^T(K+\lambda W^{-1})F_1\boldsymbol{a}_1\right)$$

Recall that the solution for the first component has already been found to be $\boldsymbol{a}_1 = (R^T)^{-1}T_0$ and so $\boldsymbol{l}$ has been determined.

This solution is complicated but still much can be gleaned from its form. At this point the benefit of carrying along the σ^2 and ρ in the derivation is apparent. The estimator only depends on their *ratio*, λ, not on the separate values. Considering the equations for the two components of $\boldsymbol{a}$ it follows that

$$\hat{f}(\boldsymbol{x}_0) = \boldsymbol{l}^T\boldsymbol{Y} = T_0^T\hat{\boldsymbol{\beta}} + \boldsymbol{k}_0^T\hat{\boldsymbol{\delta}} = \sum_{j=1,t}\phi_j(\boldsymbol{x}_0)\hat{\boldsymbol{\beta}}_j + \sum_{i=1,n}\psi_i(\boldsymbol{x}_0)\hat{\delta}_i$$

where the basis functions, $\{\psi_i\}$ are defined in terms of the covariance function: $\psi_i(\boldsymbol{x}) = Cov(h(\boldsymbol{x}), h(\boldsymbol{x}_i))$. Here both $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\delta}}$ are linear, albeit complicated, combinations of $\boldsymbol{Y}$ that do *not* depend on $\boldsymbol{x}_0$. Thus the Kriging estimate can be interpreted as the sum of basis functions where the coefficients are linear functions of the observed $\boldsymbol{Y}$. Taking this a step further, there is a "hat" matrix $A(\lambda)$ like the one for splines such that $\hat{\boldsymbol{Y}} = A(\lambda)\boldsymbol{Y}$.

1.3.2 Kriging estimates are a type of spline

Because of similar notation, the astute reader may have already anticipated another key result for this estimate. Symbolically the form of the spatial process estimator is identical to the spline estimator presented in the previous section. This fact, is established by comparing the estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ implied in the equations with those in Section 2, specifically, 1.10. Therefore, the Kriging estimator can be characterized as the solution to the minimization problem at 1.6 provided that the right identifications are made for the basis functions and the K matrix. Because of this equivalence the cross-validation formulas presented in Section 3 are also valid for the Kriging estimate. This characterization, in the form of a minimization problem, suggests a route to computing spatial process estimates for large data sets. The minimization of 1.6 can be solved using iterative numerical techniques, such as conjugate gradient methods.

1.3.3 Thin-plate splines as spatial process estimators

If Kriging estimates also solve a variational problem, going in the other direction, perhaps a thin-plate spline is just a spatial process estimate? The matrix K derived from radial basis function is not itself a covariance matrix. In fact its diagonal elements are zero! However, following the discussion in Wahba (1990a) one can construct a covariance function of the form

$$k(\boldsymbol{x}, \boldsymbol{x}') = E_m(||\boldsymbol{x}, \boldsymbol{x}'||) + Q(\boldsymbol{x}, \boldsymbol{x}')$$

The modification, Q is symmetric and involves combinations of the radial basis functions and polynomials up to degree $m-1$. It also has the following important property: if $\boldsymbol{u}$ is any vector orthogonal to the columns of T, then

$$\sum_{j=1}^{n} Q(\boldsymbol{x}, \boldsymbol{x}_j)\boldsymbol{u}_j$$

is a polynomial of degree $m-1$ or less in $\boldsymbol{x}$.

Based on this covariance there is of course a corresponding Kriging estimate. Furthermore one can show that this estimate reduces to the usual thin-plate spline (Wahba 1990a). Although the argument will not be reproduced here the basic ingredient is that K and $\boldsymbol{k}_0$ enter all the expressions for $\boldsymbol{a}_2$ post multiplied by F_2^T. The columns of F_2 are of course orthogonal to T and so the terms only differ from the radial basis function kernel by a polynomial term of degree $m-1$ or less. These polynomial extra terms are absorbed into the null space term.

After going to the trouble to construct a valid covariance function from the radial basis function kernel ($E_m(||\boldsymbol{x}-\boldsymbol{x}'||$), we see that the modification, (i.e. Q), has no impact on the estimator. One can use the radial basis functions alone, pretending that it is a valid covariance, and one will recover the same estimate from the Kriging formulas. In fact, the radial basis function kernel does act like a covariance kernel for certain linear combinations of h. This stochastic model has a rigorous foundation stemming from the work of Matheron (1973) on intrinsic random functions and the reader is referred to this work and Wahba (1990a) for the development of this theory. However, in summary, a spline can be identified as a spatial process estimate with a fairly peculiar covariance function.

1.3.4 Profiling the likelihood function for parameters of the covariance function

The QR decomposition also has use in simplifying the likelihood function in order to estimate parameters of the covariance function. Assuming a Gaussian distribution for h and $\boldsymbol{\epsilon}$ in equations 1.1 and 1.2, $F_2^T Y$ will be multivariate normal with mean zero and covariance matrix, $F_2^T(\rho K + \sigma^2 W^{-1})F_2$. Although a realistic spatial covariance may depend on several parameters we will emphasize some reductions for the parameters that are linear in the covariance. Recall that we have reparameterized the covariance so that $\sigma^2 = \lambda\rho$. The log likelihood for λ and ρ up to a constant term is

$$l(\lambda, \rho) = -(1/2)log(|M|\rho^{n-t}) - \frac{Y_2^T(M)^{-1}Y_2}{2\rho} + C$$

where $M = F_2^T(K + \lambda W^{-1})F_2$ and $Y_2 = F_2^T Y$. C is a generic symbol for an expression that does not depend on ρ or λ. Maximizing over ρ for fixed λ gives

$$\hat{\rho} = Y_2^T(M)^{-1}Y_2/(n-t)$$

and substituting this back into the likelihood, we have up to a constant term,

$$l(\lambda,\hat{\rho}) = -(1/2)log\left(\left(Y_2^T(M)^{-1}Y_2\right)^{n-t}|M|\right) + C$$

A useful detail from the discussion of equation 1.9 is that the residual vector, $(I - A(\lambda))\boldsymbol{Y}$, is equal to $\lambda W^{-1}\hat{\delta}$. Now going back to equation 1.10 and using the fact that $\hat{\boldsymbol{\delta}} = F_2\hat{\boldsymbol{\omega}}_2$ we are lead to the relationship that

$$\lambda W^{-1}F_2M^{-1}F_2^T = I - A(\lambda)$$

Furthermore, one can show that $|M^{-1}|$ is equal to the product of the nonzero eigenvalues of $I - A(\lambda)$. Denote this determinant by $|I - A(\lambda)|_+$ and so we have found the profile likelihood with respect to λ

$$l(\lambda,\hat{\rho}) = -(1/2)log\left(\frac{\left(\boldsymbol{Y}^T(I - A(\lambda))\boldsymbol{Y}\right)^{n-t}}{|I - A(\lambda)|_+}\right) + C$$

Implicit in these expressions is the possibility that the covariance function, k, may depend on other parameters besides the leading multiplier ρ. For example, a useful covariance family analogous to those in the design and analysis of computer experiments (Sacks et al. 1989 and Welch et al. 1992) is

$$k(\boldsymbol{x},\boldsymbol{x}') = \rho e^{-\left(\frac{||\boldsymbol{x}-\boldsymbol{x}'||}{\theta}\right)^p} \tag{1.17}$$

where θ controls the range of correlations and $1 \leq p \leq 2$ determines the smoothness of the random process. Parameters such as the range and smoothness in the example above enter in an inherently nonlinear manner. There is a rich literature in estimating the parameters of the covariance function and the treatment in Cressie (1991) is a fine introduction. However, given the emphasis in this chapter we will see that it is possible to rapidly evaluate the likelihood or GCV function for different values of λ, holding other parameters in the covariance fixed. This separation suggests that the nonlinear parameters such as p and θ in 1.17, could be estimated separately, say, by nonlinear least squares fitting of a variogram. This would be followed by estimating ρ and λ using maximum likelihood with the formulas derived above. In practice, spatial estimates are often much more sensitive to the choice of λ than to the choice of range or smoothness parameters.

