Springer Series in Statistics

Chong Gu

Smoothing Spline ANOVA Models

Second Edition

Nonparametric function estimation with stochastic data, otherwise known as smoothing, has been studied by several generations of statisticians. Assisted by the ample computing power in today's servers, desktops, and laptops, smoothing methods have been finding their ways into everyday data analysis by practitioners. While scores of methods have proved successful for univariate smoothing, ones practical in multivariate settings number far less. Smoothing spline ANOVA models are a versatile family of smoothing methods derived through roughness penalties, that are suitable for both univariate and multivariate problems.

In this book, the author presents a treatise on penalty smoothing under a unified framework. Methods are developed for (i) regression with Gaussian and non-Gaussian responses as well as with censored life time data; (ii) density and conditional density estimation under a variety of sampling schemes; and (iii) hazard rate estimation with censored life time data and covariates. The unifying themes are the general penalized likelihood method and the construction of multivariate models with built-in ANOVA decompositions. Extensive discussions are devoted to model construction, smoothing parameter selection, computation, and asymptotic convergence.

Most of the computational and data analytical tools discussed in the book are implemented in R, an open-source platform for statistical computing and graphics. Suites of functions are embodied in the R package gss, and are illustrated throughout the book using simulated and real data examples.

This monograph will be useful as a reference work for researchers in theoretical and applied statistics as well as for those in other related disciplines. It can also be used as a text for graduate level courses on the subject. Most of the materials are accessible to a second year graduate student with a good training in calculus and linear algebra and working knowledge in basic statistical inferences such as linear models and maximum likelihood estimates.

Chong Gu received his Ph.D. from University of Wisconsin–Madison in 1989, and has been on the faculty in Department of Statistics, Purdue University since 1990. At various times during his career, he has held visiting appointments at University of British Columbia, University of Michigan, and National Institute of Statistical Sciences.

Statistics

ISBN 978-1-4614-5368-0

9 781461 453680

▶ **springer.com**

Contents

Preface to the First Edition vii

Preface ix

1 Introduction 1

1.1 Estimation Problem and Method 2

1.1.1 Cubic Smoothing Spline 2

1.1.2 Penalized Likelihood Method 4

1.2 Notation 5

1.3 Decomposition of Multivariate Functions 6

1.3.1 ANOVA Decomposition and Averaging Operator . . 6

1.3.2 Multiway ANOVA Decomposition 7

1.3.3 Multivariate Statistical Models 10

1.4 Case Studies 12

1.4.1 Water Acidity in Lakes 12

1.4.2 AIDS Incubation 14

1.4.3 Survival After Heart Transplant 15

1.5 Scope 17

1.6 Bibliographic Notes 19

1.7 Problems 20

2 Model Construction **23**
2.1 Reproducing Kernel Hilbert Spaces 24
2.1.1 Hilbert Spaces and Linear Subspaces 24
2.1.2 Riesz Representation Theorem 29
2.1.3 Reproducing Kernel and Non-Negative Definite Function 29
2.2 Smoothing Splines on $\{1, \ldots, K\}$ 32
2.3 Polynomial Smoothing Splines on $[0, 1]$ 34
2.3.1 A Reproducing Kernel in $\mathcal{C}^{(m)}[0, 1]$ 34
2.3.2 Computation of Polynomial Smoothing Splines . . . 36
2.3.3 Another Reproducing Kernel in $\mathcal{C}^{(m)}[0, 1]$ 37
2.4 Smoothing Splines on Product Domains 40
2.4.1 Tensor Product Reproducing Kernel Hilbert Spaces 40
2.4.2 Reproducing Kernel Hilbert Spaces on $\{1, \ldots, K\}^2$ 41
2.4.3 Reproducing Kernel Hilbert Spaces on $[0, 1]^2$ 42
2.4.4 Reproducing Kernel Hilbert Spaces on $\{1, \ldots, K\} \times [0, 1]$ 44
2.4.5 Multiple-Term Reproducing Kernel Hilbert Spaces: General Form 45
2.5 Bayes Model 48
2.5.1 Shrinkage Estimates as Bayes Estimates 48
2.5.2 Polynomial Splines as Bayes Estimates 49
2.5.3 Smoothing Splines as Bayes Estimates 51
2.6 Minimization of Penalized Functional 51
2.6.1 Existence of Minimizer 52
2.6.2 Penalized and Constrained Optimization 53
2.7 Bibliographic Notes 54
2.8 Problems 56

