Fieldguide

CHAPTER 1

D. Nychka, R. Furrer, S. Sain

1 Introduction

A goal in statistical science is to infer properties of a population or process from a sample of incomplete observations. Our interest in this book is in using limited measurements of a curve or surface to estimate the function and to also provide measures of uncertainty for this estimate. This process involves three parts: 1) A statistical model for the unknown function and data, 2) a method of estimating the function and finally 3) a strategy for quantifying the uncertainty in the estimate. These three steps can often be unified through Bayesian statistics but we should also note that many practical statistical methods can be obtained by simpler approaches and involve substantially less computation. One reason for this book is to explore the differences and similarities among statistical methods for estimating curves and surfaces. Being part of the UseR! series we will marry our statistical exposition with statistical functions in the R fields package that implement the methods. To this end, we feel that we are not only providing a survey of these methods but also a tutorial on the computational foundations of these approaches through a large body of structured and accessible source code. The data sets that we use throughout this book are drawn from substantial problems that may be of interest on their own and can serve to motivate further research.

1.1 An example and an overview

A fundamental problem in the climate sciences is to understand the distribution of weather at a given location under our present atmospheric and ocean conditions. Knowledge of our present climate forms a foundation for more complex studies and models of climate change. A simple climate example that illustrates curve fitting is how springtime average temperatures depends on elevation for locations in Colorado. Figure 1 indicates the locations in a region centered on Colorado where surface temperatures have been measured during the period 1961-1990. The color scale indicates the average daily spring (March, April and May) temperature over the years 1961- 1990. The contours indicate topography of this region. A sobering perspective on these data is: This is it. There are not really any other di-
rect long-term surface observations in this region. Any additional inference about mean spring temperatures in between these locations must rely on other information and statistics. In particular, understanding the climate in between station locations or for the spatial average over subregions must be inferred from statistical and physical models based on these sparse data.

Figure 2 plots the average daily minimum and maximum temperatures as a function of elevation. Daily minimum and maximum temperatures are often more important to capture diurnal variation and one might expect that a changing climate might influence one more than the other. It is not surprising that both maximum and minimum temperatures decrease as a function of elevation but this not a simple linear relationship. Moreover the shape appears to be different between the maximum and minimum temperatures.

These climate data illustrate several statistical problems tackled in this
book. From Figure 1 we see some coherence in the temperatures across Colorado. These data motivate the classic problem of interpolating scattered data to form a surface: How does one estimate the values at locations where there are not observations and how does one characterize the uncertainty in this estimate? From Figure 2 we see there is also dependence on elevation. How does one combine elevation and the geographical locations to improve the interpolation? These data are actually multivariate: each location has both a minimum and maximum mean daily temperature and it is important to preserve the relationship among these two climate variables. How should this pair of variables be interpolated to reflect the joint distribution of mean minimum and maximum daily temperatures? Finally, suppose one had several thousand spatial locations across the entire United States rather than several hundred shown here. How can we implement similar statistical methods for much larger spatial data sets?

These questions are addressed in this book in three different levels: 1) practical recipes and examples, 2) a friendly implementation in R, and 3) a general framework to organize the basic principles. For the general reader interested in a gentle introduction to this useful area of statistics we give many examples using data and R functions. And in keeping with this pragmatic goal, we really do answer some of the questions for Colorado climate stated above.

At a more detailed level this book also describes the statistical computations and their implementation in R. The R functions in fields are modular and have numerous comments. They reinforce the algorithmic descriptions in this book as working examples of how to do the computations. Two important features that simplify the code in fields are a flexible description of the covariance function and overloading of the usual R functions for linear algebra with sparse matrix methods. This allows for a single source code where the covariance functions can include non-stationarity, multivariate responses and sparse matrix computations. The analysis of large spatial datasets in an interactive computational environment such as R is an emerging area. In this book we approximate the underlying covariance by a tapered version that is identically zero outside a specified range. This creates matrices that have many zeroes, known as sparse matrices, and can greatly speed calculations. The overloading for sparse matrices also allows for a single readable source code for computing the estimators with the details of the sparse matrix methods being hidden.

