Springer Series in Statistics

Smoothing Spline ANOVA Models

Second Edition

Springer

Springer Series in Statistics 297

For further volumes:
http://www.springer.com/series/692

Chong Gu

Smoothing Spline ANOVA Models

Second Edition

Springer

Chong Gu
Department of Statistics
Purdue University
West Lafayette, IN 47907
USA

ISSN 0172-7397
ISBN 978-1-4614-5368-0 ISBN 978-1-4614-5369-7 (eBook)
DOI 10.1007/978-1-4614-5369-7
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012950795

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To my father
For the books and the bookcases

Preface to the First Edition

Thirty years have passed since the pioneering work of Kimeldorf and Wahba (1970a, 1970b, 1971) and Good and Gaskins (1971), and during this time, a rich body of literature has been developed on smoothing methods with roughness penalties. There have been two books solely devoted to the subject prior to this one, of which Wahba (1990) compiled an excellent synthesis for work up to that date, and Green and Silverman (1994) provided a mathematically gentler introduction to the field through regression models that are largely univariate.

Much has happened in the past decade, and more has been done with the penalty method than just regression. In this book, I have tried to assemble a comprehensive treatment of penalty smoothing under a unified framework. Treated are (i) regression with Gaussian and non-Gaussian responses as well as with censored lifetime data, (ii) density and conditional density estimation under a variety of sampling schemes, and (iii) hazard rate estimation with censored lifetime data and covariates. The unifying themes are the general penalized likelihood method and the construction of multivariate models with certain ANOVA decompositions built in. Extensive discussions are devoted to model (penalty) construction, smoothing parameter selection, computation, and asymptotic convergence. There are, however, many omissions, and the selection and treatment of topics solely reflect my personal preferences and views. Most of the materials have appeared in the literature, but a few items are new, as noted in the bibliographic notes at the end of the chapters.

An adequate treatment of model construction in the context requires some elementary knowledge of reproducing kernel Hilbert spaces, of which a self-contained introduction is included early in the book; the materials should be accessible to a second-year graduate student with a good training in calculus and linear algebra. Also assumed is a working knowledge of basic statistical inference such as linear models, maximum likelihood estimates, etc. To better understand materials on hazard estimation, prior knowledge of basic survival analysis would also help.

Most of the computational and data analytical tools discussed in the book are implemented in R, an open-source clone of the popular S/Splus language. Code for regression is reasonably polished and user-friendly and has been distributed in the R package `gss` available through CRAN, the Comprehensive R Archive Network, with the master site at

http://cran.r-project.org

The use of `gss` facilities is illustrated in the book through simulated and real-data examples.

Remaining on my wish list are (i) polished, user-friendly software tools for density estimation and hazard estimation, (ii) fast computation via approximate solutions of penalized likelihood problems, and (iii) handling of parametric random effects such as those appearing in longitudinal models and hazard models with frailty. All of the above are under active development and could be addressed in a later edition of the book or, sooner than that, in later releases of `gss`.

The book was conceived in Spring 1996 when I was on leave at the Department of Statistics, University of Michigan, which offered me the opportunity to teach a course on the subject. Work on the book has been on and off since then, with much of the progress being made in the 1997–1998 academic year during my visit at the National Institute of Statistical Sciences, and in Fall 2000 when I was teaching a course on the subject at Purdue.

I am indebted to Grace Wahba, who taught me smoothing splines, and to Doug Bates, who taught me statistical computing. Bill Studden carefully read various drafts of Chaps. 1, 2, and 4; his questions alerted me to numerous accounts of mathematical sloppiness in the text and his suggestions led to much improved presentations. Detailed comments and suggestions by Nancy Heckman on a late draft helped me to fix numerous problems throughout the first five chapters and to shape the final organization of the book (e.g., the inclusion of §1.4). For various ways in which they helped, I would also like to thank Mary Ellen Bock, Jerry Davis, Nels Grevstad, Wensheng Guo, Alan Karr, Youngju Kim, Ping Ma, Jerry Sacks, Jingyuan Wang, Yuedong Wang, Jeff Wu, Dong Xiang, Liqing Yan, and the classes at Michigan and Purdue. Last but not least, I would like to thank the R Core Team, for creating a most enjoyable platform for statistical computing.

West Lafayette, Indiana Chong Gu
July 2001

Preface

When the first edition was published a decade ago, I wrote in the Preface:

> Remaining on my wish list are (i) polished, user-friendly software tools for density estimation and hazard estimation, (ii) fast computation via approximate solutions of penalized likelihood problems, and (iii) handling of parametric random effects such as those appearing in longitudinal models and hazard models with frailty.

I am happy to report that the wishes have been fulfilled, plus some more, and it is time to present an updated treatise on smoothing methods with roughness penalties.

The developments of software tools embodied in an R package `gss` have gone a long way in the past decade, with the user-interface polished, functionality expanded, and/or numerical efficiency improved from release to release. The primary objective of this new edition is to introduce extensive software illustrations to complement the theoretical and methodological discussions, so the reader not only can read about the methods but also can use them in everyday data analysis.