3 Regression with Gaussian-Type Responses **61**
3.1 Preliminaries 62
3.2 Smoothing Parameter Selection 64
3.2.1 Unbiased Estimate of Relative Loss 65
3.2.2 Cross-Validation and Generalized Cross-Validation 67
3.2.3 Restricted Maximum Likelihood Under Bayes Model 70
3.2.4 Weighted and Replicated Data 72
3.2.5 Empirical Performance 74
3.3 Bayesian Confidence Intervals 75
3.3.1 Posterior Distribution 76
3.3.2 Confidence Intervals on Sampling Points 78
3.3.3 Across-the-Function Coverage 78

3.4 Computation: Generic Algorithms 79
3.4.1 Algorithm for Fixed Smoothing Parameters 80
3.4.2 Algorithm for Single Smoothing Parameter 80
3.4.3 Algorithm for Multiple Smoothing Parameters . . . 82
3.4.4 Calculation of Posterior Variances 84
3.5 Efficient Approximation 85
3.5.1 Preliminaries 85
3.5.2 Bayes Model 86
3.5.3 Computation 88
3.5.4 Empirical Choice of q 90
3.5.5 Numerical Accuracy 92
3.6 Software 93
3.6.1 RKPACK 93
3.6.2 R Package `gss`: `ssanova` and `ssanova0` Suites 94
3.7 Model Checking Tools 98
3.7.1 Cosine Diagnostics 98
3.7.2 Examples 99
3.7.3 Concepts and Heuristics 103
3.8 Square Error Projection 104
3.9 Case Studies 106
3.9.1 Nitrogen Oxides in Engine Exhaust 106
3.9.2 Ozone Concentration in Los Angeles Basin 107
3.10 Computation: Special Algorithms 111
3.10.1 Fast Algorithm for Polynomial Splines 112
3.10.2 Iterative Algorithms and Monte Carlo Cross-Validation 114
3.11 Bibliographic Notes 115
3.12 Problems 118

4 More Splines **125**
4.1 Partial Splines 126
4.2 Splines on the Circle 127
4.2.1 Periodic Polynomial Splines 127
4.2.2 Splines as Low-Pass Filters 128
4.2.3 More on Asymptotics of §3.2 130
4.3 Thin-Plate Splines 134
4.3.1 Semi-Kernels for Thin-Plate Splines 135
4.3.2 Reproducing Kernels for Thin-Plate Splines 136
4.3.3 Tensor Product Splines with Thin-Plate Marginals 139
4.3.4 Case Study: Water Acidity in Lakes 140
4.4 Splines on the Sphere 143
4.4.1 Spherical Harmonics 143
4.4.2 Laplacian on the Sphere and Spherical Splines . . . 144

4.4.3 Reproducing Kernels in Closed Forms 146
4.4.4 Case Study: Global Temperature Map 147
4.5 L-Splines 149
4.5.1 Trigonometric Splines 150
4.5.2 Chebyshev Splines 153
4.5.3 General Construction 157
4.5.4 Case Study: Weight Loss of Obese Patient 161
4.5.5 Fast Algorithm 165
4.6 Bibliographic Notes 166
4.7 Problems 167