1.3.5 Measures of uncertainty in the estimates

One advantage of the spatial process model is that it provides measures of uncertainty for the function estimates. The equations for the prediction variance at 1.13 hold for any unbiased prediction of $f(\boldsymbol{x}_0)$. So the prediction mean squared error can be found by substituting the specific $\boldsymbol{l}$ into this equation. Furthermore, the prediction variance associated with the minimum variance estimator yields some simplifications. However, in practice there is the additional uncertainty in the estimate due to estimating ρ, σ and other covariance parameters. For a more comprehensive treatment including such parameters, the reader is referred to Handcock and Stein (1993), Qian (1997) and Pilz et al. (1997).

The uncertainty measures based on a random field can be very different from the computed *variance* of a thin-plate spline. The variance associated with the estimate of a fixed function is just $\sigma^2 \boldsymbol{l}^t W^{-1} \boldsymbol{l}$, just one part of the expression for mean squared error at 1.13. (The remaining part will also be nonnegative because it is the mean squared error for a estimate when $\sigma = 0$.)

The variance term alone can be substantially smaller than the actual uncertainty especially when σ is close to zero. This is not a defect in the spline estimate but an inadequate accounting of the estimates error. The problem is that variance alone only takes into account the variability in fitting the spline to the sample. The other component of error is approximating the true function in terms of a spline. As a complement to the spatial process model, we will illustrate this problem in the context of f as a fixed function. In this case the spline will be a biased estimate of f. It is helpful to study the bias by breaking it into two components. Let f_S be the thin-plate spline that interpolates f at the data points. The bias is

$$f(\boldsymbol{x}) - E(\hat{f}(\boldsymbol{x})) = f(\boldsymbol{x}) - f_S(\boldsymbol{x}) + f_S(\boldsymbol{x}) - E(\hat{f}(\boldsymbol{x}))$$

The first term will only be zero if $\boldsymbol{x}$ is a data point. This first term is the interpolation error and will only be zero for all $\boldsymbol{x}$ if the unknown function is *exactly* a spline. A highly unlikely prospect! The second term will be zero if the spline interpolates the data ($\lambda = 0$) but nonzero if some smoothing is done $\lambda > 0$. Thus this second term can be thought of as the bias incurred by smoothing. Both of these biases can cause error in the spline above the variance. In the extreme case of interpolation, the variance of the spline estimate is zero. But never-the-less there may exist substantial interpolation error when the estimate is evaluated at points where observations have not been made. One approach to determining the bias is to use bounds based on the approximation properties of splines. Another approach is to fall back on the equivalent spatial model associated with thin-plate splines and derive uncertainties using the generalized covariance. This later case is in the spirit of Wahba's “ Bayesian ”confidence intervals for splines and is a practical method even for fixed function inference (Nychka 1988).

1.4 Ridge regression estimates and shrinkage

Both the thin-plate spline and the spatial process estimators have a common form as ridge regressions. This general form is at the heart of the numerical procedures for both methods and leads to the construction of the orthogonal series and efficient formulas for determining the smoothing parameter, λ.

Assume the model

$$Y = X\omega + \boldsymbol{e}$$

and let H be a nonnegative definite penalty matrix. The general form of the penalized (and weighted) least squares problem is minimization of

$$(Y - X\omega)^T W (Y - X\omega) + \lambda \omega^T H \omega \tag{1.18}$$

over $\omega \in \Re^n$. The solution is

$$\omega = (X^T W X + \lambda H)^{-1} X^T W Y \quad .$$

1.4.1 A natural basis

Given this general framework, the specific applications to thin-plate splines and spatial process estimates involve identifying the relevant parameterization to find the X matrix and the penalty matrix, H. In both cases the constraint on the parameters, $T^T\boldsymbol{\delta} = 0$, provides a route to reduce the $t + n$ basis functions to just n. Denote this basis as $\{\eta_i\}$ where the first t functions are equal to the polynomial terms, $\{\phi_i\}$ and the remaining $n - t$ terms are linear combinations of $\{\psi_j\}$ based on the columns of F_2. Explicitly we have $\eta_{i+t} = \sum_{k=1}^{n} F_{2k,i}\psi_i$ for these last members of the basis. We will refer to this as the *natural* basis because it automatically builds in the constraint necessary for a solution. A more practical feature is that we have reduced an over-determined linear system with linear constraints to a system with as many equations as parameters. To the general reader there is really nothing "natural" about this basis but it takes its name from the properties in the one-dimensional case. In the case of a one dimensional, mth order spline this constraint results in the higher derivatives $j = m, m+1, ...2m-1$ of the spline being zero at the boundaries. These can be interpreted in a physical system as free or natural boundary conditions. In the case of cubic smoothing splines, $(d = 1, m = 2)$, the estimate will be a linear function at the endpoints of data and thus extrapolates as a linear function beyond the range of the observed data.

Given the natural basis, the representation of the estimate is $f(\boldsymbol{x}) = \sum_{i=1}^{n} \omega_i \eta_i(\boldsymbol{x})$ thus the relevant regression matrix is $X_{ij} = \eta_j(\boldsymbol{x}_i)$, and the associated penalty matrix is

$$H = \begin{pmatrix} 0 & 0 \\ 0 & F_2^T K F_2 \end{pmatrix}$$

1.4.2 Demmler-Reinsh basis

The last step is to consider a linear transformation of the natural basis to produce an orthogonal one. This is accomplished by finding a matrix, G, that will diagonalize both X^TWX and H. Let B denote the inverse square root of X^TWX and let UDU^T be the singular value decomposition of BHB^T. Now set $G = U^TB$ and it is straightforward to verify that $G^T(X^TWX)G = I$ and $G^T(H)G = D$, where D is diagonal. Finally we define the new, orthogonal basis as

$$g_\nu(\boldsymbol{x}) = \sum_{i=1,n} G_{i\nu}\eta_i(\boldsymbol{x})$$

This basis, known as the Demmler-Reinsch (DR) basis in the context of splines. Assume that $f(\boldsymbol{x}) = \sum_{\nu=1}^n \alpha_\nu g_\nu(\boldsymbol{x})$ We have the following properties for this basis.

1. $\{g_\nu\}$ spans the same subspace of functions as the natural basis, $\{\eta_\nu\}$
2. $$\sum_{i=1}^n g_\nu(\boldsymbol{x}_i)W_i g_\mu(\boldsymbol{x}_i) = \begin{cases} 0 & \mu \neq \nu \\ 1 & \mu = \nu \end{cases}$$

   and so

   $$\sum_{i=1}^n f(\boldsymbol{x}_i)^2 W_i = \sum_{\nu=1}^n \alpha_\nu^2$$

3. Setting $\boldsymbol{\omega} = G\boldsymbol{\alpha}$

   $$f(\boldsymbol{x}) = \sum_{\nu=1}^n \omega_\nu \eta_\nu(\boldsymbol{x})$$

   and

   $$\omega^T H \omega = \boldsymbol{\alpha}^T D \boldsymbol{\alpha} = \sum_{\nu=1}^n D_\nu \alpha_\nu^2.$$

These properties can be easily proved from the construction of the transformation matrix, G. An important consequence from the second property is a simple representation for the interpolating function. Given the sequences $(\boldsymbol{x}_i, Y_i)$ for $1 \leq i \leq n$, let $u_\nu = \sum_{i=1,n} g_\nu(\boldsymbol{x}_i)W_iY_i$. The function $\sum_{\nu=1}^n u_\nu g_\nu(\boldsymbol{x})$ will interpolate these data.

1.4.3 Simplifications due to the Demmler-Reinsch basis

Using the parameterization with respect to the Demmler-Reinsch basis, assume that $f(\boldsymbol{x}) = \sum_{\nu=1}^n \alpha_\nu g_\nu(\boldsymbol{x})$ and let $\boldsymbol{u}$ be the coefficients that interpolate the data pairs $(\boldsymbol{x}_i, \boldsymbol{Y}_i)$.

$$S_\lambda(f) = \sum_{\nu=1}^n (u_\nu - \alpha_\nu)^2 + \lambda \sum_{\nu=1}^n D_\nu \alpha_\nu^2$$

Note that now the parameters are decoupled and each term can just be minimized independently to give the solution

$$\hat{\alpha}_\nu = \frac{u_\nu}{1 + \lambda D_\nu}$$

leading to the function estimate presented in the introduction.