Newly developed theoretical, methodological, and computational techniques are integrated in a few new chapters and new sections, along with some previously omitted entries; due modifications are made in related chapters and sections to maintain coherence. Empirical studies are expanded, reorganized, and mostly rerun using the latest software.

Two appendices are also added. One appendix outlines the overall design of the R package `gss`. The other presents some conceptual critiques on a few issues concerning smoothing methods at large, which are potentially controversial.

Much of the new materials that went into this edition were taken from or inspired by collaborations or communications with Pang Du, Anouschka Foltz, Chun Han, Young-Ju Kim, Yi Lin, Ping Ma, Christophe Pouzat, Jingyuan Wang, and Tonglin Zhang, to whom I owe thanks. I can not thank enough the R Core Team, for creating and maintaining a most enjoyable platform for statistical computing.

West Lafayette, Indiana Chong Gu
August 2011

Contents

1
Introduction

Data and models are two sources of information in a statistical analysis. Data carry noise but are "unbiased," whereas models, effectively a set of constraints, help to reduce noise but are responsible for "biases." Representing the two extremes on the spectrum of "bias-variance" trade-off are standard parametric models and constraint-free nonparametric "models" such as the empirical distribution for a probability density. In between the two extremes, there exist scores of nonparametric or semiparametric models, of which most are also known as smoothing methods. A family of such nonparametric models in a variety of stochastic settings can be derived through the penalized likelihood method, forming the subject of this book.

The general penalized likelihood method can be readily abstracted from the cubic smoothing spline as the solution to a minimization problem, and its applications in regression, density estimation, and hazard estimation set out the subject of study (§1.1). Some general notation is set in §1.2. Multivariate statistical models can often be characterized through function decompositions similar to the classical analysis of variance (ANOVA) decomposition, which we discuss in §1.3. To illustrate the potential applications of the methodology, previews of selected case studies are presented in §1.4. Brief summaries of the chapters to follow are given in §1.5.

C. Gu, *Smoothing Spline ANOVA Models*, Springer Series in Statistics 297, DOI 10.1007/978-1-4614-5369-7_1,

1.1 Estimation Problem and Method

The problem to be addressed in this book is flexible function estimation based on stochastic data. To allow for flexibility in the estimation of η, say, soft constraints of the form $J(\eta) \leq \rho$ are used in lieu of the rigid constraints of parametric models, where $J(\eta)$ quantifies the roughness of η and ρ sets the allowance; an example of $J(\eta)$ for η on $[0,1]$ is $\int_0^1 (d^2\eta/dx^2)^2 dx$. Solving the constrained maximum likelihood problem by the Lagrange method, one is led to the penalized likelihood method.

In what follows, a brief discussion of the cubic smoothing spline helps to motivate the idea, and a simple simulation illustrates the role of ρ through the Lagrange multiplier, better known as the smoothing parameter in the context. Following a straightforward abstraction, the penalized likelihood method is exemplified in regression, density estimation, and hazard estimation.

1.1.1 Cubic Smoothing Spline

Consider a regression problem $Y_i = \eta(x_i) + \epsilon_i$, $i = 1, \ldots, n$, where $x_i \in [0,1]$ and $\epsilon_i \sim N(0, \sigma^2)$. In a classical parametric regression analysis, η is assumed to be of form $\eta(x, \boldsymbol{\beta})$, known up to the parameters $\boldsymbol{\beta}$, which are to be estimated from the data. When $\eta(x, \boldsymbol{\beta})$ is linear in $\boldsymbol{\beta}$, one has a standard linear model. A parametric model characterizes a set of rigid constraints on η. The dimension of the model space (i.e., the number of unknown parameters) is typically much smaller than the sample size n.

To avoid possible model misspecification in a parametric analysis, otherwise known as bias, an alternative approach to estimation is to allow η to vary in a high-dimensional (possibly infinite) function space, leading to various nonparametric or semiparametric estimation methods. A popular approach to the nonparametric estimation of η is via the minimization of a penalized least squares score,

$$\frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2 + \lambda \int_0^1 \ddot{\eta}^2 dx, \tag{1.1}$$

with $\ddot{\eta} = d^2\eta/dx^2$, where the first term discourages the lack of fit of η to the data, the second term penalizes the roughness of η, and the smoothing parameter λ controls the trade-off between the two conflicting goals. The minimization of (1.1) is implicitly over functions with square integrable second derivatives. The minimizer η_λ of (1.1) is called a cubic smoothing spline. As $\lambda \to 0$, η_λ approaches the minimum curvature interpolant. As $\lambda \to \infty$, η_λ approaches the simple linear regression line. Note that the linear polynomials $\{f : f = \beta_0 + \beta_1 x\}$ form the so-called null space of the roughness penalty $\int_0^1 \ddot{f}^2 dx$, $\{f : \int_0^1 \ddot{f}^2 dx = 0\}$.