5 Regression with Responses from Exponential Families 175
5.1 Preliminaries 176
5.2 Smoothing Parameter Selection 177
5.2.1 Performance-Oriented Iteration 178
5.2.2 Direct Cross-Validation 181
5.3 Inferential Tools 184
5.3.1 Approximate Bayesian Confidence Intervals 185
5.3.2 Kullback-Leibler Projection 186
5.4 Software, Customization, and Empirical Performance . . . 187
5.4.1 R Package gss: gssanova, gssanova0, and gssanova1 Suites 187
5.4.2 Binomial Family 188
5.4.3 Poisson Family 191
5.4.4 Gamma Family 193
5.4.5 Inverse Gaussian Family 196
5.4.6 Negative Binomial Family 199
5.5 Case Studies 202
5.5.1 Eruption Time of Old Faithful 202
5.5.2 Spectrum of Yearly Sunspots 203
5.5.3 Progression of Diabetic Retinopathy 205
5.5.4 Colorectal Cancer Mortality Rate 208
5.6 Bibliographic Notes 210
5.7 Problems 212

6 Regression with Correlated Responses 215
6.1 Models for Correlated Data 216
6.1.1 Random Effects 216
6.1.2 Stationary Time Series 216
6.2 Mixed-Effect Models and Penalized Joint Likelihood 217
6.2.1 Smoothing Matrices 218
6.2.2 Bayes Model 219
6.2.3 Optimality of Generalized Cross-Validation 219
6.2.4 Empirical Performance 221

6.2.5 Non-Gaussian Regression 222
6.2.6 R Package gss: Optional Argument random 222
6.3 Penalized Likelihood with Correlated Data 223
6.3.1 Bayes Model 223
6.3.2 Extension of Cross-Validation 225
6.3.3 Optimality of Cross-Validation 226
6.3.4 Empirical Performance 228
6.3.5 R Package gss: ssanova9 Suite 230
6.4 Case Studies 231
6.4.1 Treatment of Bacteriuria 231
6.4.2 Ozone Concentration in Los Angeles Basin 232
6.5 Bibliographic Notes 233
6.6 Problems 235

7 Probability Density Estimation 237

7.1 Preliminaries 238
7.2 Poisson Intensity 242
7.3 Smoothing Parameter Selection 243
7.3.1 Kullback-Leibler Loss 243
7.3.2 Cross-Validation 244
7.3.3 Empirical Performance 246
7.4 Computation, Inference, and Software 247
7.4.1 Newton Iteration 247
7.4.2 Numerical Integration 248
7.4.3 Kullback-Leibler Projection 250
7.4.4 R Package gss: ssden Suite 250
7.5 Case Studies 253
7.5.1 Buffalo Snowfall 253
7.5.2 Eruption Time of Old Faithful 254
7.5.3 AIDS Incubation 255
7.6 Biased Sampling and Random Truncation 257
7.6.1 Biased and Truncated Samples 257
7.6.2 Penalized Likelihood Estimation 258
7.6.3 Empirical Performance 260
7.6.4 R Package gss: ssden Suite 260
7.6.5 Case Study: AIDS Incubation 262
7.7 Conditional Densities 263
7.7.1 Penalized Likelihood Estimation 263
7.7.2 Empirical Performance of Cross-validation 265
7.7.3 Kullback-Leibler Projection 266
7.7.4 R Package gss: sscden Suite 266
7.7.5 Case Study: Penny Thickness 268
7.8 Regression with Cross-Classified Responses 269
7.8.1 Logistic Regression 269
7.8.2 Log-Linear Regression Models 271

7.8.3 Bayesian Confidence Intervals for y-Contrasts 271
7.8.4 Mixed-Effect Models for Correlated Data 272
7.8.5 Empirical Performance of Cross-Validation 273
7.8.6 R Package gss: ssllrm Suite 274
7.8.7 Case Study: Eyetracking Experiments 275
7.9 Response-Based Sampling 278
7.9.1 Response-Based Samples 278
7.9.2 Penalized Likelihood Estimation 279
7.10 Bibliographic Notes 280
7.11 Problems 282