The last component of this book is a unified description of the methods based on ridge regression. This short book is able to treat a broad sweep of methods by tying ideas from geostatistics, splines and Bayesian methods
Figure 2: Minimum (red) and maximum (blue) average daily temperature for the months March, April and May in degrees centigrade for the period 1961-1990 plotted against elevation in meters.
together in a basic statistical framework. From geostatistics and Bayesian principles, curve and surface estimators can be motivated from the assumption that the unknown function is a realization of a random process. We call these spatial process estimators. There is also parallel development where the estimator is found by maximizing a penalized likelihood criterion over a large set of possible functions. This second perspective includes splines estimators and wavelets and we term these variational methods because they are expressed as the solution to a variational problem. In either case we will show that the resulting estimate is a sum of basis functions, tailored to the problem and information at hand, with coefficients estimated from the data. This form allows for a common set of algorithms, the fields package, but still will allow us to consider a broad range of curve and surface fitting problems. These include additive models that are sums of curves and surfaces based on different variables and multivariate models where the response at each location is more than one variable. Moreover, because the coefficients are found by penalized least squares much of the methodology can be understood from multivariate regression. The measures of uncertainty for these estimates follow from the spatial process interpretation. Coupled with Monte Carlo simulation this approach allows one to make complex but relevant inferences on the estimated function.

These levels also help us to weave the statistical ideas in this book. Different methods and models are integrated by revisiting the same data sets to aid in comparison. A unified source code helps to emphasize the common computational elements across different problems. And finally, grounding all these problems in the same statistical framework helps to tie together seemingly different problems and different estimators. Moreover, we believe that the interplay between data analysis, statistical computation and modeling imparts a contemporary – and enjoyable – perspective on using statistics to solve scientific problems that involve curves and surfaces.

2 The Basic Statistical Model

The main form of data considered throughout this book is one where a discrete set of observations are made on a curve or surface. The observations might be exact or can contain substantial measurement error. The observations locations are typically irregularly spaced but one can also consider locations on regular grids. With these data the goal is to estimate the complete function, not just the values where it is observed, and also provide a companion measure of the uncertainty. We start with a description of
the basic statistical model for the observations and the assumptions made on the unknown function. More complex models are developed with the applications in subsequent chapters.

For many researchers outside of statistical science the estimate of a curve or surface is often adopted without a broader consideration of statistical assumptions. For example one might choose a cubic smoothing spline to estimate a one-dimensional curve without exploring assumptions on the shape or smoothness of the underlying function. In practice the choice of the statistical curve estimate is often supported by physical intuition and prior experience using a particular method for specific applications. Although this approach to data analysis is adequate it does not support choosing methods for completely new kinds of data. In addition, a statistical model becomes vital when one wants to determine measures of uncertainty in the estimator. For example, in Chapter 3, we suggest generating a Monte Carlo sample of possible curve estimates that reflect the sampling uncertainty in the unknown function. Such ensembles would be difficult to generate without a statistical model. Taken a step further, our perspective is that the exact form or algorithm for a curve or surface estimate is less important than the statistical model used to describe the observations and the assumptions or constraints on the unknown function.

Much of this book is concerned with statistical problems where a function is measured with additive error and so to fix ideas we present this basic model.

Suppose $g$ is the function that one wants to determine and there are $n$ pairs of observations $(x_i, y_i), i = 1, \ldots, n$ available where

$$ y_i = g(x_i) + \epsilon_i $$

The $x_i$ are locations or variables where the function is measured or where observations are taken. In a geoscience application $x$ might be a geographic location such as the longitude and latitude. However, in general it need not be tied to a physical location and can include other variables. But to simplify our discussion we will just refer to $x$ as observation “locations”. The $\epsilon_i$ represent random departures of the observations from $g$, usually termed measurement errors. The statistical goal is to estimate the function $g$ given the observations.