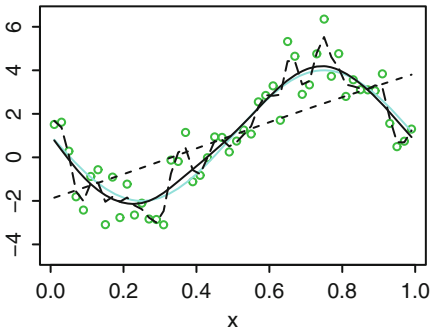

FIGURE 1.1. Cubic smoothing splines. The test function is in the *faded line* and the estimates are in the *solid*, *dashed*, and *long-dashed lines*. The data are superimposed as *circles*.

To illustrate, consider a simple simulation with $x_i = (i - 0.5)/50$, $i = 1, \ldots, 50$, $\eta(x) = 1 + 3\sin(2\pi x - \pi)$, and $\sigma^2 = 1$. The estimate η_λ was calculated at $\log_{10} n\lambda = 0, -3, -6$. Plotted in Fig. 1.1 are the test function (faded line), the estimates (solid, dashed, and long-dashed lines), and the data (circles). The rough fit corresponds to $\log_{10} n\lambda = -6$, the near straight line to $\log_{10} n\lambda = 0$, and the close fit to $\log_{10} n\lambda = -3$.

An alternative derivation of the cubic smoothing spline is through a constrained least squares problem, which solves

$$\min \frac{1}{n}\sum_{i=1}^{n}\big(Y_i - \eta(x_i)\big)^2, \quad \text{subject to} \quad \int_0^1 \ddot{\eta}^2 dx \le \rho, \tag{1.2}$$

for some $\rho \ge 0$. The solution to (1.2) usually falls on the boundary of the permissible region, $\int_0^1 \ddot{\eta}^2 dx = \rho$, and by the Lagrange method, it can be calculated as the minimizer of (1.1) with an appropriate Lagrange multiplier λ. Thus, up to the choices of λ and ρ, a penalized least squares problem with a penalty proportional to $\int_0^1 \ddot{\eta}^2 dx$ is equivalent to a constrained least squares problem subject to a soft constraint of the form $\int_0^1 \ddot{\eta}^2 dx \le \rho$; see, e.g., Schoenberg (1964). See also §2.6.2.

Defined as the solution to a penalized optimization problem, a smoothing spline is also known as a natural spline in the numerical analysis literature. The minimizer η_λ of (1.1) is called a cubic spline because it is a piecewise cubic polynomial. It is three times differentiable, with the third derivative jumping at the knots $\xi_1 < \xi_2 < \cdots < \xi_q$, the ordered distinctive sampling points x_i, and it is linear beyond the first knot ξ_1 and the last knot ξ_q. See Schumaker (1981, Chap. 8) for a comprehensive treatment of smoothing splines from a numerical analytical perspective. See also de Boor (1978).

1.1.2 Penalized Likelihood Method

The cubic smoothing spline of (1.1) is a specialization of the general penalized likelihood method in univariate Gaussian regression. To estimate a function of interest η on a generic domain $\mathcal{X}$ using stochastic data, one may use the minimizer of

$$L(\eta|\text{data}) + \frac{\lambda}{2} J(\eta), \tag{1.3}$$

where $L(\eta|\text{data})$ is usually taken as the minus log likelihood of the data and $J(f)$ is a quadratic roughness functional with a null space $\mathcal{N}_J = \{f : J(f) = 0\}$ of low dimension; see §2.1.1 for the definition of quadratic functional. The solution of (1.3) is the maximum likelihood estimate in a model space $\mathcal{M}_\rho = \{f : J(f) \le \rho\}$ for some $\rho \ge 0$, and the smoothing parameter λ in (1.3) is the Lagrange multiplier. See §2.6.2 for a detailed discussion of the role of λ as a Lagrange multiplier.

A few examples of penalized likelihood estimation follow.

Example 1.1 (Response data regression) Assume

$$Y|x \sim \exp\big\{\big(y\eta(x) - b(\eta(x))\big)/a(\phi) + c(y,\phi)\big\},$$

an exponential family density with a modeling parameter η and a possibly unknown nuisance parameter ϕ. Observing independent data (x_i, Y_i), $i = 1, \ldots, n$, the method estimates η via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\big\{Y_i\eta(x_i) - b(\eta(x_i))\big\} + \frac{\lambda}{2}J(\eta). \tag{1.4}$$

When the density is Gaussian, (1.4) reduces to a penalized least squares problem; see Problem 1.1. Penalized least squares regression for Gaussian-type responses is the subject of Chap. 3. Penalized likelihood regression for non-Gaussian responses will be studied in Chap. 5. □

Example 1.2 (Density estimation) Observing independent and identically distributed samples X_i, $i = 1, \ldots, n$ from a probability density $f(x)$ supported on a bounded domain $\mathcal{X}$, the method estimates f by $e^\eta / \int_{\mathcal{X}} e^\eta dx$, where η minimizes

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\eta(X_i) - \log\int_{\mathcal{X}} e^{\eta(x)}dx\right\} + \frac{\lambda}{2}J(\eta). \tag{1.5}$$

