

Contents

Preface	xvii
I Nonparametric Kernel Methods	1
1 Density Estimation	3
1.1 Univariate Density Estimation	4
1.2 Univariate Bandwidth Selection: Rule-of-Thumb and Plug-In Methods	14
1.3 Univariate Bandwidth Selection: Cross-Validation Methods	15
1.3.1 Least Squares Cross-Validation	15
1.3.2 Likelihood Cross-Validation	18
1.3.3 An Illustration of Data-Driven Bandwidth Selection	19
1.4 Univariate CDF Estimation	19
1.5 Univariate CDF Bandwidth Selection: Cross-Validation Methods	23
1.6 Multivariate Density Estimation	24
1.7 Multivariate Bandwidth Selection: Rule-of-Thumb and Plug-In Methods	26
1.8 Multivariate Bandwidth Selection: Cross-Validation Methods	27
1.8.1 Least Squares Cross-Validation	27
1.8.2 Likelihood Cross-Validation	28
1.9 Asymptotic Normality of Density Estimators	28
1.10 Uniform Rates of Convergence	30
1.11 Higher Order Kernel Functions	33
1.12 Proof of Theorem 1.4 (Uniform Almost Sure Convergence)	35
1.13 Applications	40

1.13.1	Female Wage Inequality	41
1.13.2	Unemployment Rates and City Size	43
1.13.3	Adolescent Growth	44
1.13.4	Old Faithful Geyser Data	44
1.13.5	Evolution of Real Income Distribution in Italy, 1951–1998	45
1.14	Exercises	47
2	Regression	57
2.1	Local Constant Kernel Estimation	60
2.1.1	Intuition Underlying the Local Constant Kernel Estimator	64
2.2	Local Constant Bandwidth Selection	66
2.2.1	Rule-of-Thumb and Plug-In Methods	66
2.2.2	Least Squares Cross-Validation	69
2.2.3	AIC_c	72
2.2.4	The Presence of Irrelevant Regressors	73
2.2.5	Some Further Results on Cross-Validation	78
2.3	Uniform Rates of Convergence	78
2.4	Local Linear Kernel Estimation	79
2.4.1	Local Linear Bandwidth Selection: Least Squares Cross-Validation	83
2.5	Local Polynomial Regression (General p th Order)	85
2.5.1	The Univariate Case	85
2.5.2	The Multivariate Case	88
2.5.3	Asymptotic Normality of Local Polynomial Estimators	89
2.6	Applications	92
2.6.1	Prestige Data	92
2.6.2	Adolescent Growth	92
2.6.3	Inflation Forecasting and Money Growth	93
2.7	Proofs	97
2.7.1	Derivation of (2.24)	98
2.7.2	Proof of Theorem 2.7	100
2.7.3	Definitions of $A_{l,p+1}$ and V_l Used in Theorem 2.10	106
2.8	Exercises	108
3	Frequency Estimation with Mixed Data	115
3.1	Probability Function Estimation with Discrete Data	116

3.2	Regression with Discrete Regressors	118
3.3	Estimation with Mixed Data: The Frequency Approach	118
3.3.1	Density Estimation with Mixed Data	118
3.3.2	Regression with Mixed Data	119
3.4	Some Cautionary Remarks on Frequency Methods	120
3.5	Proofs	122
3.5.1	Proof of Theorem 3.1	122
3.6	Exercises	123
4	Kernel Estimation with Mixed Data	125
4.1	Smooth Estimation of Joint Distributions with Discrete Data	126
4.2	Smooth Regression with Discrete Data	131
4.3	Kernel Regression with Discrete Regressors: The Irrelevant Regressor Case	134
4.4	Regression with Mixed Data: Relevant Regressors	136
4.4.1	Smooth Estimation with Mixed Data	136
4.4.2	The Cross-Validation Method	138
4.5	Regression with Mixed Data: Irrelevant Regressors	140
4.5.1	Ordered Discrete Variables	144
4.6	Applications	145
4.6.1	Food-Away-from-Home Expenditure	145
4.6.2	Modeling Strike Volume	147
4.7	Exercises	150
5	Conditional Density Estimation	155
5.1	Conditional Density Estimation: Relevant Variables	155
5.2	Conditional Density Bandwidth Selection	157
5.2.1	Least Squares Cross-Validation: Relevant Variables	157
5.2.2	Maximum Likelihood Cross-Validation: Relevant Variables	160
5.3	Conditional Density Estimation: Irrelevant Variables	162
5.4	The Multivariate Dependent Variables Case	164
5.4.1	The General Categorical Data Case	167
5.4.2	Proof of Theorem 5.5	168
5.5	Applications	171
5.5.1	A Nonparametric Analysis of Corruption	171
5.5.2	Extramarital Affairs Data	172
5.5.3	Married Female Labor Force Participation	175

