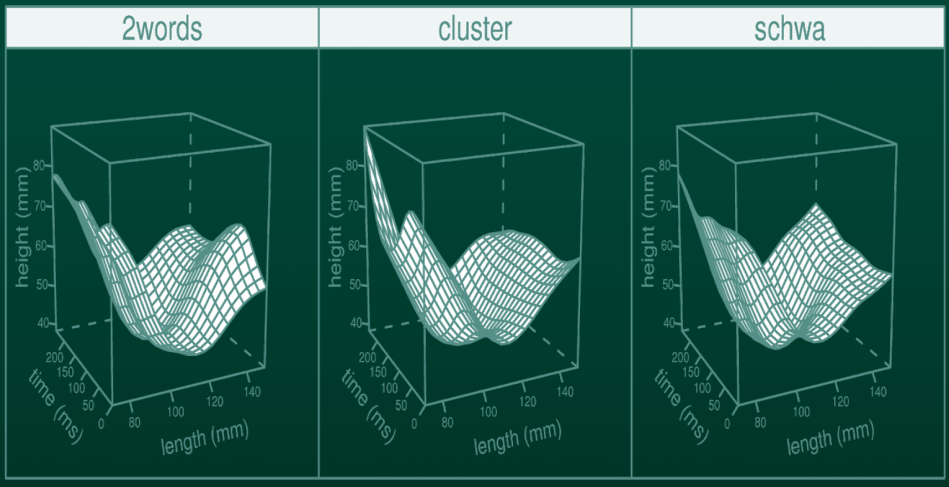
Monographs on Statistics and Applied Probability 121

Smoothing Splines

Methods and Applications



Yuedong Wang

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Smoothing Splines

Methods and Applications

MONOGRAPHS ON STATISTICS AND APPLIED PROBABILITY

General Editors

F. Bunea, V. Isham, N. Keiding, T. Louis, R. L. Smith, and H. Tong

1 Stochastic Population Models in Ecology and Epidemiology *M.S. Barlett* (1960)
2 Queues *D.R. Cox and W.L. Smith* (1961)
3 Monte Carlo Methods *J.M. Hammersley and D.C. Handscomb* (1964)
4 The Statistical Analysis of Series of Events *D.R. Cox and P.A.W. Lewis* (1966)
5 Population Genetics *W.J. Ewens* (1969)
6 Probability, Statistics and Time *M.S. Barlett* (1975)
7 Statistical Inference *S.D. Silvey* (1975)
8 The Analysis of Contingency Tables *B.S. Everitt* (1977)
9 Multivariate Analysis in Behavioural Research *A.E. Maxwell* (1977)
10 Stochastic Abundance Models *S. Engen* (1978)
11 Some Basic Theory for Statistical Inference *E.J.G. Pitman* (1979)
12 Point Processes *D.R. Cox and V. Isham* (1980)
13 Identification of Outliers *D.M. Hawkins* (1980)
14 Optimal Design *S.D. Silvey* (1980)
15 Finite Mixture Distributions *B.S. Everitt and D.J. Hand* (1981)
16 Classification *A.D. Gordon* (1981)
17 Distribution-Free Statistical Methods, 2nd edition *J.S. Maritz* (1995)
18 Residuals and Influence in Regression *R.D. Cook and S. Weisberg* (1982)
19 Applications of Queueing Theory, 2nd edition *G.F. Newell* (1982)
20 Risk Theory, 3rd edition *R.E. Beard, T. Pentikäinen and E. Pesonen* (1984)
21 Analysis of Survival Data *D.R. Cox and D. Oakes* (1984)
22 An Introduction to Latent Variable Models *B.S. Everitt* (1984)
23 Bandit Problems *D.A. Berry and B. Fristedt* (1985)
24 Stochastic Modelling and Control *M.H.A. Davis and R. Vinter* (1985)
25 The Statistical Analysis of Composition Data *J. Aitchison* (1986)
26 Density Estimation for Statistics and Data Analysis *B.W. Silverman* (1986)
27 Regression Analysis with Applications *G.B. Wetherill* (1986)
28 Sequential Methods in Statistics, 3rd edition
G.B. Wetherill and K.D. Glazebrook (1986)
29 Tensor Methods in Statistics *P. McCullagh* (1987)
30 Transformation and Weighting in Regression
R.J. Carroll and D. Ruppert (1988)
31 Asymptotic Techniques for Use in Statistics
O.E. Bandorff-Nielsen and D.R. Cox (1989)
32 Analysis of Binary Data, 2nd edition *D.R. Cox and E.J. Snell* (1989)
33 Analysis of Infectious Disease Data *N.G. Becker* (1989)
34 Design and Analysis of Cross-Over Trials *B. Jones and M.G. Kenward* (1989)
35 Empirical Bayes Methods, 2nd edition *J.S. Maritz and T. Lwin* (1989)
36 Symmetric Multivariate and Related Distributions
K.T. Fang, S. Kotz and K.W. Ng (1990)
37 Generalized Linear Models, 2nd edition *P. McCullagh and J.A. Nelder* (1989)
38 Cyclic and Computer Generated Designs, 2nd edition
J.A. John and E.R. Williams (1995)
39 Analog Estimation Methods in Econometrics *C.F. Manski* (1988)
40 Subset Selection in Regression *A.J. Miller* (1990)
41 Analysis of Repeated Measures *M.J. Crowder and D.J. Hand* (1990)
42 Statistical Reasoning with Imprecise Probabilities *P. Walley* (1991)
43 Generalized Additive Models *T.J. Hastie and R.J. Tibshirani* (1990)
44 Inspection Errors for Attributes in Quality Control
N.L. Johnson, S. Kotz and X. Wu (1991)
45 The Analysis of Contingency Tables, 2nd edition *B.S. Everitt* (1992)