A side condition, say $\int_{\mathcal{X}} \eta\, dx = 0$, shall be imposed on η for a one-to-one transform $f \leftrightarrow e^\eta / \int_{\mathcal{X}} e^\eta dx$. Penalized likelihood density estimation is the subject of Chap. 7. □

Example 1.3 (Hazard estimation) Let T be the lifetime of an item with survival function $S(t|u) = P(T > t|u)$, possibly dependent on a covariate U. The hazard function is defined as $e^{\eta(t,u)} = -\partial \log S(t|u)/\partial t$. Let Z be the left-truncation time and C be the right-censoring time, independent of T and of each other. Observing $(U_i, Z_i, X_i, \delta_i)$, $i = 1, \ldots, n$, where $X = \min(T, C)$, $\delta = I_{[T \le C]}$, and $Z < X$, the method estimates the log hazard η via the minimization of

$$-\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_i \eta(X_i, U_i) - \int_{Z_i}^{X_i} e^{\eta(t,U_i)}dt\right\} + \frac{\lambda}{2}J(\eta); \tag{1.6}$$

see Problem 1.2 for the derivation of the likelihood. Penalized likelihood hazard estimation will be studied in Chap. 8. □

The two basic components of a statistical model, the deterministic part and the stochastic part, are well separated in (1.3). The structure of the deterministic part is determined by the construction of $J(\eta)$ for η on a domain $\mathcal{X}$, of which a comprehensive treatment is presented in Chap. 2. The stochastic part is reflected in the likelihood $L(\eta|\text{data})$ and determines, among other things, the natural measures with which the performance of the estimate is to be assessed. The minimizer of (1.3) with a varying λ defines a family of estimates, and from the cubic spline simulation shown in Fig. 1.1, we have seen how differently the family members may behave. Data-driven procedures for the proper selection of the smoothing parameter are crucial to the practicability of penalized likelihood estimation, to which extensive discussion will be devoted in the settings of regression, density estimation, and hazard estimation in their respective chapters.

1.2 Notation

Listed below is some general notation used in this book. Context-specific or subject-specific notation may differ from that listed here, in which case every effort will be made to avoid possible confusion.

Domains are usually denoted by $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, etc., or subscripted as $\mathcal{X}_1$, $\mathcal{X}_2$, etc. Points on domains are usually denoted by $x \in \mathcal{X}$, $y \in \mathcal{Y}$, or $x_1, x_2, y \in \mathcal{X}$. Points on product domains are denoted by $x_1, x_2, y \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$, with $x_{1\langle 1\rangle}, x_{2\langle 1\rangle}, y_{\langle 1\rangle} \in \mathcal{X}_1$ and $x_{1\langle 2\rangle}, x_{2\langle 2\rangle}, y_{\langle 2\rangle} \in \mathcal{X}_2$, or by $z = (x, y) \in \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$, with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Ordinary subscripts are used to denote multiple points on a domain, but *not* coordinates of a point on a product domain.

Function spaces are usually denoted by $\mathcal{H}$, $\mathcal{G}$, etc. Functions in function spaces are usually denoted by $f, g, h \in \mathcal{H}$, $\eta, \phi, \xi \in \mathcal{H}$, etc. Derivatives of a univariate function $f(x)$ are denoted by $\dot{f} = df/dx$, $\ddot{f} = d^2 f/dx^2$,

or by the general notation $f^{(m)} = d^m f/dx^m$. Derivatives of multivariate functions $f(x_{\langle 1\rangle}, x_{\langle 2\rangle})$ on $\mathcal{X}_1 \times \mathcal{X}_2$ or $g(x, y)$ on $\mathcal{X} \times \mathcal{Y}$ are denoted by $f^{(3)}_{\langle 112\rangle} = \partial^3 f/\partial x^2_{\langle 1\rangle} \partial x_{\langle 2\rangle}$, $\ddot{g}_{\langle xy\rangle} = \partial^2 g/\partial x \partial y$, etc.

Matrices are denoted by the standard notation of uppercase letters. Vectors, however, are often *not* denoted by boldface letters in this book. For a point on a product domain $\mathcal{X} = \prod_{\gamma=1}^{\Gamma} \mathcal{X}_\gamma$, we write $x = (x_{\langle 1\rangle}, \ldots, x_{\langle \Gamma\rangle})$. For a function on domain $\mathcal{X} = \{1, \ldots, K\}$, we write $f = \big(f(1), \ldots, f(K)\big)^T$, which may be used as a vector in standard matrix arithmetic. Boldface vectors are used where confusion may result otherwise. For example, $\mathbf{1} = (1, \ldots, 1)^T$ is used to denote a vector of all one's, and $\mathbf{c} = (c_1, \ldots, c_n)^T$ is used to encapsulate subscripted coefficients. In formulas concerning matrix computation, vectors are always set in boldface.

The standard O_p, o_p notation is used in the asymptotic analyses of §§3.2, 4.2.3, 5.2, 6.2, 6.3, Chap. 9, §§10.2, and 10.5. If $P\big(|X| > KY\big) \to 0$ for some constant $K < \infty$, we write $X = O_p(Y)$, and when $P\big(|X| > \epsilon Y\big) \to 0$, $\forall \epsilon > 0$, we denote $X = o_p(Y)$.