The DR basis also yields simple expressions for the residual sums of squares and the trace of the hat matrix. The i^{th} residual is

$$Y_i - \hat{f}(\boldsymbol{x}_i) = \sum_{\nu=1}^{n} \left(u_\nu - \frac{u_\nu}{1+\lambda D_\nu} \right) g_\nu(\boldsymbol{x}_i) = \sum_{\nu=1}^{n} \frac{\lambda D_\nu u_\nu g_\nu(\boldsymbol{x}_i)}{1+\lambda D_\nu}$$

Using the orthogonality properties it follows that

$$\sum_{i=1}^{n} W_i (Y_i - \hat{f}(\boldsymbol{x}_i))^2 = \sum_{\nu=1}^{n} \left(\frac{\lambda D_\nu u_\nu}{1+\lambda D_\nu} \right)^2$$

The hat matrix has elements

$$A(\lambda)_{ij} = \sum_{\nu=1}^{n} \frac{g_\nu(\boldsymbol{x}_i) g_\nu(\boldsymbol{x}_j)}{1+\lambda D_\nu} W_j$$

Again, using the orthogonality property it follows that the trace of this matrix is

$$trace(A(\lambda)) = \sum_{\nu=1}^{n} \frac{1}{1+\lambda D_\nu} \quad .$$

In each case, once the DR basis has been computed, the residual sums of squares and the trace can be evaluated rapidly in $O(n)$ operations. Provided that the independent variables are fixed, this feature makes it efficient to estimate the function, to evaluate the cross-validation function, or to evaluate the profiled likelihood for different realizations of the dependent variable.

One important application of the DR form is in the context of resampling. For example, to test for the presence of the nonparametric smooth component or to quantify the uncertainty in estimating λ (see Nychka 1991), it is necessary to use a bootstrap for approximating the relevant sampling distribution. Under the assumption of normal errors, the interpolation coefficients, $\boldsymbol{u}$ can be generated directly. This eliminates the need for re-evaluating the basis functions and makes all the computations possible with respect to the DR coefficients. Let Γ denote the matrix with elements $\Gamma_{\nu,i} = g_\nu(\boldsymbol{x}_i) W_i$ so that $\boldsymbol{u} = \Gamma \boldsymbol{Y}$. Under the Gaussian assumptions, $\boldsymbol{Y}$ will be multivariate normal and thus so will $\boldsymbol{u}$. If f is assumed to be a *fixed* function then $E(\boldsymbol{u}) = \boldsymbol{\alpha}$ and $Cov(\boldsymbol{u}) = \Gamma(\sigma^2 W^{-1})\Gamma^T = \sigma^2 I$. Thus $\boldsymbol{u} \sim MN(\boldsymbol{\alpha}, \sigma^2 I)$. In the case of the spatial statistics model with a random function, it is easy to verify that $\{u_\nu\}$ are independent, zero mean normal random variables with variances $\{\rho D_\nu + \sigma^2\}$.

1.4.4 Reducing the number of basis functions

One advantage of this ridge regression formalism is that it is easy to see how to reduce the size of the basis but still retain the qualitative, statistical features of these smoothers. There is often strong incentive for doing this. The storage requirements for the estimate involve matrices that are of size $(n-t) \times (n-t)$. Moreover, the operation count for computing the estimate [2] for a fixed λ is $O(n^2)$ and to find the orthogonal basis [3] it increases to $O(n^3)$.

Thus the computation burden increases rapidly as the sample size is increased and it is difficult to fit these estimates to data sets having more than several hundred unique locations. Working with the complete basis yields estimates with the highest possible spatial resolution. However, for situations where the function does not have finely detailed structure and when measurement error is present one does not need the full resolution provided by an expansion using all of the basis functions. If the weighting function in equation 1.4 is nearly zero for large values of $\{D_\nu\}$ it is not clear why these terms even need to be computed. Thus it seems reasonable in some cases to work with a truncated basis.

A natural strategy is to restrict the size of the basis independently of the sample size. Start with the full basis $\{\phi_i\}$ and $\{\psi_j\}$ and consider a subset of these functions of size N that will give a good approximation to the full expansion. Because $\{\psi_j\}$ are generated from the observed locations one principle is to select a subset derived from locations that are widely separated. One automatic way to accomplish this is by selecting the subset of $\{\boldsymbol{x}_k\}$ based on a space-filling design criterion. The FUNFITS manual gives an example of this strategy.

Even if a subset of the natural basis functions is used, the ridge regression development is largely unchanged and one will still obtain an orthogonal representation with the same properties. But there are two differences that should be noted. The index for the basis, ν is restricted to $[1, N]$ with $N < n$. Thus X will be $n \times N$ and H is $N \times N$. Also, the residual sums of squares should include another term associated with the estimate when $\lambda = 0$, $\hat{f}_0$. This is just the result from the (weighted) regression of $\boldsymbol{Y}$ on X.

1.5 A response surface example

To help interpret the orthogonal series estimates the DR basis and the weighting functions will be computed for a specific example using thin-plate splines. It is informative to compare the resulting spline to a kriging estimate and also investigate the effect of changing the null space. The main goal is to give some intuition behind the constructions presented in the previous sections.

[2] Because one must solve an $n \times n$ linear system.

[3] Because one must perform an eigenvalue/eigenvector decomposition on a symmetric matrix of size $n \times n$.

1.5.1 Demmler Reinsch basis functions for the flame data set

The data set `flame` is a two-dimensional response surface and is included as part of FUNFITS. These measurements come from an experiment in analytical chemistry to relate the intensity of light from the ionization of Lithium ions in aqueous solution to variables controlling the ionizing flame. The independent variables are the flow rates of the fuel and and the oxygen supplied to the burner used to ionize the sample. The data consists of 89 different settings of these two flow rates.

Figure **1** illustrates several of the 89 DR basis functions associated with a thin-plate spline with a third derivative roughness penalty ($m = 3$). The basis functions are in order of increasing D_ν and thus the first six have zero eigenvalues and span the subspace of second degree polynomial functions in $\Re^2$. Basis functions beyond the null space terms exhibit increasing numbers of modes and informally we equate basis functions with larger D_ν with higher spatial frequencies. This leads to associating these higher terms with more detailed structure in the surface. Figure **2** illustrates how the coefficients are filtered to produce a smooth surface. The top plot indicates the weight sequence $1/(1 + \lambda D_\nu)$ for three values of λ and the middle plot indicates the effect of tapering the interpolating coefficients, $\boldsymbol{u}$ using these weights. Note that the first six weights are always one, indicating the fact that no smoothing is done on the quadratic null space. The resulting smoothed surfaces and one that interpolates the data are displayed in the first column of plots in Figure **3**. The basic feature of this response surface is a broad ridge that is oriented obliquely to the independent variables. There is also some finer structure along the ridge line.

As mentioned in Section 2 the amount of smoothing can be estimated from the data using GCV and Figure **4** gives some diagnostic plots from fitting splines with different size null spaces ($m = 2, 3, 4$) and roughness penalties. It appears that when $m = 4$, a cubic null model, the spline estimate has a slightly lower GCV value than $m = 3$ and fewer degrees of freedom associated with the fitted surface. A smaller number of degrees of freedom might be preferable because it gives a more parsimonious representation of the response surface.

1.5.2 Comparison with a spatial process estimate

An exponential covariance model (equation 1.17 with $p = 1$) was also estimated for these data using cross-validation for λ but fixing the range parameter, θ at 5 (Figure **5b**). This was done largely for comparison. In fact, using cross-validation to choose θ suggests a much larger value that will make the estimator very similar to a spline. Even with this smaller value for the range parameter, there is little difference between the spatial process estimate and the thin-plate spline. However, if we consider measures of uncertainty, the prediction variance of the spline is substantially different. This is due to the fact that under the kriging model one obtains a much smaller estimate for σ. The two image plots in Figure **6** illustrate these prediction variances over the design region. The spline variances form a smooth surface reflecting an estimate that smooths the data rather than interpolating. The spatial process estimate is much closer to an interpolatory surface and so prediction variances will be nearly zero at the observed points.

1.5.3 A multiresolution analysis

There has been much recent interest in using wavelet and other related basis functions for nonparametric regression and smoothing (see for example Nason and Silverman 1994 and Chapter 7 in this volume). One contribution from a wavelet analysis is the decomposition of the estimated function into a sequence of curves that represent the structure at different levels of resolution. This is easy in the case of wavelets (for equally spaced data) because the basis functions themselves naturally fall into a hierarchy of resolution levels. The orthogonal series representation also lends itself to a multi-resolution analysis. While it is not as precise or compact as the wavelet representation, it does extend to irregularly spaced observations and serves as a useful summary of how the estimated surface changes as a function of the smoothing parameter.