46 The Analysis of Quantal Response Data *B.J.T. Morgan* (1992)

47 Longitudinal Data with Serial Correlation—A State-Space Approach
R.H. Jones (1993)

48 Differential Geometry and Statistics *M.K. Murray and J.W. Rice* (1993)

49 Markov Models and Optimization *M.H.A. Davis* (1993)

50 Networks and Chaos—Statistical and Probabilistic Aspects
O.E. Barndorff-Nielsen, J.L. Jensen and W.S. Kendall (1993)

51 Number-Theoretic Methods in Statistics *K.-T. Fang and Y. Wang* (1994)

52 Inference and Asymptotics *O.E. Barndorff-Nielsen and D.R. Cox* (1994)

53 Practical Risk Theory for Actuaries
C.D. Daykin, T. Pentikäinen and M. Pesonen (1994)

54 Biplots *J.C. Gower and D.J. Hand* (1996)

55 Predictive Inference—An Introduction *S. Geisser* (1993)

56 Model-Free Curve Estimation *M.E. Tarter and M.D. Lock* (1993)

57 An Introduction to the Bootstrap *B. Efron and R.J. Tibshirani* (1993)

58 Nonparametric Regression and Generalized Linear Models
P.J. Green and B.W. Silverman (1994)

59 Multidimensional Scaling *T.F. Cox and M.A.A. Cox* (1994)

60 Kernel Smoothing *M.P. Wand and M.C. Jones* (1995)

61 Statistics for Long Memory Processes *J. Beran* (1995)

62 Nonlinear Models for Repeated Measurement Data
M. Davidian and D.M. Giltinan (1995)

63 Measurement Error in Nonlinear Models
R.J. Carroll, D. Rupert and L.A. Stefanski (1995)

64 Analyzing and Modeling Rank Data *J.J. Marden* (1995)

65 Time Series Models—In Econometrics, Finance and Other Fields
D.R. Cox, D.V. Hinkley and O.E. Barndorff-Nielsen (1996)

66 Local Polynomial Modeling and its Applications *J. Fan and I. Gijbels* (1996)

67 Multivariate Dependencies—Models, Analysis and Interpretation
D.R. Cox and N. Wermuth (1996)

68 Statistical Inference—Based on the Likelihood *A. Azzalini* (1996)

69 Bayes and Empirical Bayes Methods for Data Analysis
B.P. Carlin and T.A Louis (1996)

70 Hidden Markov and Other Models for Discrete-Valued Time Series
I.L. MacDonald and W. Zucchini (1997)

71 Statistical Evidence—A Likelihood Paradigm *R. Royall* (1997)

72 Analysis of Incomplete Multivariate Data *J.L. Schafer* (1997)

73 Multivariate Models and Dependence Concepts *H. Joe* (1997)

74 Theory of Sample Surveys *M.E. Thompson* (1997)

75 Retrial Queues *G. Falin and J.G.C. Templeton* (1997)

76 Theory of Dispersion Models *B. Jørgensen* (1997)

77 Mixed Poisson Processes *J. Grandell* (1997)

78 Variance Components Estimation—Mixed Models, Methodologies and Applications *P.S.R.S. Rao* (1997)

79 Bayesian Methods for Finite Population Sampling
G. Meeden and M. Ghosh (1997)

80 Stochastic Geometry—Likelihood and computation
O.E. Barndorff-Nielsen, W.S. Kendall and M.N.M. van Lieshout (1998)

81 Computer-Assisted Analysis of Mixtures and Applications—
Meta-analysis, Disease Mapping and Others *D. Böhning* (1999)

82 Classification, 2nd edition *A.D. Gordon* (1999)

83 Semimartingales and their Statistical Inference *B.L.S. Prakasa Rao* (1999)

84 Statistical Aspects of BSE and vCJD—Models for Epidemics
C.A. Donnelly and N.M. Ferguson (1999)

85 Set-Indexed Martingales *G. Ivanoff and E. Merzbach* (2000)

86 The Theory of the Design of Experiments *D.R. Cox and N. Reid* (2000)

87 Complex Stochastic Systems
O.E. Barndorff-Nielsen, D.R. Cox and C. Klüppelberg (2001)

88 Multidimensional Scaling, 2nd edition *T.F. Cox and M.A.A. Cox* (2001)

89 Algebraic Statistics—Computational Commutative Algebra in Statistics
G. Pistone, E. Riccomagno and H.P. Wynn (2001)

90 Analysis of Time Series Structure—SSA and Related Techniques
N. Golyandina, V. Nekrutkin and A.A. Zhigljavsky (2001)

91 Subjective Probability Models for Lifetimes
Fabio Spizzichino (2001)

92 Empirical Likelihood *Art B. Owen* (2001)

93 Statistics in the 21st Century
Adrian E. Raftery, Martin A. Tanner, and Martin T. Wells (2001)

94 Accelerated Life Models: Modeling and Statistical Analysis
Vilijandas Bagdonavicius and Mikhail Nikulin (2001)

95 Subset Selection in Regression, Second Edition *Alan Miller* (2002)

96 Topics in Modelling of Clustered Data
Marc Aerts, Helena Geys, Geert Molenberghs, and Louise M. Ryan (2002)

97 Components of Variance *D.R. Cox and P.J. Solomon* (2002)

98 Design and Analysis of Cross-Over Trials, 2nd Edition
Byron Jones and Michael G. Kenward (2003)

99 Extreme Values in Finance, Telecommunications, and the Environment
Bärbel Finkenstädt and Holger Rootzén (2003)

100 Statistical Inference and Simulation for Spatial Point Processes
Jesper Møller and Rasmus Plenge Waagepetersen (2004)

101 Hierarchical Modeling and Analysis for Spatial Data
Sudipto Banerjee, Bradley P. Carlin, and Alan E. Gelfand (2004)

102 Diagnostic Checks in Time Series *Wai Keung Li* (2004)

103 Stereology for Statisticians *Adrian Baddeley and Eva B. Vedel Jensen* (2004)

104 Gaussian Markov Random Fields: Theory and Applications
Håvard Rue and Leonhard Held (2005)

105 Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition
Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu (2006)

106 Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood
Youngjo Lee, John A. Nelder, and Yudi Pawitan (2006)

107 Statistical Methods for Spatio-Temporal Systems
Bärbel Finkenstädt, Leonhard Held, and Valerie Isham (2007)

108 Nonlinear Time Series: Semiparametric and Nonparametric Methods
Jiti Gao (2007)

109 Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis
Michael J. Daniels and Joseph W. Hogan (2008)

110 Hidden Markov Models for Time Series: An Introduction Using R
Walter Zucchini and Iain L. MacDonald (2009)

111 ROC Curves for Continuous Data
Wojtek J. Krzanowski and David J. Hand (2009)

112 Antedependence Models for Longitudinal Data
Dale L. Zimmerman and Vicente A. Núñez-Antón (2009)

113 Mixed Effects Models for Complex Data
Lang Wu (2010)

114 Intoduction to Time Series Modeling
Genshiro Kitagawa (2010)

115 Expansions and Asymptotics for Statistics
Christopher G. Small (2010)

116 Statistical Inference: An Integrated Bayesian/Likelihood Approach
Murray Aitkin (2010)

117 Circular and Linear Regression: Fitting Circles and Lines by Least Squares
Nikolai Chernov (2010)

118 Simultaneous Inference in Regression *Wei Liu* (2010)

119 Robust Nonparametric Statistical Methods, Second Edition *Thomas P. Hettmansperger and Joseph W. McKean* (2011)

120 Statistical Inference: The Minimum Distance Approach
Ayanendranath Basu, Hiroyuki Shioya, and Chanseok Park (2011)

121 Smoothing Splines : Methods and Applications *Yuedong Wang* (2011)

Monographs on Statistics and Applied Probability 121

Smoothing Splines

Methods and Applications

Yuedong Wang
University of California
Santa Barbara, California, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20110429

International Standard Book Number-13: 978-1-4200-7756-8 (eBook - PDF)