5.5.4	Labor Productivity	177
5.5.5	Multivariate Y Conditional Density Example: GDP Growth and Population Growth Conditional on OECD Status	178
5.6	Exercises	180
6	Conditional CDF and Quantile Estimation	181
6.1	Estimating a Conditional CDF with Continuous Covariates without Smoothing the Dependent Variable	182
6.2	Estimating a Conditional CDF with Continuous Covariates Smoothing the Dependent Variable	184
6.3	Nonparametric Estimation of Conditional Quantile Functions	189
6.4	The Check Function Approach	191
6.5	Conditional CDF and Quantile Estimation with Mixed Discrete and Continuous Covariates	193
6.6	A Small Monte Carlo Simulation Study	196
6.7	Nonparametric Estimation of Hazard Functions	198
6.8	Applications	200
6.8.1	Boston Housing Data	200
6.8.2	Adolescent Growth Charts	202
6.8.3	Conditional Value at Risk	202
6.8.4	Real Income in Italy, 1951–1998	206
6.8.5	Multivariate Y Conditional CDF Example: GDP Growth and Population Growth Conditional on OECD Status	206
6.9	Proofs	209
6.9.1	Proofs of Theorems 6.1, 6.2, and 6.4	209
6.9.2	Proofs of Theorems 6.5 and 6.6 (Mixed Covariates Case)	214
6.10	Exercises	215
II	Semiparametric Methods	219
7	Semiparametric Partially Linear Models	221
7.1	Partially Linear Models	222
7.1.1	Identification of β	222
7.2	Robinson's Estimator	222
7.2.1	Estimation of the Nonparametric Component	228

7.3	Andrews's MINPIN Method	230
7.4	Semiparametric Efficiency Bounds	233
7.4.1	The Conditionally Homoskedastic Error Case	233
7.4.2	The Conditionally Heteroskedastic Error Case	235
7.5	Proofs	238
7.5.1	Proof of Theorem 7.2	238
7.5.2	Verifying Theorem 7.3 for a Partially Linear Model	244
7.6	Exercises	246
8	Semiparametric Single Index Models	249
8.1	Identification Conditions	251
8.2	Estimation	253
8.2.1	Ichimura's Method	253
8.3	Direct Semiparametric Estimators for β	258
8.3.1	Average Derivative Estimators	258
8.3.2	Estimation of $g(\cdot)$	262
8.4	Bandwidth Selection	263
8.4.1	Bandwidth Selection for Ichimura's Method	263
8.4.2	Bandwidth Selection with Direct Estimation Methods	265
8.5	Klein and Spady's Estimator	266
8.6	Lewbel's Estimator	267
8.7	Manski's Maximum Score Estimator	269
8.8	Horowitz's Smoothed Maximum Score Estimator	270
8.9	Han's Maximum Rank Estimator	270
8.10	Multinomial Discrete Choice Models	271
8.11	Ai's Semiparametric Maximum Likelihood Approach	272
8.12	A Sketch of the Proof of Theorem 8.1	275
8.13	Applications	277
8.13.1	Modeling Response to Direct Marketing Catalog Mailings	277
8.14	Exercises	281
9	Additive and Smooth (Varying) Coefficient Semiparametric Models	283
9.1	An Additive Model	283
9.1.1	The Marginal Integration Method	284
9.1.2	A Computationally Efficient Oracle Estimator	286
9.1.3	The Ordinary Backfitting Method	289