1.3 Decomposition of Multivariate Functions

An important aspect of statistical modeling, which distinguishes it from mere function approximation, is the interpretability of the results. Of great utility are decomposition of multivariate functions similar to the classical analysis of variance (ANOVA) decomposition and the associated notions of main effect and interaction. Higher-order interactions are often excluded in practical estimation to control model complexity; the exclusion of all interactions yields the popular additive models. Selective exclusion of certain interactions also characterizes many interesting statistical models in a variety of stochastic settings.

Casting the classical one-way ANOVA decomposition as the decomposition of functions on a discrete domain, a simple averaging operator is introduced to facilitate the generalization of the notion to arbitrary domains. Multiway ANOVA decomposition is then defined, with the identifiability of the terms assured by side conditions specified through the averaging operators. Examples are given and a proposition is proved concerning certain intrinsic structures that are independent of the side conditions. The utility and implication of selective term trimming in an ANOVA decomposition are then briefly discussed in the context of regression, density estimation, and hazard estimation.

1.3.1 ANOVA Decomposition and Averaging Operator

Consider a standard one-way ANOVA model, $Y_{ij} = \mu_i + \epsilon_{ij}$, where μ_i are the treatment means at treatment levels $i = 1, \ldots, K$ and ϵ_{ij} are

independent normal errors. Writing $\mu_i = \mu + \alpha_i$, one has the "overall mean" μ and the treatment effect α_i. The identifiability of μ and α_i are assured through a side condition, of which common choices include $\alpha_1 = 0$ with level 1 treated as the control and $\sum_{i=1}^{K} \alpha_i = 0$ with all levels treated symmetrically.

The one-way ANOVA model can be recast as $Y_j = f(x_j) + \epsilon_j$, where $f(x)$ is defined on the discrete domain $\mathcal{X} = \{1, \ldots, K\}$; the treatment levels are now coded by x and the subscript j labels the observations. The ANOVA decomposition $\mu_i = \mu + \alpha_i$ in the standard ANOVA model notation can be written as

$$f(x) = Af + (I - A)f = f_\emptyset + f_x,$$

where A is an averaging operator that "averages out" the argument x to return a constant function and I is the identity operator. For example, with $Af = f(1)$, one has $f(x) = f(1) + \{f(x) - f(1)\}$, corresponding to $\alpha_1 = 0$. With $Af = \sum_{x=1}^{K} f(x)/K = \bar{f}$, one has $f(x) = \bar{f} + \big(f(x) - \bar{f}\big)$, corresponding to $\sum_{i=1}^{K} \alpha_i = 0$. Note that applying A to a constant function returns that constant, hence the name "averaging." It follows that $A(Af) = Af$, $\forall f$, or, simply, $A^2 = A$. The constant term $f_\emptyset = Af$ is the "overall mean" and the term $f_x = (I - A)f$ is the treatment effect, or "contrast," that satisfies the side condition $Af_x = 0$.

On a continuous domain, say $\mathcal{X} = [a, b]$, one may similarly define an ANOVA decomposition $f(x) = Af + (I - A)f = f_\emptyset + f_x$ through an appropriately defined averaging operator A, where f_x satisfies the side condition $Af_x = 0$. For example, with $Af = f(a)$, one has $f(x) = f(a) + \{f(x) - f(a)\}$. Similarly, with $Af = \int_a^b f dx/(b - a)$, one has $f(x) = \int_a^b f dx/(b - a) + \{f(x) - \int_a^b f dx/(b - a)\}$.

1.3.2 Multiway ANOVA Decomposition

Now consider a function $f(x) = f(x_{\langle 1 \rangle}, \ldots, x_{\langle \Gamma \rangle})$ on a product domain $\mathcal{X} = \prod_{\gamma=1}^{\Gamma} \mathcal{X}_\gamma$, where $x_{\langle \gamma \rangle} \in \mathcal{X}_\gamma$ denotes the γth coordinate of $x \in \mathcal{X}$. Let A_γ be an averaging operator on $\mathcal{X}_\gamma$ that averages out $x_{\langle \gamma \rangle}$ from the active argument list and satisfies $A_\gamma^2 = A_\gamma$; $A_\gamma f$ is constant on the $\mathcal{X}_\gamma$ axis but not necessarily an overall constant function. An ANOVA decomposition of f can be defined as

$$f = \left\{ \prod_{\gamma=1}^{\Gamma} (I - A_\gamma + A_\gamma) \right\} f = \sum_{\mathcal{S}} \left\{ \prod_{\gamma \in \mathcal{S}} (I - A_\gamma) \prod_{\gamma \notin \mathcal{S}} A_\gamma \right\} f = \sum_{\mathcal{S}} f_{\mathcal{S}}, \quad (1.7)$$