One simple way to accomplish this goal is to consider differences of spline estimates for a decreasing sequence of smoothing parameters. Let $\{\lambda_j\}$, $0 \leq j \leq K$ be a decreasing sequence of smoothing parameters where $\lambda_0 = \infty$ and $\lambda_K = 0$. Let $\hat{d}_0(\boldsymbol{x}) \equiv \hat{f}_{\lambda_0}(\boldsymbol{x})$ and for $j \geq 1$

$$\hat{d}_j(\boldsymbol{x}) = \hat{f}_{\lambda_j}(\boldsymbol{x}) - \hat{f}_{\lambda_{(j-1)}}(\boldsymbol{x})$$

These differences, $\hat{d}_j$, will be referred to as the detail surfaces because they capture the detail between the estimates at two different levels of smoothing. By the telescoping property of the differences, $Y_i = \sum_{j=0}^{K} \hat{d}_j(\boldsymbol{x}_i)$.

In terms of the orthogonal series representation

$$\hat{d}_j(\boldsymbol{x}) = \sum_{\nu=1}^{n} \frac{D_\nu(\lambda_j - \lambda_{j-1})}{(1 + D_\nu\lambda_j)(1 + D_\nu\lambda_{j-1})} u_\nu g_\nu(\boldsymbol{x}) = \sum_{\nu=1}^{n} w_{\nu,j} u_\nu g_\nu(\boldsymbol{x})$$

so we see that this difference is just applying a different type of weighting to the interpolation coefficients. From a signal processing / image analysis perspective the spline estimate is a low pass filter applied to the data. Thus, the difference of two spline smooths should behave as a bandpass filter[4]. The third plot in Figure **2** illustrates the band pass weights for a sequence of smoothing parameters chosen to have effective model degrees of freedom 8, 43.5 and 60. The smoothing parameter sequence has been chosen so that there is minimal overlap among the weights for the detail surfaces. Because the interpolation coefficients will be independent under

[4] A filter that passes components in a range of frequencies but omits frequencies that are either very low or very high. A bandpass filter can be constructed by the appropriate combination of a low and high pass filter.

a Gaussian model for the errors, the different resolution levels will only be weakly correlated with each other. The second and third columns in Figure **3** illustrate the sequence of detail surfaces for the `flame` data. (Note that the image plots are not on a common color scale.) Most of the structure is well explained by the low resolution surface. The isolated local maxima along the ridge, however, are features that only emerge at the highest resolution levels, but these are small adjustments to the overall estimate.

1.6 Predicting Ambient Ozone

The FUNFITS dataset `ozone2` contains the daily, 8-hour average ozone measurements during the summer of 1987 for approximately 160 locations in the Midwest (Figure **7**). One use for these data is to compare these observations to the output from a large numerical model for predicting surface ozone (Regional Oxidant Model). In order to carry out this validation, it is important to estimate the ozone concentrations from the observational data at the regular grid points matching with the model output. For the purposes of this chapter we will focus on the 16th day in this data set (June 19, 1987) and compare estimates based on a thin-plate spline and a more comprehensive spatial model.

The ozone field is not a stationary spatial process. This conclusion is not only based on the physical theories for the formation and transport of ozone but is also supported by the observed data. The covariance function for the ozone field can be estimated empirically under the assumption that each day is an independent realization of the random field. For example, in this data set we have 89 separate measurements for each station[5].

One simple departure from stationarity is to consider a covariance that has different marginal variances but isotropic correlations. Specifically consider

$$COV(h(\boldsymbol{x}), h(\boldsymbol{x}')) = \rho w(\boldsymbol{x}) w(\boldsymbol{x}') \left(e^{||\boldsymbol{x}-\boldsymbol{x}'||/\theta} \right)$$

and a nugget (or measurement error) variance with $Var(e_k) = \sigma^2 w(\boldsymbol{x}_k)$. As a first step we can estimate the marginal standard deviation surface, $w(\boldsymbol{x})$, using a thin-plate spline fit to the sample standard deviations for each location over the 89 days of the summer. In the covariance model $w(\boldsymbol{x})$ is taken to be this spline estimate and ρ and θ are found by GCV. This covariance model has been set up so that if the marginal variance on this day agrees with the average variance found for the summer $\sigma^2 + \rho = 1$. Day 16 is apparently different from the summer average and the estimated variability is larger than expected. The estimated ozone surface estimate using a thin-plate spline and the spatial process estimate are depicted in Figure **8**. Note that even though these estimates are based on very different covariance

[5]Actually some days are missing for particular stations and so 89 is an upper bound on the number of observations.

structures the surfaces are very similar. The standard errors for these estimates are different as we see in Figure **9**. The the spline has much lower standard errors. This is due to the difference between the generalized covariance associated with the spline and the exponential covariance. In particular the range parameter (θ) can limit the extent of the spatial correlations in the exponential model, implying a rougher surface. In this case, the spatial estimate will have a larger standard error. This example suggests that for inference, one should be careful about modeling the covariance of the surface and evaluate the sensitivity to different models. Forced to choose, the more general spatial model probably gives more believable measures of uncertainty.

1.7 Future Directions

All of the smoothing methods discussed in this chapter use a single global smoothing parameter and so every part of the data is smoothed an equal amount. This constraint may miss local features in data and may cause inferences about the estimate to be unreliable. Wavelet and multiresolution ideas provide one way of local smoothing and the example in this chapter is one possible implementation for splines. An alternative is to select the basis functions that define the estimate. The idea is to concentrate basis functions in areas where the function has fine structure and eliminate basis functions where the function does not have features (see for example Luo and Wahba 1997). In the spatial context local smoothing is equivalent to a nonstationary covariance function and the modeling and use of nonstationary spatial covariances is an area of very active research.

Another important area is in nonparametric design. In the context of design for computer experiments there has been some work on locating points that are efficient for estimating a class of functions and not just a parametric family. There are also opportunities for design in response surface problems and in the more traditional area of spatial sampling.

In a regression problem one would like measures of the contribution of different variables to the predictive power of the model. Also one would like to quantify the importance of the smooth component in the model above the polynomial part. Being able to make inferences about the value of a variable is important in response surface problems and identifying key process variables can sharpen the predictions and also make the models easier to interpret. Testing for the significance of a variable usually results in test statistics where the reference distribution must be determined by simulation and so the computational efficiency of using the DR representation is important. The work by Gu and Wahba (1991) has relevance here because different smoothing parameters for different parts of a model or different variables can indicate the amount of structure associated with each component.

1.8 References

Bates, D.M., Lindstrom, M.J., Wahba, G., and Yandell, B.S. (1987). GCVPACK - Routines for generalized cross-validation. *Communications in Statistics - Simulation and Computation* **16**, 263-297.

Cressie, N.A.C. (1990). Reply to Wahba's letter. *American Statistician*, **44**, 256-258.

Cressie, N.A.C. (1991). *Statistics for Spatial Data.* Wiley, New York.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobelev spaces. In *Constructive Theory of Several Variables*, Berlin, 85-100.

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

Gu, C. and Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM Journal of Scientific and Statistical Computing* **12**, 383-398.

Gu, C. (1998). ????. In *Smoothing and Regression* ed. M. Schimek, Wiley, New York.

Handcock, M. and Stein, M. (1993). A Bayesian Analysis of Kriging. *Technometrics*, **35** 403-410.

Luo, Z. and Wahba, G. (1997). Hybrid adaptive splines, *Journal of the American Statistical Association*, **92** 107-116.

Kent, J.T. and Mardia, K.V. (1994). The link between Kriging and Thin Plate Splines. *Probability, Statistics and Optimization*, ed: F.P. Kelly, Wiley, New York, 324-339.

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

Meinguet, J.(1979). Multivariate interpolation at arbitrary points made simple, *J. Appl. Math. Phys.* (ZAMP), **30**, 292-304.

Messer, K. (1991). A comparison of a spline estimate to its equivalent kernel estimate. *Annals of Statistics* **19**, 817-829.