8 Hazard Rate Estimation 285
8.1 Preliminaries 286
8.2 Smoothing Parameter Selection 288
8.2.1 Kullback-Leibler Loss and Cross-Validation . . . 289
8.2.2 Empirical Performance 291
8.3 Inference and Software 292
8.3.1 Bayesian Confidence Intervals 292
8.3.2 Kullback-Leibler Projection 293
8.3.3 Frailty Models for Correlated Data 293
8.3.4 R Package gss: sshzd Suite 293
8.4 Case Studies 295
8.4.1 Treatments of Gastric Cancer 295
8.4.2 Survival After Heart Transplant 297
8.5 Penalized Partial Likelihood 299
8.5.1 Partial Likelihood and Biased Sampling 299
8.5.2 Inference 300
8.5.3 R Package gss: sscox Suite 300
8.5.4 Case Study: Survival After Heart Transplant . . . 302
8.6 Models Parametric in Time 303
8.6.1 Location-Scale Families and Accelerated Life Models 303
8.6.2 Kullback-Leibler and Cross-Validation 305
8.6.3 Weibull Family 305
8.6.4 Log Normal Family 309
8.6.5 Log Logistic Family 311
8.6.6 Case Study: Survival After Heart Transplant . . . 314
8.7 Bibliographic Notes 316
8.8 Problems 317

9 Asymptotic Convergence 319
9.1 Preliminaries 319
9.2 Rates for Density Estimates 322
9.2.1 Linear Approximation 323
9.2.2 Approximation Error and Main Results 325

9.2.3 Efficient Approximation 327
9.2.4 Convergence Under Incorrect Model 330
9.2.5 Estimation Under Biased Sampling 331
9.2.6 Estimation of Conditional Density 332
9.2.7 Estimation Under Response-Based Sampling 332
9.3 Rates for Hazard Estimates 333
9.3.1 Martingale Structure 333
9.3.2 Linear Approximation 334
9.3.3 Approximation Error and Main Results 335
9.3.4 Efficient Approximation 338
9.3.5 Convergence Under Incorrect Model 341
9.4 Rates for Regression Estimates 341
9.4.1 General Formulation 341
9.4.2 Linear Approximation 342
9.4.3 Approximation Error and Main Result 343
9.4.4 Efficient Approximation 345
9.4.5 Convergence Under Incorrect Model 347
9.5 Bibliographic Notes 348
9.6 Problems 349

10 Penalized Pseudo Likelihood 351

10.1 Density Estimation on Product Domains 352
10.1.1 Pseudo and Genuine Likelihoods 352
10.1.2 Preliminaries 353
10.1.3 Smoothing Parameter Selection 354
10.1.4 Square Error Projection 357
10.1.5 R Package gss: ssden1 Suite 358
10.1.6 Case Study: Transcription Factor Association 359
10.2 Density Estimation: Asymptotic Convergence 360
10.2.1 Linear Approximation 361
10.2.2 Approximation Error and Main Results 361
10.2.3 Efficient Approximation 363
10.3 Conditional Density Estimation 364
10.3.1 Preliminaries 365
10.3.2 Smoothing Parameter Selection 366
10.3.3 Square Error Projection 369
10.3.4 R Package gss: sscden1 Suite 369
10.3.5 Case Study: Penny Thickness 371
10.3.6 Asymptotic Convergence 372
10.4 Hazard Estimation 372
10.4.1 Preliminaries 373
10.4.2 Smoothing Parameter Selection 374
10.4.3 Inference 375
10.4.4 R Package gss: sshzd1 Suite 376
10.4.5 Case Study: Survival After Heart Transplant 378

10.5 Hazard Estimation: Asymptotic Convergence 378
10.5.1 Linear Approximation 379
10.5.2 Approximation Error and Main Results 380
10.5.3 Efficient Approximation 381
10.6 Bibliographic Notes 383
10.7 Problems 384

A R Package gss 387
A.1 Model Construction 387
A.1.1 Marginal Configurations 388
A.1.2 Construction of Interaction Terms 389
A.1.3 Custom Types 390
A.2 Modeling and Data Analytical Tools 391
A.3 Numerical Engines 393

B Conceptual Critiques 395
B.1 Model Indexing 395
B.2 Optimal and Cross-Validation Indices 397
B.3 Loss, Risk, and Smoothing Parameter Selection 398
B.4 Degrees of Freedom 400

References 403

Author Index 417

Subject Index 421