where $\mathcal{S} \subseteq \{1, \ldots, \Gamma\}$ enlists the active arguments in $f_{\mathcal{S}}$ and the summation is over all of the 2^Γ subsets of $\{1, \ldots, \Gamma\}$. The term $f_\emptyset = \prod A_\gamma f$ is a constant, the term $f_\gamma = f_{\{\gamma\}} = (I - A_\gamma) \prod_{\alpha \neq \gamma} A_\alpha f$ is the $x_{\langle \gamma \rangle}$ main effect,

the term $f_{\gamma,\delta} = f_{\{\gamma,\delta\}} = (I - A_\gamma)(I - A_\delta)\prod_{\alpha\neq\gamma,\delta} A_\alpha f$ is the $x_{\langle\gamma\rangle}$-$x_{\langle\delta\rangle}$ interaction, and so forth. The terms of such a decomposition satisfy the side conditions $A_\gamma f_{\mathcal{S}} = 0, \forall \mathcal{S} \ni \gamma$. The choices of A_γ, or the side conditions on each axes, are open to specification.

The domains $\mathcal{X}_\gamma$ are generic in the above discussion; in particular, they can be product domains themselves. As a matter of fact, the ANOVA decomposition of (1.7) can also be defined recursively through a series of nested constructions with $\Gamma = 2$; see, e.g., Problem 1.3.

The ANOVA decomposition can be built into penalized likelihood estimation through the proper construction of the roughness functional $J(f)$; details are to be found in §2.4.

Example 1.4 When $\Gamma = 2$, $\mathcal{X}_1 = \{1, \ldots, K_1\}$, and $\mathcal{X}_2 = \{1, \ldots, K_2\}$, the decomposition reduces to a standard two-way ANOVA decomposition. With averaging operators $A_1 f = f(1, x_{\langle 2\rangle})$ and $A_2 f = f(x_{\langle 1\rangle}, 1)$, one has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = f(1,1), \\ f_1 &= (I - A_1)A_2 f = f(x_{\langle 1\rangle}, 1) - f(1,1), \\ f_2 &= A_1(I - A_2)f = f(1, x_{\langle 2\rangle}) - f(1,1), \\ f_{1,2} &= (I - A_1)(I - A_2)f \\ &= f(x_{\langle 1\rangle}, x_{\langle 2\rangle}) - f(x_{\langle 1\rangle}, 1) - f(1, x_{\langle 2\rangle}) + f(1,1). \end{aligned}$$

With $A_\gamma f = \sum_{x_{\langle\gamma\rangle}=1}^{K_\gamma} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_\gamma$, $\gamma = 1, 2$, one similarly has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = f_{\cdot\cdot}, \\ f_1 &= (I - A_1)A_2 f = f_{x_{\langle 1\rangle}\cdot} - f_{\cdot\cdot}, \\ f_2 &= A_1(I - A_2)f = f_{\cdot x_{\langle 2\rangle}} - f_{\cdot\cdot}, \\ f_{1,2} &= (I - A_1)(I - A_2)f \\ &= f(x_{\langle 1\rangle}, x_{\langle 2\rangle}) - f_{x_{\langle 1\rangle}\cdot} - f_{\cdot x_{\langle 2\rangle}} + f_{\cdot\cdot}, \end{aligned}$$

where $f_{\cdot\cdot} = \sum_{x_{\langle 1\rangle}, x_{\langle 2\rangle}} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_1 K_2$, $f_{x_{\langle 1\rangle}\cdot} = \sum_{x_{\langle 2\rangle}} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_2$, and $f_{\cdot x_{\langle 2\rangle}} = \sum_{x_{\langle 1\rangle}} f(x_{\langle 1\rangle}, x_{\langle 2\rangle})/K_1$. One may also use different averaging operators on different axes; see Problem 1.4. □

Example 1.5 Consider $\Gamma = 2$ and $\mathcal{X}_1 = \mathcal{X}_2 = [0,1]$. With $A_1 f = f(0, x_{\langle 2\rangle})$ and $A_2 f = f(x_{\langle 1\rangle}, 0)$, one has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = f(0,0), \\ f_1 &= (I - A_1)A_2 f = f(x_{\langle 1\rangle}, 0) - f(0,0), \\ f_2 &= A_1(I - A_2)f = f(0, x_{\langle 2\rangle}) - f(0,0), \\ f_{1,2} &= (I - A_1)(I - A_2)f \\ &= f(x_{\langle 1\rangle}, x_{\langle 2\rangle}) - f(x_{\langle 1\rangle}, 0) - f(0, x_{\langle 2\rangle}) + f(0,0). \end{aligned}$$

With $A_\gamma f = \int_0^1 f dx_{\langle\gamma\rangle}$, $\gamma = 1, 2$, one has

$$\begin{aligned} f_\emptyset &= A_1 A_2 f = \textstyle\int_0^1 \int_0^1 f dx_{\langle 1\rangle} dx_{\langle 2\rangle}, \\ f_1 &= (I - A_1) A_2 f = \textstyle\int_0^1 (f - \int_0^1 f dx_{\langle 1\rangle}) dx_{\langle 2\rangle}, \\ f_2 &= A_1 (I - A_2) f = \textstyle\int_0^1 (f - \int_0^1 f dx_{\langle 2\rangle}) dx_{\langle 1\rangle}, \\ f_{1,2} &= (I - A_1)(I - A_2) f \\ &= f - \textstyle\int_0^1 f dx_{\langle 2\rangle} - \int_0^1 f dx_{\langle 1\rangle} + \int_0^1 \int_0^1 f dx_{\langle 1\rangle} dx_{\langle 2\rangle}. \end{aligned}$$