TO

YAN, CATHERINE, AND KEVIN

This page intentionally left blank

Contents

This page intentionally left blank

List of Tables

This page intentionally left blank

List of Figures

Symbol Description

$(x)_+$	$\max\{x, 0\}$
$x \wedge z$	$\min\{x, z\}$
$x \vee z$	$\max\{x, z\}$
$\det^+$	Product of the nonzero eigenvalues
$k_r(x)$	Scaled Bernoulli polynomials
$(\cdot, \cdot)$	Inner product
$\|\cdot\|$:	Norm
$\mathcal{X}$	Domain of a function
$\mathcal{S}$	Unit sphere
$\mathcal{H}$	Function space
$\mathcal{L}$	Linear functional
$\mathcal{N}$	Nonlinear functional
P	Projection
$\mathcal{A}$	Averaging operator
$\mathcal{M}$	Model space
$R(x, z)$	Reproducing kernel
$\mathbb{R}^d$	Euclidean d-space
$NS^{2m}(t_1, \cdots, t_k)$	Natural polynomial spline space
$W_2^m[a, b]$	Sobolev space on $[a, b]$
$W_2^m(per)$	Sobolev space on unit circle
$W_2^m(\mathbb{R}^d)$	Thin-plate spline model space
$W_2^m(\mathcal{S})$	Sobolev space on unit sphere
$\oplus$	Direct sum of function spaces
$\otimes$	Tensor product of function spaces

This page intentionally left blank

Preface

Statistical analysis often involves building mathematical models that examine the relationship between dependent and independent variables. This book is about a general class of powerful and flexible modeling techniques, namely, *spline smoothing*.

Research on smoothing spline models has attracted a great deal of attention in recent years, and the methodology has been widely used in many areas. This book provides an introduction to some basic smoothing spline models, including polynomial, periodic, spherical, thin-plate, L-, and partial splines, as well as an overview of more advanced models, including smoothing spline ANOVA, extended and generalized smoothing spline ANOVA, vector spline, nonparametric nonlinear regression, semiparametric regression, and semiparametric mixed-effects models. Methods for model selection and inference are also presented.

The general forms of nonparametric/semiparametric linear/nonlinear fixed/mixed smoothing spline models in this book provide unified frameworks for estimation, inference, and software implementation. This book draws on the theory of reproducing kernel Hilbert space (RKHS) to present various smoothing spline models in a unified fashion. On the other hand, the subject of smoothing spline in the context of RKHS and regularization is often regarded as technical and difficult. One of my main goals is to make the advanced smoothing spline methodology based on RKHS more accessible to practitioners and students. With this in mind, the book focuses on methodology, computation, implementation, software, and application. It provides a gentle introduction to the RKHS, keeps theory at the minimum level, and provides details on how the RKHS can be used to construct spline models.

User-friendly software is key to the routine use of any statistical method. The `assist` library in R implements methods presented in this book for fitting various nonparametric/semiparametric linear/nonlinear fixed/mixed smoothing spline models. The `assist` library can be obtained at

`http://www.r-project.org`

Much of the exposition is based on the analysis of real examples. Rather than formal analysis, these examples are intended to illustrate the power and versatility of the spline smoothing methodology. All data analyses are performed in R, and most of them use functions in the

`assist` library. Codes for all examples and further developments related to this book will be posted on the web page

`http://www.pstat.ucsb.edu/faculty/yuedong/book.html`

This book is intended for those wanting to learn about smoothing splines. It can be a reference book for statisticians and scientists who need advanced and flexible modeling techniques. It can also serve as a text for an advanced-level graduate course on the subject. In fact, topics in Chapters 1–4 were covered in a quarter class at the University of California — Santa Barbara, and the University of Science and Technology of China.

I was fortunate indeed to have learned the smoothing spline from Grace Wahba, whose pioneering work has paved the way for much ongoing research and made this book possible. I am grateful to Chunlei Ke, my former student and collaborator, for developing the `assist` package. Special thanks goes to Anna Liu for reading the draft carefully and correcting many mistakes. Several people have helped me over various phases of writing this book: Chong Gu, Wensheng Guo, David Hinkley, Ping Ma, and Wendy Meiring. I must thank my editor, David Grubbes, for his patience and encouragement. Finally, I would like to thank several researchers who kindly shared their data sets for inclusion in this book; they are cited where their data are introduced.

Yuedong Wang
Santa Barbara
December 2010

Chapter 1

Introduction

1.1 Parametric and Nonparametric Regression

Regression analysis builds mathematical models that examine the relationship of a dependent variable to one or more independent variables. These models may be used to predict responses at unobserved and/or future values of the independent variables. In the simple case when both the dependent variable y and the independent variable x are scalar variables, given observations (x_i, y_i) for $i = 1, \ldots, n$, a *regression model* relates dependent and independent variables as follows:

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where f is the *regression function* and ϵ_i are zero-mean independent random errors with a common variance σ^2. The goal of regression analysis is to construct a model for f and estimate it based on noisy data.

For example, for the Old Faithful geyser in Yellowstone National Park, consider the problem of predicting the waiting time to the next eruption using the length of the previous eruption. Figure 1.1(a) shows the scatter plot of waiting time to the next eruption ($y =$ `waiting`) against duration of the previous eruption ($x =$ `duration`) for 272 observations from the Old Faithful geyser. The goal is to build a mathematical model that relates the waiting time to the duration of the previous eruption. A first attempt might be to approximate the regression function f by a straight line

$$f(x) = \beta_0 + \beta_1 x. \tag{1.2}$$

The least squares straight line fit is shown in Figure 1.1(a). There is no apparent sign of lack-of-fit. Furthermore, there is no clear visible trend in the plot of residuals in Figure 1.1(b).

Often f is nonlinear in x. A common approach to dealing with nonlinear relationship is to approximate f by a *polynomial of order m*

$$f(x) = \beta_0 + \beta_1 x + \cdots + \beta_{m-1} x^{m-1}. \tag{1.3}$$

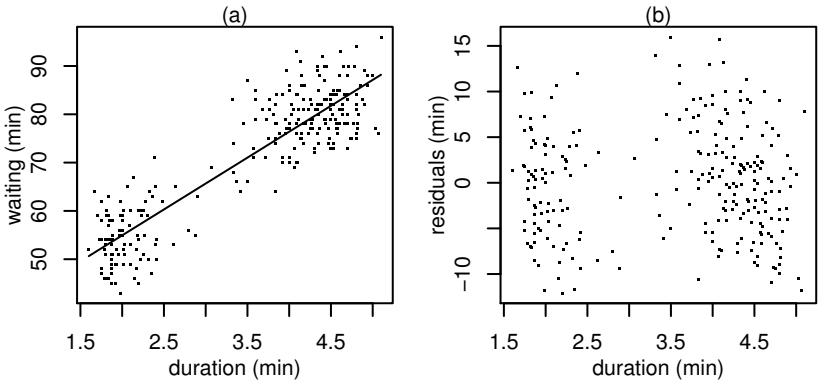


FIGURE 1.1 Geyser data, plots of (a) observations and the least squares straight line fit, and (b) residuals.