As part of this introduction is it useful to explain the statistical assumptions typically made with this model. Here we just discuss the ideas in broad terms and subsequent sections will make the terminology precise with statistical expressions. At the heart of the statistical assumptions is
that the unknown \( g \) be smooth or regular in a way that allows one to borrow strength across observations and to interpolate in between observation locations. Without some notion of continuity or a consistent relationship among the values of \( g \), estimating \( g \) at a point that is not observed is difficult or even impossible. There are two useful statistical descriptions for \( g \) that at first seem very different. One, common in the analysis of spatial data, assumes that \( g \) is a realization of a smooth Gaussian stochastic process: the {\it spatial process} approach. The other point of view has its roots in numerical analysis and requires that a penalty for \( g \) based on its derivatives or other measures of roughness be small: the {\it variational approach}. In this second case one considers \( g \) as a fixed but unknown function, rather than a stochastic process. In either case, however, the interpolation of \( g \) between observed locations makes sense. The stochastic interpretation asserts that values for \( g \) are highly correlated if the separation of the locations is small. In the second approach, constraining the roughness of \( g \) also implies that the variation of \( g \) is small when locations are closely spaced.

Typically the errors, \( \{ \epsilon_i \} \), are assumed to have an expected value of zero and variances that are proportional to known set of values. Although we refer to this component as measurement error there are many other reasons to expect departures between the observations and the unknown function. For example, a rain gauge might be very accurate for the micro climate where it is sited but may not be as accurate for a larger region surrounding the gauge. In this situation it may be useful to model this departure as a "measurement error" even though it is not directly tied to faults in the measuring instrument. It is sometimes the case that the errors are small or even zero and so \( g \) is observed accurately at the observation locations. Even in this situation the interpolation of \( g \) to unobserved locations is still important and carries some uncertainty. A important advantage of the ridge regression approach is that it handles this situation as well as the one with more significant observational errors.

Both the spatial process and variational models for \( g \) lead to useful statistical estimators based on a discrete set of observations. The spatial process point of view is associated with Kriging or geostatistical methods and the roughness penalty idea motivates regularized estimators such as smoothing or thin-plate splines. A surprisingly fact is that these two approaches can lead to the same estimator. Splines are geostatistical estimators and Kriging is a kind of spline estimator. The main contribution in this book is to exploit this common form of the estimator using one computational framework. In this introduction we sketch how this equivalence comes about for one dimensional curve fitting and explore more general cases in subsequent
chapters. All of these ideas are incorporated in the fields package, and we
end this chapter by a brief tour of fields.

3 Smoothing Splines and Kriging

As a way to introduce the unifying theme of this book we return the Colorado
climate data and present a variational estimate and a stochastic process
estimate. Or, in other terms, we fit the data in Figure 2 using a
cubic smoothing spline fit and also a Kriging estimate.

Spline functions take many forms but a common thread is that they are
solutions to a minimization (variational) problem. Given the additive model
\( \sum_{i=1}^{n} (y_i - g(x_i))^2 + \int \lambda \left( g^{(2)}(u) \right)^2 du \) (2)
The sums of squares in the first term is a standard statistical measure of how
well the function \( g \) matches the observations. The second term, the integral
of the squared second derivative, is a measure of roughness in the function
and gives greater penalties for more wiggly curves. Splines take their name
from a drafting tool made from a flexible strip of metal or rubber used
to draw smooth curves between fixed points. The estimator follows this
physical analogy. The roughness penalty is interpreted as the resistance of
the strip to bending and the sums of squares as a series of springs that pull
the strip to the data points. The solution minimizes the total energy stored
as bending energy in the strip and in the stretched springs. Thin-plate
splines presented in Chapter 3 extend this idea to more than one dimension.

The parameter \( \lambda \geq 0 \) is termed the smoothing parameter and controls
the relative weight given to the residual sums of squares and the roughness
penalty. The minimizing curve is a compromise between the fit to the data
and smoothness and is readily computed using functions in fields and also
by other R packages. For nonzero lambda it will give a smooth curve that
visually smoothes the data. The estimate is a cubic polynomial in between
observation locations and the polynomial segments are joined so that \( \hat{g} \) has
two continuous derivatives. Figure 3 is the Colorado mean spring temper-
atures with cubic smoothing splines for \( \lambda = 10^6, 10^4, 10^2 \). The smoother
curve uses a value of \( \lambda \) determined by generalized cross-validation, a data-
based method for estimating this parameter.