Similar results with different averaging operators on different axes are also straightforward; see Problem 1.5. □

In standard ANOVA models, higher-order terms are frequently eliminated, whereas main effects and lower-order interactions are estimated from the data. One learns not to drop the $x_{\langle 1\rangle}$ and $x_{\langle 2\rangle}$ main effects if the $x_{\langle 1\rangle}$-$x_{\langle 2\rangle}$ interaction is considered, however, and not to drop the $x_{\langle 1\rangle}$-$x_{\langle 2\rangle}$ interaction when the $x_{\langle 1\rangle}$-$x_{\langle 2\rangle}$-$x_{\langle 3\rangle}$ interaction is included. Although the ANOVA decomposition as defined in (1.7) obviously depends on the averaging operators A_γ, certain structures are independent of the particular choices of A_γ. Specifically, for any index set $\mathcal{I}$, if $f_\mathcal{S} = 0$, $\forall \mathcal{S} \supseteq \mathcal{I}$ with a particular set of A_γ, then the structure also holds for any other choices of A_γ, as the following proposition asserts.

Proposition 1.1 *For any two sets of averaging operators A_γ and $\tilde{A}_\gamma$ satisfying $A_\gamma^2 = A_\gamma$ and $\tilde{A}_\gamma^2 = \tilde{A}_\gamma$, $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ if and only if $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma) f = 0$, where $\mathcal{I}$ is any index set.*

Note that the condition $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ means that $f_\mathcal{S} = 0$, $\forall \mathcal{S} \supseteq \mathcal{I}$. For example, $(I - A_1) f = 0$ implies that all terms involving $x_{\langle 1\rangle}$ vanish, and $(I - A_1)(I - A_2) f = 0$ means that all terms involving both $x_{\langle 1\rangle}$ and $x_{\langle 2\rangle}$ disappear. Model structures that can be characterized through constraints of the form $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ permit a term $f_\mathcal{S}$ only when all of its "subset terms," $f_{\mathcal{S}'}$ for $\mathcal{S}' \subset \mathcal{S}$, are permitted. A simple corollary of the proposition is the obvious fact that an additive model remains an additive model regardless of the side conditions.

Proof of Proposition 1.1: It is easy to see that $(I - \tilde{A}_\gamma) A_\gamma = 0$. Suppose $\prod_{\gamma\in\mathcal{I}}(I - A_\gamma) f = 0$ and define the ANOVA decomposition in (1.7) using A_γ. Now, for any nonzero term $f_\mathcal{S}$ in (1.7), one has $\mathcal{S} \not\supseteq \mathcal{I}$, so there exists $\gamma \in \mathcal{I}$ but $\gamma \notin \mathcal{S}$, hence $f_\mathcal{S} = [\cdots A_\gamma \cdots] f$. The corresponding $(I - \tilde{A}_\gamma)$ in $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma)$ then annihilates the term. It follows that all nonzero ANOVA terms in (1.7) are annihilated by $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma)$, so $\prod_{\gamma\in\mathcal{I}}(I - \tilde{A}_\gamma) f = 0$. The converse is true by symmetry. □

1.3.3 Multivariate Statistical Models

Many multivariate statistical models can be characterized by selective term elimination in an ANOVA decomposition. Some of such models are discussed below.

Curse of Dimensionality and Additive Models

Recall the classical ANOVA models with $\mathcal{X}_\gamma$ discrete. In practical data analysis, one usually includes only the main effects, with the possible addition of a few lower-order interactions. Higher-order interactions are less interpretable yet more difficult to estimate, as they usually consume many more degrees of freedom than the lower-order terms. Models with only main effects included are called additive models.

The difficulty associated with function estimation in high-dimensional spaces may be perceived through the sparsity of the space. Take $\mathcal{X}_\gamma = [0, 1]$, for example, a k-dimensional cube with each side of length 0.5 has volume 0.5^k. Assume a uniform distribution of the data and consider a piecewise constant function with jumps only possible at $x_{\langle\gamma\rangle} = 0.5$. To estimate such a function in 1 dimension with two pieces, one has information from 50 % of the data per piece, in 2 dimensions with four pieces, 25 % per piece, in 3 dimensions with eight pieces, 12.5 % per piece, etc. The lack of data due to the sparsity of high-dimensional spaces is often referred to as the curse of dimensionality. Alternatively, the curse of dimensionality may also be characterized by the explosive increase in the number of parameters, or the degrees of freedom, that one would need to approximate a function well in a high-dimensional space. To achieve the flexibility of a five-piece piecewise polynomial in 1 dimension, for example, one would end up with 125 pieces in 3 dimensions by taking products of the pieces in 1 dimension.