Figure 1.2 shows the scatter plot of acceleration ($y =$ `acceleration`) against time after impact ($x =$ `time`) from a simulated motorcycle crash experiment on the efficacy of crash helmets. It is clear that a straight line cannot explain the relationship between acceleration and time. Polynomials with $m = 1, \ldots, 20$ are fitted to the data, and Figure 1.2 shows the best fit selected by *Akaike's information criterion* (AIC). There are waves in the fitted curve at both ends of the range. The fit is still not completely satisfactory even when polynomials up to order 20 are considered. Unlike the linear regression model (1.2), except for small m, coefficients in model (1.3) no longer have nice interpretations.

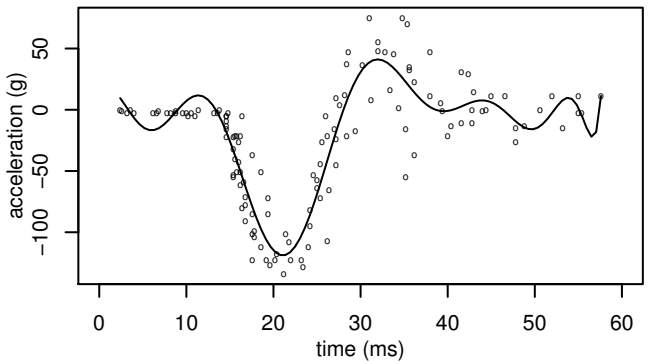


FIGURE 1.2 Motorcycle data, plot of observations, and a polynomial fit.

In general, a *parametric regression* model assumes that the form of f is known except for finitely many unknown parameters. The specific form of f may come from scientific theories and/or approximations to mechanics under some simplified assumptions. The assumptions may be too restrictive and the approximations may be too crude for some applications. An inappropriate model can lead to systematic bias and misleading conclusions. In practice, one should always check the assumed form for the function f.

It is often difficult, if not impossible, to obtain a specific functional form for f. A *nonparametric regression* model does not assume a predetermined form. Instead, it makes assumptions on qualitative properties of f. For example, one may be willing to assume that f is "smooth", which does not reduce to a specific form with finite number of parameters. Rather, it usually leads to some infinite dimensional collections of functions. The basic idea of nonparametric regression is to let the data speak for themselves. That is to let the data decide which function fits the best without imposing any specific form on f. Consequently, nonparametric methods are in general more flexible. They can uncover structure in the data that might otherwise be missed.

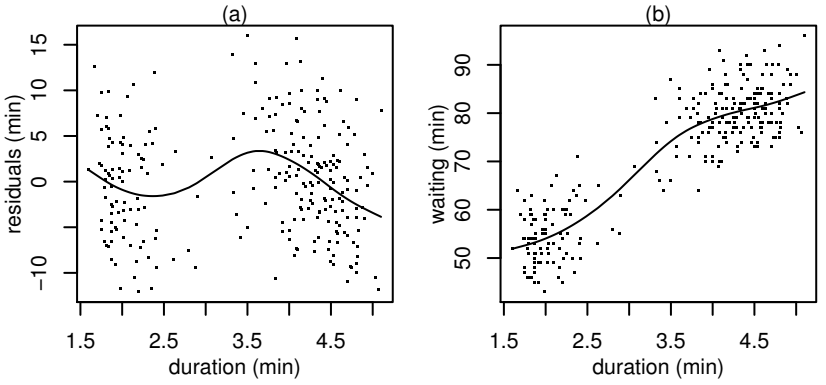


FIGURE 1.3 Geyser data, plots of (a) residuals from the straight line fit and the cubic spline fit to the residuals, and (b) the cubic spline fit to the original data.

For illustration, we fit cubic splines to the geyser data. The cubic spline is a special nonparametric regression model that will be introduced in Section 1.2. A cubic spline fit to residuals from the linear model (1.2) reveals a nonzero trend in Figure 1.3(a). This raises the question of

whether a simple linear regression model is appropriate for the geyser data. A cubic spline fit to the original data is shown in Figure 1.3(b). It reveals that there are two clusters in the independent variable, and a different linear model may be required for each cluster. Sections 2.10, 3.8, and 3.9 contain more analysis of the geyser data. A cubic spline fit to the motorcycle data is shown in Figure 1.4. It fits data much better than the polynomial model. Sections 2.10, 3.8, 5.4.1, and 6.4 contain more analysis of the motorcycle data.

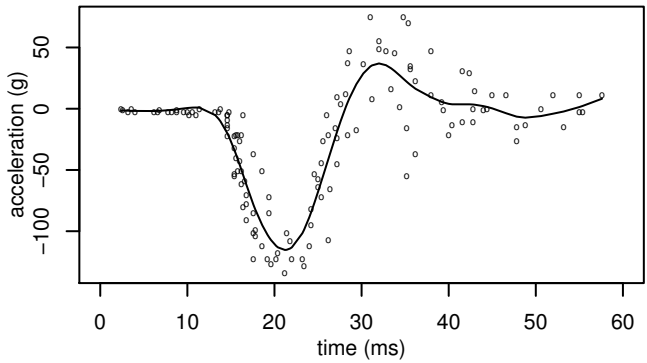


FIGURE 1.4 Motorcycle data, plot of observations, and the cubic spline fit.

The above simple exposition indicates that the nonparametric regression technique can be applied to different steps in regression analysis: data exploration, model building, testing parametric models, and diagnosis. In fact, as illustrated throughout the book, *spline smoothing* is a powerful and versatile tool for building statistical models to exploit structures in data.

1.2 Polynomial Splines

The polynomial (1.3) is a global model which makes it less adaptive to local variations. Individual observations can have undue influence on the fit in remote regions. For example, in the motorcycle data, the behavior of the mean function varies drastically from one region to another.

These local variations led to oscillations at both ends of the range in the polynomial fit. A natural solution to overcome this limitation is to use piecewise polynomials, the basic idea behind *polynomial splines*.

Let $a < t_1 < \cdots < t_k < b$ be fixed points called *knots*. Let $t_0 = a$ and $t_{k+1} = b$. Roughly speaking, polynomial splines are piecewise polynomials joined together smoothly at knots. Formally, a *polynomial spline* of order r is a real-valued function on $[a, b]$, $f(t)$, such that

(i) f is a piecewise polynomial of order r on $[t_i, t_{i+1})$, $i = 0, 1, \ldots, k$;

(ii) f has $r - 2$ continuous derivatives and the $(r-1)$st derivative is a step function with jumps at knots.

Now consider even orders represented as $r = 2m$. The function f is a *natural polynomial spline* of order $2m$ if, in addition to (i) and (ii), it satisfies the *natural boundary conditions*

(iii) $f^{(j)}(a) = f^{(j)}(b) = 0,\ j = m, \ldots, 2m-1.$

The natural boundary conditions imply that f is a polynomial of order m on the two outside subintervals $[a, t_1]$ and $[t_k, b]$. Denote the function space of natural polynomial splines of order $2m$ with knots $t_1, \ldots, t_k$ as $NS^{2m}(t_1, \ldots, t_k)$.