Another approach to estimating a curve is based on a stochastic model
and we can find the expected value of \( g \) given (or conditioned) on the ob-
Figure 3: Minimum and maximum average daily temperature for the months March, April and May in degrees centigrade for the period 1961-1990 plotted against elevation in meters. Cubic smoothing spline estimates for $\lambda = 10^{-7}, 10^{-7}, 10^{-3}$. 
servations. Here we sketch the stochastic model that will give us back the same estimate as a cubic spline.

Let \( h(t) \) be a Gaussian process where

\[
    h(t) = \int_{[0,t]} W(u)du
\]

where \( W \) is a Wiener process. By this we mean that \( W(u) \) is a Gaussian process with expected value of zero and a covariance function \( E(W(u)W(v)) = \rho_{\min}(u, v) \). \( W \) is a continuous version of a one dimensional random walk and considering the integral will give a process that has at least one derivative. The process \( h \) also has an expected value of zero and the covariance function is given in the next chapter. For the moment its exact form is not important except that it has been chosen to reproduce the cubic spline estimate. The final ingredient is that we require that the errors to be uncorrelated with \( h \).

Assume that \( g \) has the form

\[
    g(x) = d_1 + d_2x + h(x).
\]

If we express any estimate of \( g(x) \) as a linear combination of the observations then one can also calculate the expected squared error of the estimate. In general we prefer this expected squared error to be small and in this case can work out the best linear, unbiased estimate of \( g(x) \). This is termed the universal Kriging estimator and the Kriging weights refer to the linear combination of the observations to give this optimal estimator. Equivalently, \( \hat{g}(x) \) is found by estimating \( d_1 \) and \( d_2 \) by generalized least squares and \( h \) by the conditional expectation of \( h \) given the observations under the assumption that that errors are distributed as multivariate normal. In either interpretation, \( \hat{g}(x) \), will be a linear combination of the observations that depends on \( x \), the covariance of \( h \) and \( \{x_i\} \). This curve estimate is identical to a cubic smoothing spline with \( \lambda = \sigma/\rho \).

Thus, from two very different perspectives we arrive at the same curve estimator. A spline estimate is easy to interpret because it simply balances fit to the data with a constraint on the smoothness and it is appropriate for a fixed, but known, \( g \). The stochastic point of view focuses on a prior specification for \( g \) and the actual estimator follows in straightforward way from conditional distributions. Also the stochastic model, if believable, can also generate measures of uncertainty for the \( \hat{g} \). The fact that these estimates can be found using the same computational framework is a strength of the \texttt{fields} package and provides an opportunity to bring both of these perspectives to curve and surface modeling.
4 A first look at fields

The fields package grew up as a data analysis tool to fit curves and surfaces to physical and environmental data. Part of its justification was to provide an efficient means to fit splines to data, make diagnostic plots and evaluate the estimate easily for plotting. Part of the emphasis was to create accessible and modular code that invited comment, improvement and extension by other users.

Another thread of development was to create a computational approach where arbitrary covariance functions, written in R code, could be incorporated into the package. In this way thin-plate splines are implemented in fields by using a particular choice of generalized covariance function. The method of generalized cross-validation to choose the smoothing parameter also makes sense for Kriging estimates and is incorporated automatically in the common code base. In this way ideas from variational methods have been incorporated into the spatial process/geostatistics point of view. The user can create new covariance models with minimal effort and then apply them to data analysis using the standard fields functions. The main function in fields is Krig a general purpose function to find a spatial process estimate given a covariance function that is written in R code. fields supplies some standard spatial covariance functions to work with Krig but it is easy to create new covariance functions for new spatial models. For example, in this book, we consider multivariate spatial models and also spatial estimators for large data sets. Both of these extensions are implemented by defining specialized covariance functions but still using the basic fields functions for estimation, evaluation and plotting.