Nason, G.P. and Silverman, B.W. (1994). The discrete wavelet transform. *Journal of Computational and Graphical Statistics*, **3**, 163-191.

Nychka, D. (1988). Bayesian 'Confidence' Intervals for Smoothing Splines. *Journal of the American Statistical Association*, **83**, 1134-1143.

Nychka, D. (1991). Choosing a range for the amount of smoothing in nonparametric regression. *Journal of the American Statistical Association*, **86**, 653-664.

Nychka, D. (1995). Smoothing Splines as Locally Weighted Averages. *Annals of Statistics*, **23**, 1175-1197.

Nychka, D., Bailey, B., Ellner, S., Haaland, P. and O'Connell, M. (1996). FUNFITS data analysis and statistical tools for estimating functions, *North Carolina Institute of Statistics Mimeoseries* No. 2289

O'Connell, M. and Wolfinger, R. (1997). Spatial regression, response surfaces and process optimization. *Journal of Computational and Graphical Statistics* **6**, 224-241.

Omre H. and Halvorsen K. (1989). The *Bayesian* bridge between simple and universal *Kriging*. *Mathematical Geology*, **21** 767-786.

Pilz, J., Schimek, M.G. and Spöck, G. (1997). Taking account of uncertainty in spatial covariance estimation. In, *Geostatistics Wollongong '96* Baafi and N. Schofield, editors, Kluwer, Dordrecht, 355-358.

Qian, S.S. (1997). Estimation the area affected by phosphorus in an Everglades wetland: a comparison of universal kriging and Bayesian kriging. *Environmental and Ecological Statistics*, **4** 1-29.

Sacks, J., Welch, W.J., Mitchell, T.J. and Wynn, H.P. (1989). Design and analysis of computer experiments. *Statistical Science* **4**, 409-435.

Zhang, B., and Stein, M. (1993). Kernel approximations for universal Kriging predictors. *Journal of Multivariate Analysis* **44**, 286-313.

Wahba, G. (1990a). *Spline Models for Observational Data.* Society for Industrial and Applied Mathematics, Philadelphia.

Wahba, G.(1990b). Letter to the editor. *American Statistician*, **44**, 256.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., and Morris, M.D. (1992). Screening, predicting and computer experiments. *Technometrics.* **34**, 15-25.

1.9 Acknowledgements

The author would like to thank Sarah Hardy and Wendy Meiring for a careful reading of this manuscript and valuable suggestions for its improvement. This work was sponsored by National Science Foundation grants DMS-9705054 and DMS-93122686.

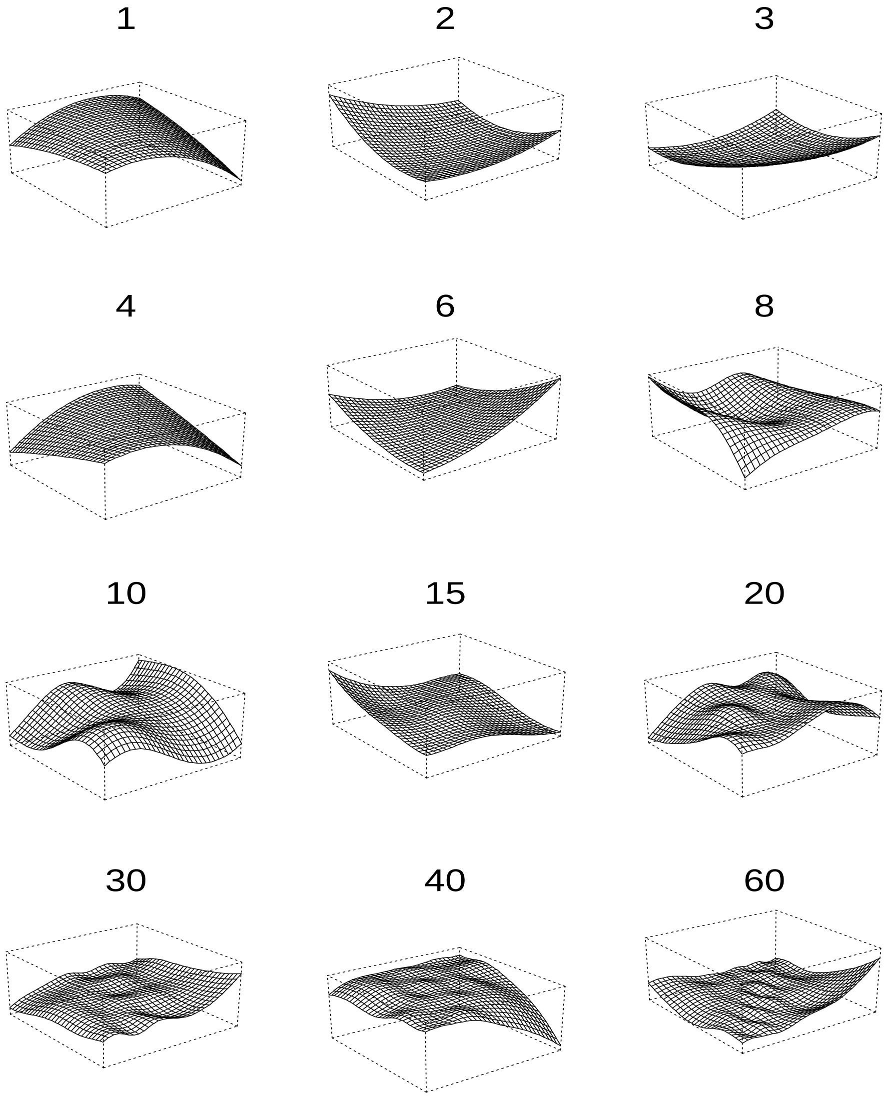

Figure 1.1: Some examples of Demmler-Reinsch basis functions for the flame data. These are based on a thin-plate spline estimate with $m = 3$, a quadratic polynomial nullspace. The numbers above each plot are the index of the associated eigenvalue when ordered from largest to smallest.

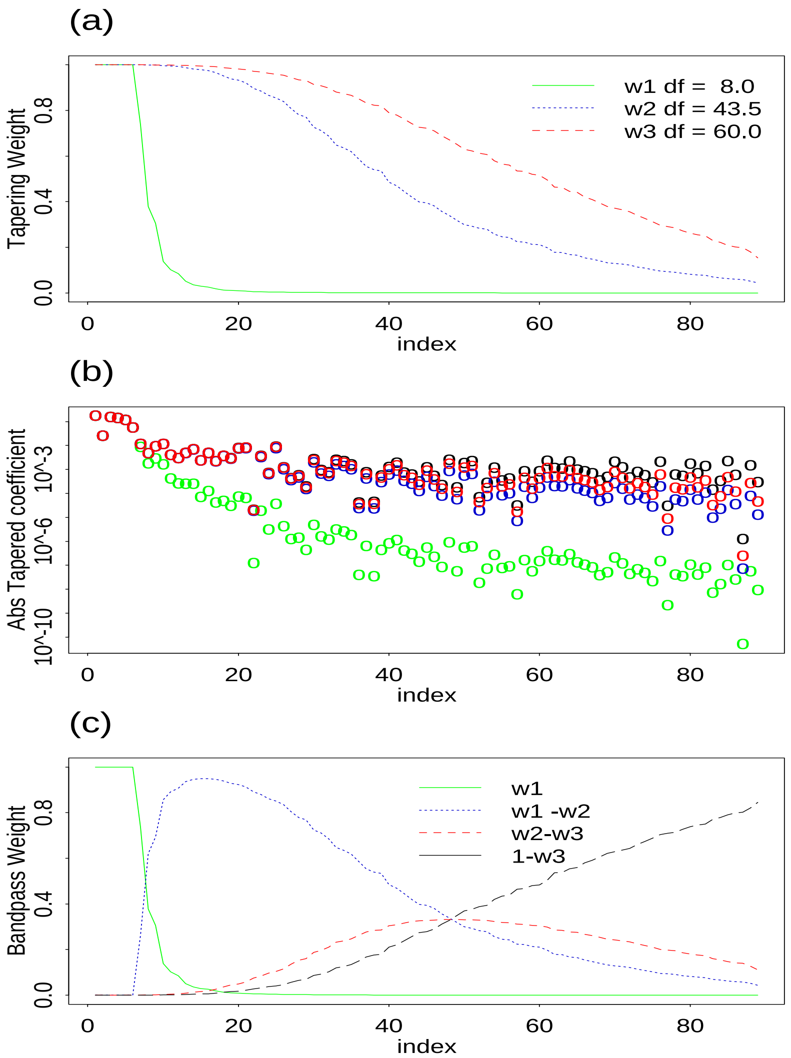

Figure 1.2: Tapering functions for the thin-plate spline fit to the flame data example. Spline taper functions w_1, w_2, w_3 for $m = 3$ are based on the Demmler-Reinsch basis and follow the form $1/(1+\lambda D_\nu)$ for three different choices of λ. The effective number of parameters associated with each smooth ($traceA(\lambda)$) appears in the legend of the first plot (a). In plot (b) are the absolute values of the coefficients $\boldsymbol{u}$ and their tapered values using the weights from (a). The third plot,(c), gives the weights to get the multiresolution decomposition in Figure 3.