One approach, known as *regression spline*, is to approximate f using a polynomial spline or natural polynomial spline. To get a good approximation, one needs to decide the number and locations of knots. This book covers a different approach known as *smoothing spline*. It starts with a well-defined model space for f and introduces a penalty to prevent overfitting. We now describe this approach for polynomial splines.

Consider the regression model (1.1). Suppose f is "smooth". Specifically, assume that $f \in W_2^m[a, b]$ where the *Sobolev space*

$$W_2^m[a,b] = \big\{ f :\ f, f', \ldots, f^{(m-1)} \text{ are absolutely continuous,} \int_a^b (f^{(m)})^2 dx < \infty \big\}. \tag{1.4}$$

For any $a \le x \le b$, Taylor's theorem states that

$$f(x) = \underbrace{\sum_{\nu=0}^{m-1} \frac{f^{(\nu)}(a)}{\nu!}(x-a)^\nu}_{\text{polynomial of order } m} + \underbrace{\int_a^x \frac{(x-u)^{m-1}}{(m-1)!} f^{(m)}(u)du}_{\text{Rem}(x)}. \tag{1.5}$$

It is clear that the polynomial regression model (1.3) ignores the remainder term $\text{Rem}(x)$ in the hope that it is negligible. It is often difficult

to verify this assumption in practice. The idea behind the spline smoothing is to let data decide how large $\text{Rem}(x)$ should be. Since $W_2^m[a,b]$ is an infinite dimensional space, a direct fit to f by minimizing the *least squares* (LS)

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{1.6}$$

leads to interpolation. Therefore, certain control over $\text{Rem}(x)$ is necessary. One natural approach is to control how far f is allowed to depart from the polynomial model. Under appropriate norms defined later in Sections 2.2 and 2.6, one measure of distance between f and polynomials is $\int_a^b (f^{(m)})^2 dx$. It is then reasonable to estimate f by minimizing the LS (1.6) under the constraint

$$\int_a^b (f^{(m)})^2 dx \le \rho \tag{1.7}$$

for a constant ρ. By introducing a Lagrange multiplier, the constrained minimization problem (1.6) and (1.7) is equivalent to minimizing the *penalized least squares* (PLS):

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_a^b (f^{(m)})^2 dx. \tag{1.8}$$

In the remainder of this book, a polynomial spline refers to the solution of the PLS (1.8) in the model space $W_2^m[a,b]$. A *cubic spline* is a special case of the polynomial spline with $m = 2$. Since it measures the roughness of the function f, $\int_a^b (f^{(m)})^2 dx$ is often referred to as a *roughness penalty*. It is obvious that there is no penalty for polynomials of order less than or equal to m. The *smoothing parameter* λ balances the trade-off between goodness-of-fit measured by the LS and roughness of the estimate measured by $\int_a^b (f^{(m)})^2 dx$.

Suppose that $n \ge m$ and $a \le x_1 < x_2 < \cdots < x_n \le b$. Then, for fixed $0 < \lambda < \infty$, (1.8) has a unique minimizer $\hat{f}$ and $\hat{f} \in NS^{2m}(x_1, \ldots, x_n)$ (Eubank 1988). This result indicates that even though we started with the infinite dimensional space $W_2^m[a,b]$ as the model space for f, the solution to the PLS (1.8) belongs to a finite dimensional space. Specifically, the solution is a natural polynomial spline with knots at distinct design points. One approach to computing the polynomial spline estimate is to represent $\hat{f}$ as a linear combination of a basis of $NS^{2m}(x_1, \ldots, x_n)$. Several basis constructions were provided in Section 3.3.3 of Eubank (1988). In particular, the R function `smooth.spline` implements this approach for the cubic spline using the B-spline basis. For example, the cubic spline fit in Figure 1.4 is derived by the following statements:

```
> library(MASS); attach(mcycle)
> smooth.spline(times, accel, all.knots=T)
```

This book presents a different approach. Instead of basis functions, representers of reproducing kernel Hilbert spaces will be used to represent the spline estimate. This approach allows us to deal with many different spline models in a unified fashion. Details of this approach for polynomial splines will be presented in Sections 2.2 and 2.6.

When $\lambda = 0$, there is no penalty, and the natural spline that interpolates observations is the unique minimizer. When $\lambda = \infty$, the unique minimizer is the mth order polynomial. As λ varies from ∞ to 0, we have a family of models ranging from the parametric polynomial model to interpolation. The value of λ decides how far f is allowed to depart from the polynomial model. Thus the choice of λ holds the key to the success of a spline estimate. We discuss how to choose λ based on data in Chapter 3.

1.3 Scope of This Book

Driven by many sophisticated applications and fueled by modern computing power, many flexible nonparametric and semiparametric modeling techniques have been developed to relax parametric assumptions and to exploit possible hidden structure. There are many different nonparametric methods. This book concentrates on one of them, *smoothing spline*. Existing books on this topic include Eubank (1988), Wahba (1990), Green and Silverman (1994), Eubank (1999), Gu (2002), and Ruppert, Wand and Carroll (2003). The goals of this book are to (a) make the advanced smoothing spline methodology based on reproducing kernel Hilbert spaces more accessible to practitioners and students; (b) provide software and examples so that the spline smoothing methods can be routinely used in practice; and (c) provide a comprehensive coverage of recently developed smoothing spline nonparametric/semiparametric linear/nonlinear fixed/mixed models. We concentrate on the methodology, implementation, software, and application. Theoretical results are stated without proofs. All methods will be demonstrated using real data sets and R functions.

The polynomial spline in Section 1.2 concerns the functions defined on the domain $[a, b]$. In many applications, the domain of the regression function is not a continuous interval. Furthermore, the regression function may only be observed indirectly. **Chapter 2** introduces gen-

eral smoothing spline regression models with reproducing kernel Hilbert spaces on general domains as model spaces. Penalized LS estimation, Kimeldorf–Wahba representer theorem, computation, and the R function `ssr` will be covered. Explicit constructions of model spaces will be discussed in detail for some popular smoothing spline models including polynomial, periodic, thin-plate, spherical, and L-splines.

Chapter 3 introduces methods for selecting the smoothing parameter and making inferences about the regression function. The impact of the smoothing parameter and basic concepts for model selection will be discussed and illustrated using an example. Connections between smoothing spline models and Bayes/mixed-effects models will be established. The unbiased risk, generalized cross-validation, and generalized maximum likelihood methods will be introduced for selecting the smoothing parameter. Bayesian and bootstrap confidence intervals will be introduced for the regression function and its components. The locally most powerful, generalized maximum likelihood and generalized cross-validation tests will also be introduced to test the hypothesis of a parametric model versus a nonparametric alternative.