In building the Krig function and related spatial functions we have also developed many auxiliary functions that simplify working with spatial data and are not supplied in the R base packages. For example, image.plot is the standard R image function with a default flexible option to add a legend for the color scale. The function quilt.plot is a handy image-style plot to visualize irregular spatial observations with a color scale. The function cover.design is a flexible function to thin a set of irregular spaced location to a subset that “covers” the remaining locations not in the subset. The functions world and US draw medium resolution maps of the world and United States. Although not as complete as the mapping packages in R these functions provide a quick means of adding a reference map to spatial plots.

Every R package struggles to balance the amount of detail and control available to the user with reasonable default choices and ease of use. Al-
Figure 4: Daily surface ozone values for June 19, 1997 represented by a color scale using `quilt.plot` and the thin plate spline surface where generalized cross-validation has been used to find the smoothing parameter.

though we give the user many places to adapt the statistical methods for specific spatial models our goal is that the top level functions such as `Tps` and `Krig` are easy to use and require a minimal amount of information. Given below is a quick tour of some `fields` functions to look at a two dimensional data set. Of course part of the brevity of this code is due to having good default choices for these functions. We will return to these examples in Chapter 4 with more explanation about changing the options for these methods.

The `fields` data set `ozone2` is a collection of daily surface ozone measurements at 153 Midwest monitoring stations for 89 days for the period June 3, 1987 through August 31, 1987. The measurements are the 8 hour maximum daily average (the maximum 8 hour average over a 24-hour period) reported in parts per billion (PPB) are organized in the y component as a 89 × 153 matrix. The locations are a 153 × 2 matrix of longitudes and latitudes in the component `lon.lat`. (See `help(ozone2)` for more details.) Here is an image plot for the actual ozone concentration measurements (PPB) on the 16th day with a US map added:

data(ozone2)
```r
quilt.plot( ozone2$lon.lat, ozone2$y[16,])
US(add=TRUE)

Fitting a thin-plate spline and plotting the estimated spline surface

Tps(ozone2$lon.lat, ozone2$y[16,]) -> out
surface(out) US(add=TRUE)

The results are shown in Figure 4. Here we have kept the example simple by using the longitude/latitude coordinates at face value instead of adjusting for the actual geographic distance. In Chapter 4 we explain how to use great circle distance or other measures of distance.

Besides graphics functions, the function `summary(out)` tells you what happened, `plot(out)` produces a sequence of diagnostic plots to assess the quality of this fit to these data and the `predict` function evaluates the spline surface at arbitrary locations.

On this particular day the estimate indicates a high concentration of ozone over Lake Michigan and lower amounts in rural areas away from large urban centers such as Chicago and St. Louis. Often surface estimates using the `Tps` function are effective in summarizing the spatial structure and can approximate the results of a much more extensive spatial analysis.

Returning to the motivating one dimensional example and cubic splines, here is an efficient way to fit a cubic smoothing spline in `fields` and to reproduce the results in Figure 3 for the maximum temperatures:

```r
out2 <- sreg(CO.elev, CO.tmax.MAM.climate)
# plot the data
plot( CO.elev, CO.tmax.MAM.climate)

# a fine grid of points for evaluating the estimate
xg <- seq(800, 3600, ,150)

# default fit is for the GCV estimate of lambda
# add the estimates for different lambdas
fits <- cbind( predict(out2, xg, lambda=1e6),
               predict(out2, xg, lambda=1e4),
               predict(out2, xg, lambda=1e3) )

matlines(xg, fits)
```
Similar to Tps one can apply the functions `summary` and `plot` to the fit object `out2` to get summaries and diagnostic plots.

In these examples some choices of the type of estimators have been made for the user and the amount of smoothing in the curves and surface has been estimated from the data using cross-validation. Thus, the succinctness of the commands mask some details that are important. Part of the reason for this book is to give the user of `fields` a more comprehensive understanding of what features are important for a spatial analysis. To get started, in the next chapter we will sketch the form of the penalized least squares estimator that leads to thin-plate splines and Kriging. This general form will help in understanding the different choices that need to made in a spatial analysis.