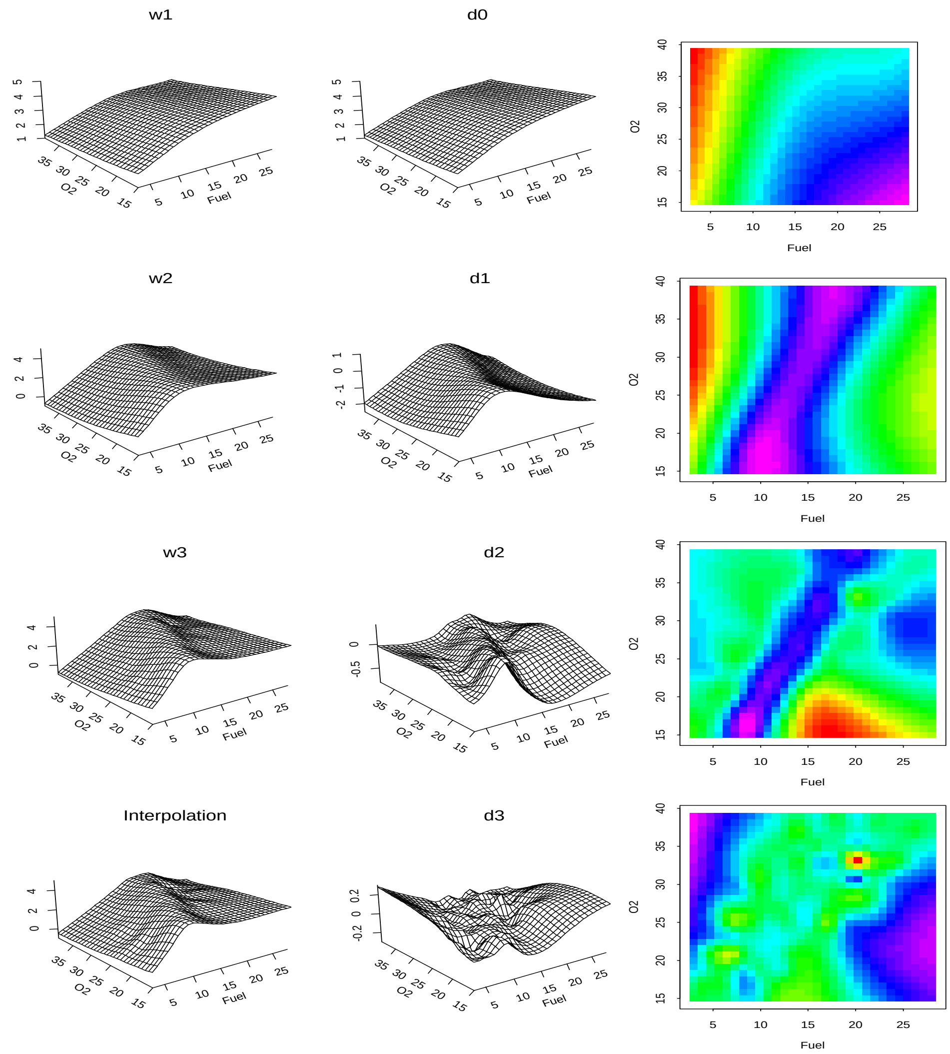

Figure 1.3: Different spline estimates ($m = 3$) and the decomposition of the flame data into several detail surfaces. In the first column are the surfaces derived from the three weight functions plotted in Figure 2 b), with effective model degrees of freedom 8.0, 43.5, and 60.0. The last plot in this column is an interpolated surface. In the second column are the detail surfaces for the multiresolution decomposition described in Section 5.3. The first detail surface is the same as the first smoothed surface. The subsequent surfaces are differences between the smoothed surfaces at two different levels of smoothing. The last column are image plots of the detail surfaces in the second column.

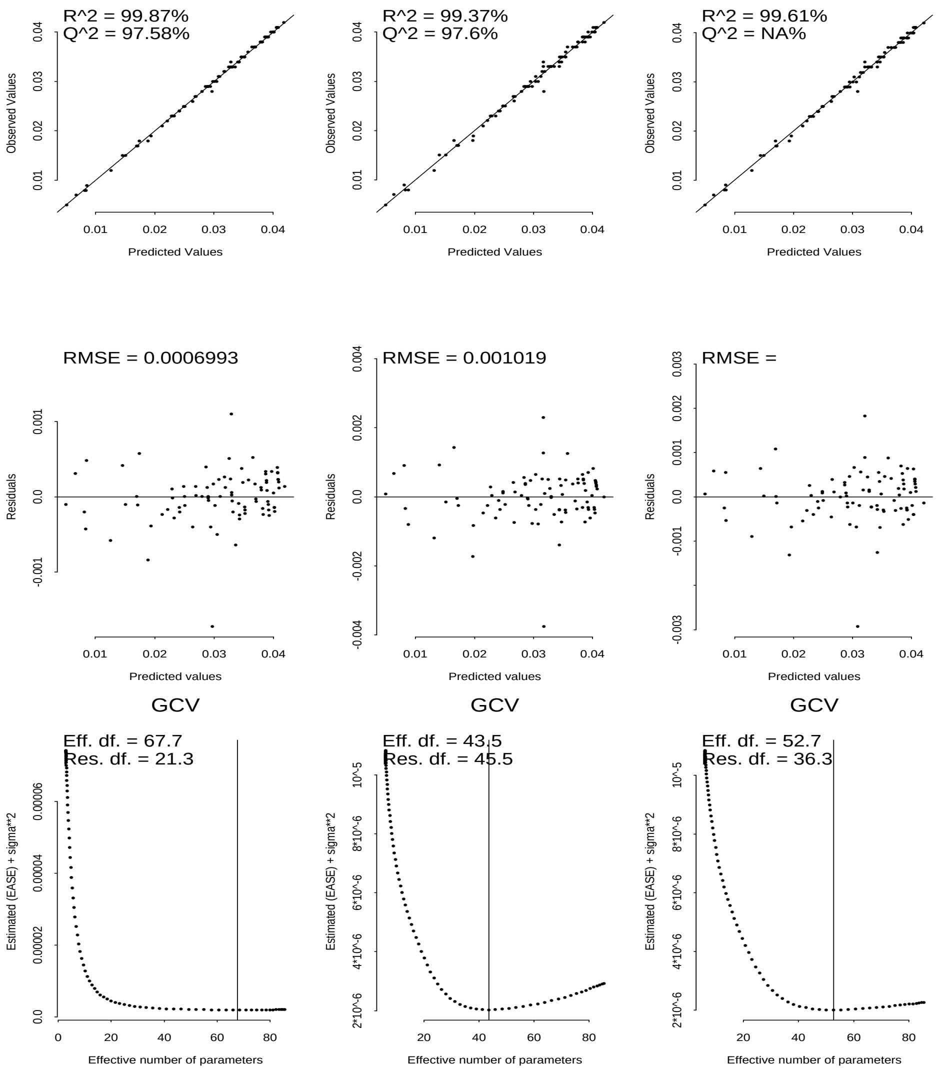

Figure 1.4: Diagnostic plots for thin-plate splines fit to the flame data using $m = 2$ (first column), $m = 3$ (second column) and $m = 4$ (third column). In the first row are scatter plots of predicted and fitted values, in the second row are residuals against predicted values, and the third row contains the GCV criteria for each model plotted against the model degrees of freedom.