Analogous to multiple regression, **Chapter 4** constructs models for multivariate regression functions based on smoothing spline analysis of variance (ANOVA) decompositions. The resulting models have hierarchical structures that facilitate model selection and interpretation. Smoothing spline ANOVA decompositions for tensor products of some commonly used smoothing spline models will be illustrated. Penalized LS estimation involving multiple smoothing parameters and componentwise Bayesian confidence intervals will be covered.

Chapter 5 presents spline smoothing methods for heterogeneous and correlated observations. Presence of heterogeneity and correlation may lead to wrong choice of the smoothing parameters and erroneous inference. Penalized weighted LS will be used for estimation. Unbiased risk, generalized cross-validation, and generalized maximum likelihood methods will be extended for selecting the smoothing parameters. Variance and correlation structures will also be discussed.

Analogous to generalized linear models, **Chapter 6** introduces smoothing spline ANOVA models for observations generated from a particular distribution in the exponential family including binomial, Poisson, and gamma distributions. Penalized likelihood will be used for estimation, and methods for selecting the smoothing parameters will be discussed. Nonparametric estimation of variance and spectral density functions will be presented.

Analogous to nonlinear regression, **Chapter 7** introduces spline smoothing methods for nonparametric nonlinear regression models where some unknown functions are observed indirectly through nonlinear function-

als. In addition to fitting theoretical and empirical nonlinear nonparametric regression models, methods in this chapter may also be used to deal with constraints on the nonparametric function such as positivity or monotonicity. Several algorithms based on Gauss–Newton, Newton–Raphson, extended Gauss–Newton and Gauss–Seidel methods will be presented for different situations. Computation and the R function `nnr` will be covered.

Chapter 8 introduces semiparametric regression models that involve both parameters and nonparametric functions. The mean function may depend on the parameters and the nonparametric functions linearly or nonlinearly. The semiparametric regression models include many well-known models such as the partial spline, varying coefficients, projection pursuit, single index, multiple index, functional linear, and shape invariant models as special cases. Estimation, inference, computation, and the R function `snr` will also be covered.

Chapter 9 introduces semiparametric linear and nonlinear mixed-effects models. Smoothing spline ANOVA decompositions are extended for the construction of semiparametric mixed-effects models that parallel the classical mixed models. Estimation and inference methods, computation, and the R functions `slm` and `snm` will be covered as well.

1.4 The `assist` Package

The `assist` package was developed for fitting various smoothing spline models covered in this book. It contains five main functions, `ssr`, `nnr`, `snr`, `slm`, and `snm` for fitting various smoothing spline models. The function `ssr` fits smoothing spline regression models in Chapter 2, smoothing spline ANOVA models in Chapter 4, extended smoothing spline ANOVA models with heterogeneous and correlated observations in Chapter 5, generalized smoothing spline ANOVA models in Chapter 6, and semiparametric linear regression models in Chapter 8, Section 8.2. The function `nnr` fits nonparametric nonlinear regression models in Chapter 7. The function `snr` fits semiparametric nonlinear regression models in Chapter 8, Section 8.3. The functions `slm` and `snm` fit semiparametric linear and nonlinear mixed-effects models in Chapter 9. The `assist` package is available at

`http://cran.r-project.org`

Figure 1.5 shows how the functions in `assist` generalize some of the existing R functions for regression analysis.

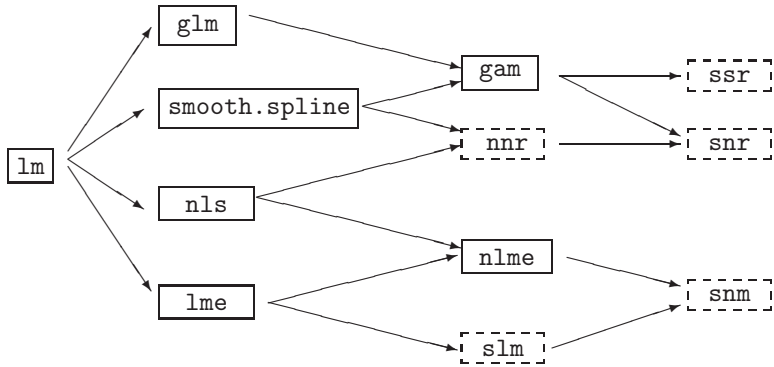


FIGURE 1.5 Functions in `assist` (dashed boxes) and some existing R functions (solid boxes). An arrow represents an extension to a more general model. `lm`: linear models. `glm`: generalized linear models. `smooth.spline`: cubic spline models. `nls`: nonlinear regression models. `lme`: linear mixed-effects models. `gam`: generalized additive models. `nlme`: nonlinear mixed-effects models. `ssr`: smoothing spline regression models. `nnr`: nonparametric nonlinear regression models. `snr`: semiparametric nonlinear regression models. `slm`: semiparametric linear mixed-effects models. `snm`: semiparametric nonlinear mixed-effects models.

References

Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Washington, DC: National Bureau of Standards.

Andrews, D. F. and Herzberg, A. M. (1985). *Data: A Collection of Problems From Many Fields for the Student and Research Worker*, Springer, Berlin.

Aronszajn, N. (1950). Theory of reproducing kernels, *Transactions of the American Mathematics Society* **68**: 337–404.

Bennett, L. H., Swartzendruber, L. J. Turchinskaya, M. J., Blendell, J. E., Habib, J. M. and Seyoum, H. M. (1994). Long-time magnetic relaxation measurements on a quench melt growth YBCO superconductor, *Journal of Applied Physics* **76**: 6950–6952.

Berlinet, A. and Thomas-Agnan, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic, Norwell, MA.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models, *Journal of the American Statistical Association* **88**: 9–25.

Carroll, R. J., Fan, J., Gijbels, I. and Wand, M. P. (1997). Generalized partial linear single-index models, *Journal of the American Statistical Association* **92**: 477–489.

Coddington, E. A. (1961). *An Introduction to Ordinary Differential Equations*, Prentice-Hall, NJ.

Cox, D. D., Koh, E., Wahba, G. and Yandell, B. (1988). Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline model, *Annals of Statistics* **16**: 113–119.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions, *Numerische Mathematik* **31**: 377–403.

Dalzell, C. J. and Ramsay, J. O. (1993). Computing reproducing kernels with arbitrary boundary constraints, *SIAM Journal on Scientific Computing* **14**: 511–518.

Davidson, L. (2006). Comparing tongue shapes from ultrasound imaging using smoothing spline analysis of variance., *Journal of the Acoustical Society of America* **120**: 407–415.

Davies, R. B. (1980). The distribution of a linear combination of χ^2 random variables, *Applied Statistics* **29**: 323–333.

Debnath, L. and Mikusiński, P. (1999). *Introduction to Hilbert Spaces with Applications*, Academic Press, London.

Douglas, A. and Delampady, M. (1990). *Eastern Lake Survey — Phase I: documentation for the data base and the derived data sets*, SIMS Technical Report 160. Department of Statistics, University of British Columbia, Vancouver.