To combat the curse of dimensionality in multivariate function estimation, one needs to eliminate higher-order interactions to control model complexity. As with classical ANOVA models, additive models with the possible addition of second-order interactions are among the most popular models used in practice.

Conditional Independence and Graphical Models

To simplify notation, the marginal domains will be denoted by $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$, etc., in the rest of the section instead of the subscripted $\mathcal{X}$ used in (1.7).

Consider a probability density $f(x)$ of a random variable X on a domain $\mathcal{X}$. Writing

$$f(x) = \frac{e^{\eta(x)}}{\int_{\mathcal{X}} e^{\eta(x)} dx}, \tag{1.8}$$

known as the logistic density transform, the log density $\eta(x)$ is free of the positivity and unity constraints, $f(x) > 0$ and $\int_{\mathcal{X}} f(x)dx = 1$, that $f(x)$

must satisfy. The transform is not one-to-one, though, as $e^{\eta(x)}/\int_{\mathcal{X}} e^{\eta(x)}dx = e^{C+\eta(x)}/\int_{\mathcal{X}} e^{C+\eta(x)}dx$ for any constant C. The transform can be made one-to-one, however, by imposing a side condition $A_x\eta = 0$ for some averaging operator A_x on $\mathcal{X}$; this can be achieved by eliminating the constant term in a one-way ANOVA decomposition $\eta = A_x\eta + (I - A_x)\eta = \eta_\emptyset + \eta_x$.

For a joint density $f(x, y)$ of random variables (X, Y) on a product domain $\mathcal{X} \times \mathcal{Y}$, one may write

$$f(x,y) = \frac{e^{\eta(x,y)}}{\int_{\mathcal{X}} dx \int_{\mathcal{Y}} e^{\eta(x,y)}dy} = \frac{e^{\eta_x+\eta_y+\eta_{x,y}}}{\int_{\mathcal{X}} dx \int_{\mathcal{Y}} e^{\eta_x+\eta_y+\eta_{x,y}}dy},$$

where η_x, η_y, and $\eta_{x,y}$ are the main effects and interaction of $\eta(x, y)$ in an ANOVA decomposition; the constant is eliminated in the rightmost expression for a one-to-one transform. The conditional distribution of Y given X has a density

$$f(y|x) = \frac{e^{\eta(x,y)}}{\int_{\mathcal{Y}} e^{\eta(x,y)}dy} = \frac{e^{\eta_y+\eta_{x,y}}}{\int_{\mathcal{Y}} e^{\eta_y+\eta_{x,y}}dy}, \tag{1.9}$$

where the logistic conditional density transform is one-to-one only for the rightmost expression with the side conditions $A_y(\eta_y + \eta_{x,y}) = 0$, $\forall x \in \mathcal{X}$, where A_y is the averaging operator on $\mathcal{Y}$ that help to define the ANOVA decomposition. The independence of X and Y, denoted by $X \perp Y$, is characterized by $\eta_{x,y} = 0$, or $(I - A_x)(I - A_y)\eta = 0$.

The domains $\mathcal{X}$ and $\mathcal{Y}$ are generic in (1.9); in particular, they can be product domains themselves. Substituting (y, z) for y in (1.9), one has

$$f(y,z|x) = \frac{e^{\eta_y+\eta_z+\eta_{y,z}+\eta_{x,y}+\eta_{x,z}+\eta_{x,y,z}}}{\int_{\mathcal{Y}} dy \int_{\mathcal{Z}} e^{\eta_y+\eta_z+\eta_{y,z}+\eta_{x,y}+\eta_{x,z}+\eta_{x,y,z}}dz},$$

where $\eta_{(y,z)}$ is expanded out as $\eta_y + \eta_z + \eta_{y,z}$ and $\eta_{x,(y,z)}$ is expanded out as $\eta_{x,y} + \eta_{x,z} + \eta_{x,y,z}$; see Problem 1.3. The conditional independence of Y and Z given X, denoted by $(Y \perp Z)\big|X$, is characterized by $\eta_{y,z} + \eta_{x,y,z} = 0$, or $(I - A_y)(I - A_z)\eta = 0$.

Now, consider the joint density of four random variables (U, V, Y, Z), with $(U \perp V)\big|(Y, Z)$ and $(Y \perp Z)\big|(U, V)$. It can be shown that such a structure is characterized by $\eta_{u,v} + \eta_{y,z} + \eta_{u,v,y} + \eta_{u,v,z} + \eta_{u,y,z} + \eta_{v,y,z} + \eta_{u,v,y,z} = 0$ in an ANOVA decomposition, or $(I - A_u)(I - A_v)\eta = (I - A_y)(I - A_z)\eta = 0$; see Problem 1.7.

As noted above, the ANOVA decompositions in the log density η that characterize conditional independence structures are all of the type covered in Proposition 1.1. The elimination of lower-order terms in (1.8) and (1.9) for one-to-one transforms only serve to remove technical redundancies introduced by the "overparameterization" of $f(x)$ or $f(y|x)$ by the corresponding unrestricted η.