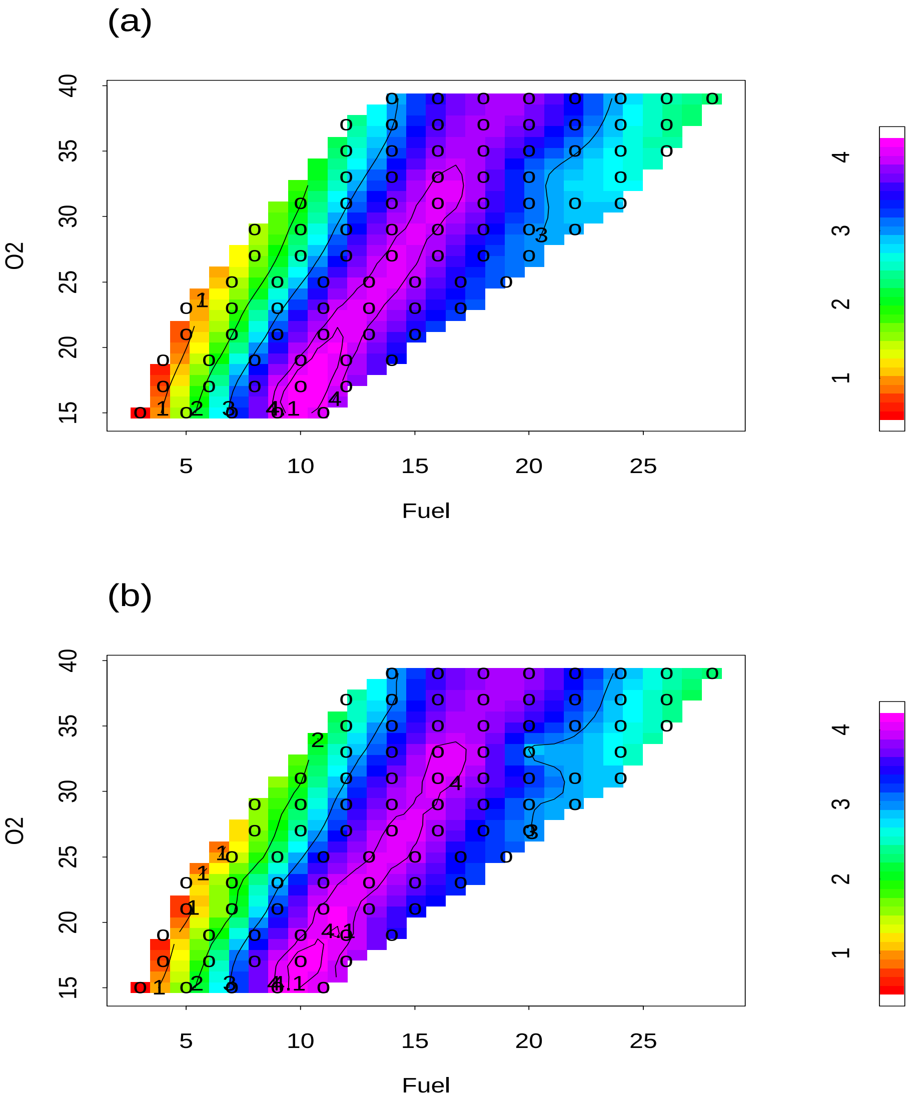

Figure 1.5: Thin-plate spline fit to the flame data (a) with $m = 3$ and kriging estimates using an exponential covariance function with range parameter $\theta = 5$ (b). Observation locations are indicated on the contour plots.

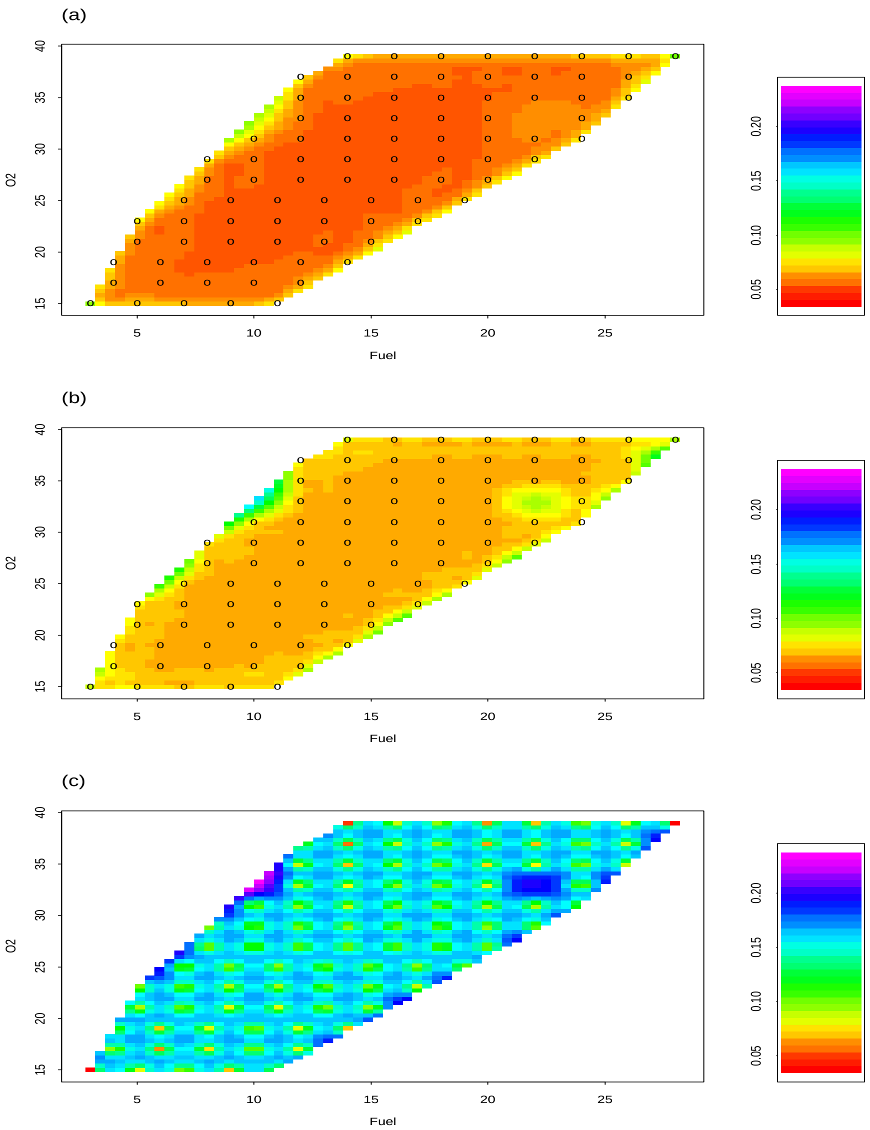

Figure 1.6: Standard errors for thin-plate spline and spatial process estimates for the flame data (see Figure 5). Plot (a) gives the standard deviation of the spline estimate assuming a fixed function and no bias. In plot (b) are the standard errors associated with the spline estimate using the covariance function implied by the radial basis functions. In plot (c) are the standard errors associated with the spatial process estimate using an exponential covariance function