Duchon, J. (1977). Spline minimizing rotation-invariant semi-norms in Sobolev spaces, pp. 85–100. In *Constructive Theory of Functions of Several Variables*, W. Schemp and K. Zeller eds., Springer, Berlin.

Earn, D. J. D., Rohani, P., Bolker, B. M. and Gernfell, B. T. (2000). A simple model for complex dynamical transitions in epidemics, *Science* **287**: 667–670.

Efron, B. (2001). Selection criteria for scatterplot smoothers, *Annals of Statistics* **29**: 470–504.

Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation (with discussion), *Journal of the American Statistical Association* **99**: 619–632.

Eubank, R. (1988). *Spline Smoothing and Nonparametric Regression*, Dekker, New York.

Eubank, R. (1999). *Nonparametric Regression and Spline Smoothing, 2nd ed.*, Dekker, New York.

Evans, M. and Swartz, T. (2000). *Approximating Integrals via Monte Carlo and Deterministic Methods*, Oxford University Press, Oxford, UK.

Fisher, M. D., Nychka, D. and Zervos, D. (1995). *Fitting the term structure of interest rates with smoothing spline*, Working Paper 95-1, Finance and Eonomics Discussion Series, Federal Reserve Board.

Flett, T. M. (1980). *Differential Analysis*, Cambridge University Press, London.

Friedman, J. H. and Stuetzle, W. (1981). Projection pursuit regression, *Journal of the American Statistical Association* **76**: 817–823.

Gause, G. F. (1934). *The Struggle for Existence*, Williams & Wilkins, Baltimore, MD.

Genton, M. G. and Hall, P. (2007). Statistical inference for evolving periodic functions, *Journal of the Royal Statistical Society B* **69**: 643–657.

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

Grizzle, J. E. and Allen, D. M. (1969). Analysis of growth and dose response curves, *Biometrics* **25**: 357–381.

Gu, C. (1992). Penalized likelihood regression: A Bayesian analysis, *Statistica Sinica* **2**: 255–264.

Gu, C. (2002). *Smoothing Spline ANOVA Models*, Springer, New York.

Guo, W., Dai, M., Ombao, H. C. and von Sachs, R. (2003). Smoothing spline ANOVA for time-dependent spectral analysis, *Journal of the American Statistical Association* **98**: 643–652.

Hall, P., Kay, J. W. and Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression, *Biometrika* **77**: 521–528.

Harville, D. (1976). Extension of the Gauss-Markov theorem to include the estimation of random effects, *Annals of Statistics* **4**: 384–395.

Harville, D. A. (1997). *Matrix Algebra From A Statistician's Perspective*, Springer, New York.

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

Hastie, T. and Tibshirani, R. (1993). Varying coefficient model, *Journal of the Royal Statistical Society B* **55**: 757–796.

Heckman, N. (1997). *The theory and application of penalized least squares methods or reproducing kernel Hilbert spaces made easy*, University of British Columbia Statistics Department Technical Report number 216.

Heckman, N. and Ramsay, J. O. (2000). Penalized regression with model-based penalties, *Canadian Journal of Statistics* **28**: 241–258.

Jarrow, R., Ruppert, D. and Yu, Y. (2004). Estimating the term structure of corporate debt with a semiparametric penalized spline model, *Journal of the American Statistical Association* **99**: 57–66.

Ke, C. and Wang, Y. (2001). Semi-parametric nonlinear mixed-effects models and their applications (with discussion), *Journal of the American Statistical Association* **96**: 1272–1298.

Ke, C. and Wang, Y. (2004). Nonparametric nonlinear regression models, *Journal of the American Statistical Association* **99**: 1166–1175.

Kimeldorf, G. S. and Wahba, G. (1971). Some results on Tchebycheffian spline functions, *Journal of Mathematical Analysis and Applications* **33**: 82–94.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1988). Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy, *Journal of the American Medical Association* **260**: 2864–2871.

Klein, R., Klein, B. E. K., Moss, S. E., Davis, M. D. and DeMets, D. L. (1989). Is blood pressure a predictor of the incidence or progression of diabetic retinopathy, *Archives of Internal Medicine* **149**: 2427–2432.

Kronfol, Z., Nair, M., Zhang, Q., Hill, E. and Brown, M. (1997). Circadian immune measures in healthy volunteers: Relationship to hypothalamic-pituitary-adrenal axis hormones and sympathetic neurotransmitters, *Psychosomatic Medicine* **59**: 42–50.

Lawton, W. H., Sylvestre, E. A. and Maggio, M. S. (1972). Self-modeling nonlinear regression, *Technometrics* **13**: 513–532.

Lee, Y., Nelder, J. A. and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Chapman and Hall, London.

Li, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing, *Annals of Statistics* **14**: 1101–1112.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data, *Biometrics* **46**: 673–687.

Liu, A. and Wang, Y. (2004). Hypothesis testing in smoothing spline models, *Journal of Statistical Computation and Simulation* **74**: 581–597.

Liu, A., Meiring, W. and Wang, Y. (2005). Testing generalized linear models using smoothing spline methods, *Statistica Sinica* **15**: 235–256.

Liu, A., Tong, T. and Wang, Y. (2007). Smoothing spline estimation of variance functions, *Journal of Computational and Graphical Statistics* **16**: 312–329.

Ma, X., Dai, B., Klein, R., Klein, B. E. K., Lee, K. and Wahba, G. (2010). *Penalized likelihood regression in reproducing kernel Hilbert spaces with randomized covariate data*, University of Wisconsin Statistics Department Technical Report number 1158.

McCullagh, P. and Nelder, J. (1989). *Generalized Linear Models*, Chapman and Hall, London.

Meinguet, J. (1979). Multivariate interpolation at arbitrary points made simple, *Journal of Applied Mathematics and Physics (ZAMP)* **30**: 292–304.

Neal, D. (2004). *Introduction to Population Biology*, Cambridge University Press, Cambridge, UK.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines, *Journal of the American Statistical Association* **83**: 1134–1143.

Opsomer, J. D., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors, *Statistical Science* **16**: 134–153.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems (with discussion), *Statistical Science* **4**: 502–527.

Parzen, E. (1961). An approach to time series analysis, *Annals of Mathematical Statistics* **32**: 951–989.

Pinheiro, J. and Bates, D. M. (2000). *Mixed-effects Models in S and S-plus*, Springer, New York.

Qin, L. and Wang, Y. (2008). Nonparametric spectral analysis with applications to seizure characterization using EEG time series, *Annals of Applied Statistics* **2**: 1432–1451.

Ramsay, J. O. (1998). Estimating smooth monotone functions, *Journal of the Royal Statistical Society B* **60**: 365–375.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, 2nd ed.*, Springer, New York.

Rice, J. A. (1984). Bandwidth choice for nonparametric regression, *Annals of Statistics* **12**: 1215–1230.

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

Ruppert, D., Wand, M. P. and Carroll, R. J. (2003). *Semiparametric Regression*, Cambridge, New York.

Schumaker, L. L. (2007). *Spline Functions: Basic Theory, 3rd ed.*, Cambridge University Press, Cambridge, UK.

Smith, M. and Kohn, R. (2000). Nonparametric seemingly unrelated regression, *Journal of Econometrics* **98**: 257–281.

Speckman, P. (1995). Fitting curves with features: semiparametric change-point methods, *Computing Science and Statistics* **26**: 257–264.

Stein, M. (1990). A comparison of generalized cross-validation and modified maximum likelihood for estimating the parameters of a stochastic process, *Annals of Statistics* **18**: 1139–1157.

Tibshirani, R. and Knight, K. (1999). The covariance inflation criterion for adaptive model selection, *Journal of the Royal Statistical Society B* **61**: 529–546.

Tong, T. and Wang, Y. (2005). Estimating residual variance in nonparametric regression using least squares, *Biometrika* **92**: 821–830.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S, 4th ed.*, Springer, New York.

Wahba, G. (1980). Automatic smoothing of the log periodogram, *Journal of the American Statistical Association* **75**: 122–132.

Wahba, G. (1981). Spline interpolation and smoothing on the sphere, *SIAM Journal on Scientific Computing* **2**: 5–16.

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline, *Journal of the Royal Statistical Society B* **45**: 133–150.

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameters in the generalized spline smoothing problem, *Annals of Statistics* **4**: 1378–1402.

Wahba, G. (1987). Three topics in ill posed inverse problems, pp. 37–51. In *Inverse and Ill-Posed Problems*, M. Engl and G. Groetsch, eds. Academic Press, New York.

Wahba, G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia, PA. CBMS-NSF Regional Conference Series in Applied Mathematics, Vol. 59.

Wahba, G. and Wang, Y. (1995). Behavior near zero of the distribution of GCV smoothing parameter estimates for splines, *Statistics and Probability Letters* **25**: 105–111.

Wahba, G., Wang, Y., Gu, C., Klein, R. and Klein, B. E. K. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy, *Annals of Statistics* **23**: 1865–1895.

Wang, Y. (1994). *Smoothing Spline Analysis of Variance of Data From Exponential Families*, Ph.D. Thesis, University of Wisconsin-Madison, Department of Statistics.

Wang, Y. (1997). GRKPACK: fitting smoothing spline analysis of variance models to data from exponential families, *Communications in Statistics: Simulation and Computation* **26**: 765–782.

Wang, Y. (1998a). Mixed-effects smoothing spline ANOVA, *Journal of the Royal Statistical Society B* **60**: 159–174.

Wang, Y. (1998b). Smoothing spline models with correlated random errors, *Journal of the American Statistical Association* **93**: 341–348.

Wang, Y. and Brown, M. B. (1996). A flexible model for human circadian rhythms, *Biometrics* **52**: 588–596.

Wang, Y. and Ke, C. (2009). Smoothing spline semi-parametric nonlinear regression models, *Journal of Computational and Graphical Statistics* **18**: 165–183.

Wang, Y. and Wahba, G. (1995). Bootstrap confidence intervals for smoothing splines and their comparison to Bayesian confidence intervals, *Journal of Statistical Computation and Simulation* **51**: 263–279.

Wang, Y. and Wahba, G. (1998). Discussion of "Smoothing Spline Models for the Analysis of Nested and Crossed Samples of Curves" by Brumback and Rice, *Journal of the American Statistical Association* **93**: 976–980.

Wang, Y., Guo, W. and Brown, M. B. (2000). Spline smoothing for bivariate data with applications to association between hormones, *Statistica Sinica* **10**: 377–397.

Wang, Y., Ke, C. and Brown, M. B. (2003). Shape invariant modelling of circadian rhythms with random effects and smoothing spline ANOVA decomposition, *Biometrics* **59**: 804–812.

Wang, Y., Wahba, G., Chappell, R. and Gu, C. (1995). Simulation studies of smoothing parameter estimates and Bayesian confidence intervals in Bernoulli SS ANOVA models, *Communications in Statistics: Simulation and Computation* **24**: 1037–1059.

Wong, W. (2006). Estimation of the loss of an estimate. In *Frontiers in Statistics*, J. Fan and H. L. Koul eds. Imperial College Press, London.

Wood, S. N. (2003). Thin plate regression splines, *Journal of the Royal Statistical Society B* **65**: 95–114.

Xiang, D. and Wahba, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data, *Statistica Sinica* **6**: 675–692.

Yang, Y., Liu, A. and Wang, Y. (2005). Detecting pulsatile hormone secretions using nonlinear mixed effects partial spline models, *Biometrics* pp. 230–238.

Ye, J. M. (1998). On measuring and correcting the effects of data mining and model selection, *Journal of the American Statistical Association* **93**: 120–131.

Yeshurun, Y., Malozemoff, A. P. and Shaulov, A. (1996). Magnetic relaxation in high-temperature superconductors, *Reviews of Modern Physics* **68**: 911–949.

Yorke, J. A. and London, W. P. (1973). Recurrent outbreaks of measles, chickenpox and mumps, *American Journal of Epidemiology* **98**: 453–482.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single index models, *Journal of the American Statistical Association* **97**: 1042–1054.

Yuan, M. and Wahba, G. (2004). Doubly penalized likelihood estimator in heteroscedastic regression, *Statistics and Probability Letters* **69**: 11–20.

Statistics

A general class of powerful and flexible modeling techniques, spline smoothing has attracted a great deal of research attention in recent years and has been widely used in many application areas, from medicine to economics. **Smoothing Splines: Methods and Applications** covers basic smoothing spline models, including polynomial, periodic, spherical, thin-plate, L-, and partial splines, as well as more advanced models, such as smoothing spline ANOVA, extended and generalized smoothing spline ANOVA, vector spline, nonparametric nonlinear regression, semiparametric regression, and semiparametric mixed-effects models. It also presents methods for model selection and inference.

The book provides unified frameworks for estimation, inference, and software implementation by using the general forms of nonparametric/semiparametric, linear/nonlinear, and fixed/mixed smoothing spline models. The theory of reproducing kernel Hilbert space (RKHS) is used to present various smoothing spline models in a unified fashion. Although this approach can be technical and difficult, the author makes the advanced smoothing spline methodology based on RKHS accessible to practitioners and students. He offers a gentle introduction to RKHS, keeps theory at a minimum level, and explains how RKHS can be used to construct spline models.

Smoothing Splines offers a balanced mix of methodology, computation, implementation, software, and applications. It uses R to perform all data analyses and includes a host of real data examples from astronomy, economics, medicine, and meteorology. The codes for all examples, along with related developments, can be found on the book's web page.

CRC Press
Taylor & Francis Group
an **informa** business
www.crcpress.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

C7755

ISBN: 978-1-4200-7755-1