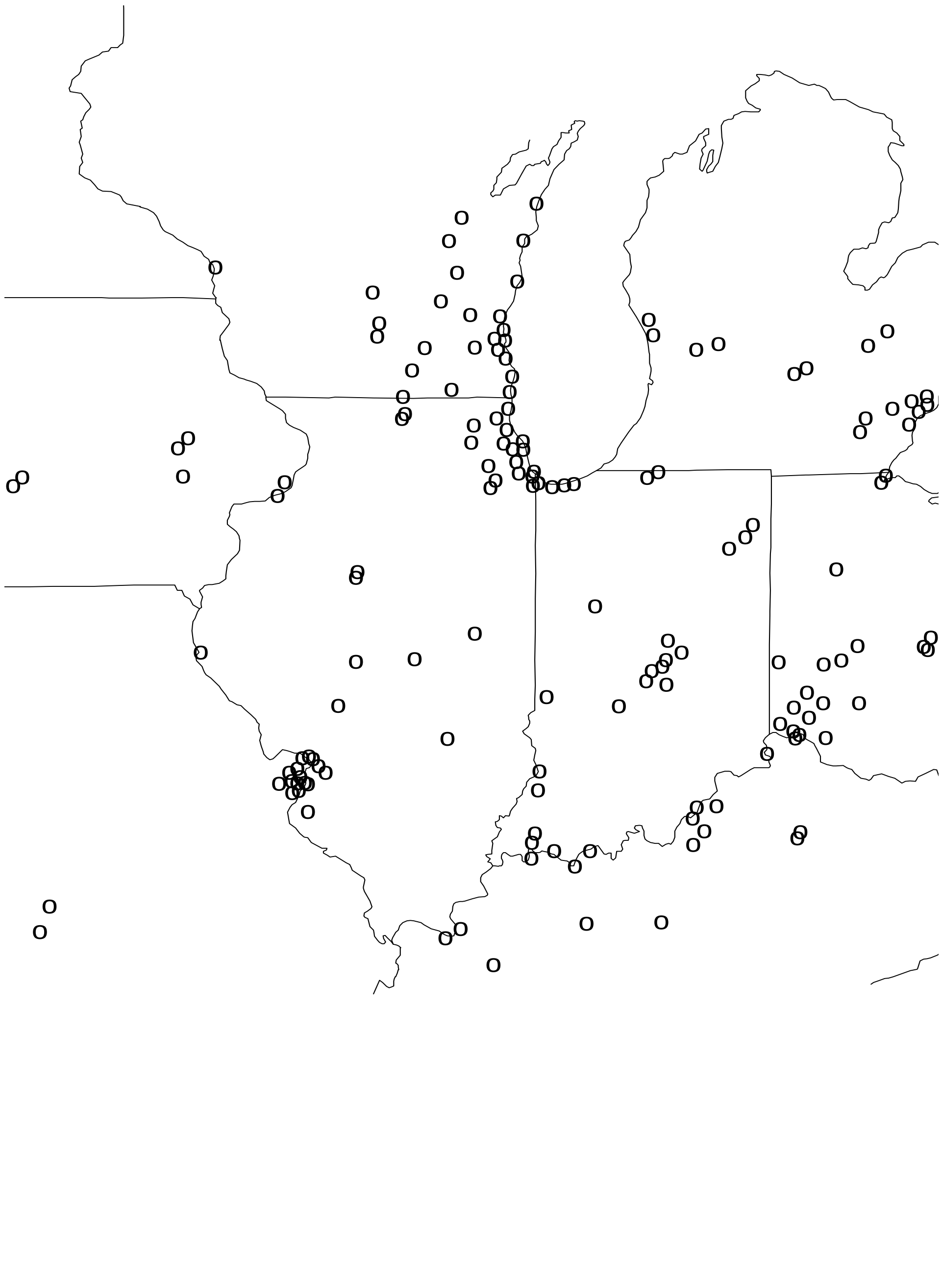

Figure 1.7: Locations of the national, state and local (NAMS/SLAMS) ozone monitoring stations operating during 1987 in the Midwestern United States.

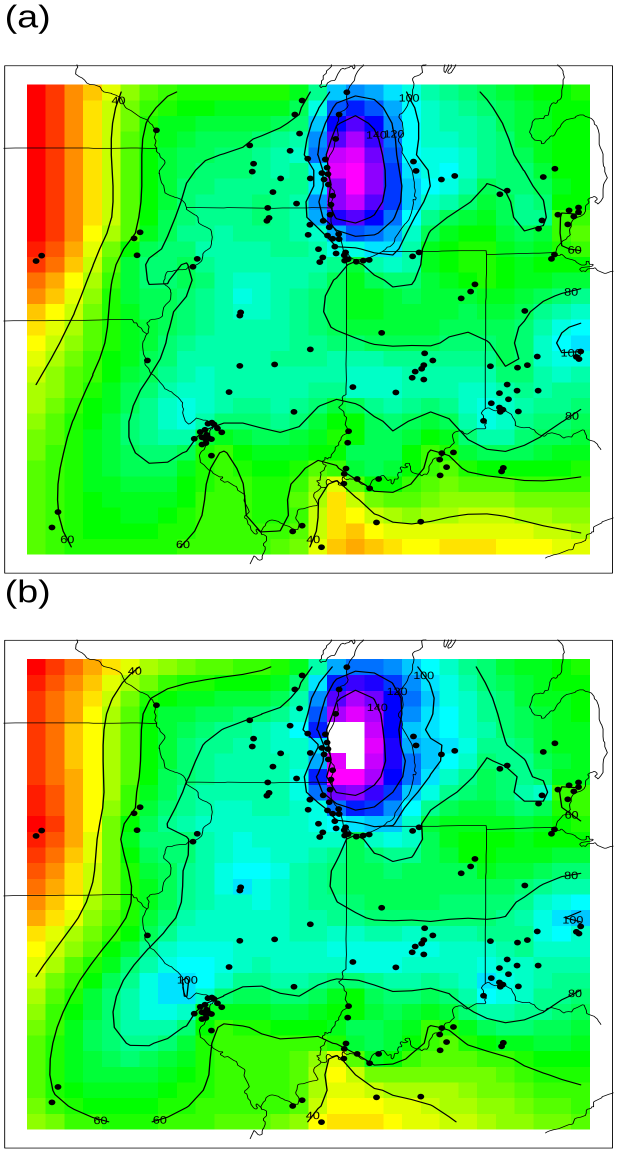

Figure 1.8: Spatial predictions for 8-hour average ozone in parts per billion (PPB) for June 19, 1987. Results for thin-plate spline (a) and kriging estimates with an isotropic, exponential correlation function but nonstationary marginal variances (b).
.

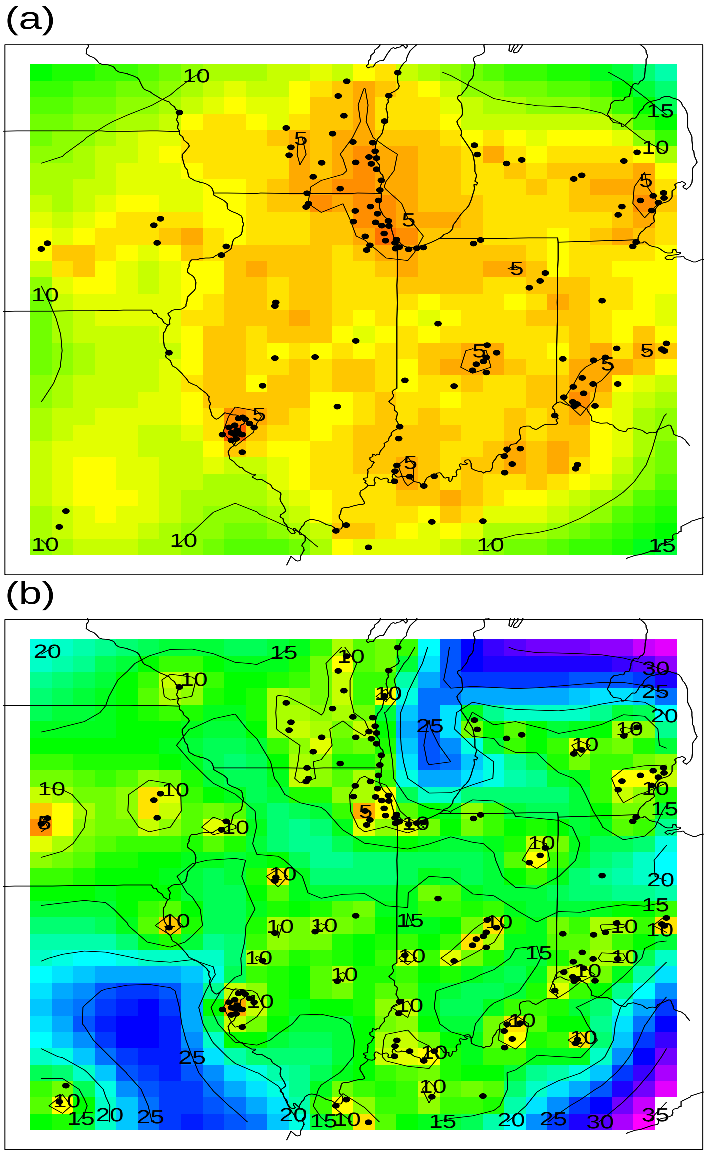

Figure 1.9: Standard errors of prediction (PPB) for the spline estimate (a) and the spatial process estimate (b) from Figure 8.