9.1.4	The Smoothed Backfitting Method	290
9.1.5	Additive Models with Link Functions	295
9.2	An Additive Partially Linear Model	297
9.2.1	A Simple Two-Step Method	299
9.3	A Semiparametric Varying (Smooth) Coefficient Model	301
9.3.1	A Local Constant Estimator of the Smooth Coefficient Function	302
9.3.2	A Local Linear Estimator of the Smooth Coefficient Function	303
9.3.3	Testing for a Parametric Smooth Coefficient Model	306
9.3.4	Partially Linear Smooth Coefficient Models	308
9.3.5	Proof of Theorem 9.3	310
9.4	Exercises	312
10	Selectivity Models	315
10.1	Semiparametric Type-2 Tobit Models	316
10.2	Estimation of a Semiparametric Type-2 Tobit Model	317
10.2.1	Gallant and Nychka's Estimator	318
10.2.2	Estimation of the Intercept in Selection Models	319
10.3	Semiparametric Type-3 Tobit Models	320
10.3.1	Econometric Preliminaries	320
10.3.2	Alternative Estimation Methods	323
10.4	Das, Newey and Vella's Nonparametric Selection Model	328
10.5	Exercises	330
11	Censored Models	331
11.1	Parametric Censored Models	332
11.2	Semiparametric Censored Regression Models	334
11.3	Semiparametric Censored Regression Models with Nonparametric Heteroskedasticity	336
11.4	The Univariate Kaplan-Meier CDF Estimator	338
11.5	The Multivariate Kaplan-Meier CDF Estimator	341
11.5.1	Nonparametric Regression Models with Random Censoring	343
11.6	Nonparametric Censored Regression	345
11.6.1	Lewbel and Linton's Approach	345
11.6.2	Chen, Dahl and Khan's Approach	346

11.7 Exercises	348
III Consistent Model Specification Tests	349
12 Model Specification Tests	351
12.1 A Simple Consistent Test for Parametric Regression Functional Form	354
12.1.1 A Consistent Test for Correct Parametric Functional Form	355
12.1.2 Mixed Data	360
12.2 Testing for Equality of PDFs	362
12.3 More Tests Related to Regression Functions	365
12.3.1 Härdle and Mammen's Test for a Parametric Regression Model	365
12.3.2 An Adaptive and Rate Optimal Test	367
12.3.3 A Test for a Parametric Single Index Model	369
12.3.4 A Nonparametric Omitted Variables Test	370
12.3.5 Testing the Significance of Categorical Variables	375
12.4 Tests Related to PDFs	378
12.4.1 Testing Independence between Two Random Variables	378
12.4.2 A Test for a Parametric PDF	380
12.4.3 A Kernel Test for Conditional Parametric Distributions	382
12.5 Applications	385
12.5.1 Growth Convergence Clubs	385
12.6 Proofs	388
12.6.1 Proof of Theorem 12.1	388
12.6.2 Proof of Theorem 12.2	389
12.6.3 Proof of Theorem 12.5	389
12.6.4 Proof of Theorem 12.9	391
12.7 Exercises	394
13 Nonsmoothing Tests	397
13.1 Testing for Parametric Regression Functional Form	398
13.2 Testing for Equality of PDFs	401
13.3 A Nonparametric Significance Test	401
13.4 Andrews's Test for Conditional CDFs	402
13.5 Hong's Tests for Serial Dependence	404

13.6	More on Nonsmoothing Tests	408
13.7	Proofs	409
13.7.1	Proof of Theorem 13.1	409
13.8	Exercises	410
 IV Nonparametric Nearest Neighbor and Series Methods		 413
 14 <i>K</i>-Nearest Neighbor Methods		 415
14.1	Density Estimation: The Univariate Case	415
14.2	Regression Function Estimation	419
14.3	A Local Linear k -nn Estimator	421
14.4	Cross-Validation with Local Constant k -nn Estimation	422
14.5	Cross-Validation with Local Linear k -nn Estimation	425
14.6	Estimation of Semiparametric Models with k -nn Methods	427
14.7	Model Specification Tests with k -nn Methods	428
14.7.1	A Bootstrap Test	431
14.8	Using Different k for Different Components of x	432
14.9	Proofs	432
14.9.1	Proof of Theorem 14.1	435
14.9.2	Proof of Theorem 14.5	435
14.9.3	Proof of Theorem 14.10	440
14.10	Exercises	444
 15 Nonparametric Series Methods		 445
15.1	Estimating Regression Functions	446
15.1.1	Convergence Rates	449
15.2	Selection of the Series Term K	451
15.2.1	Asymptotic Normality	453
15.3	A Partially Linear Model	454
15.3.1	An Additive Partially Linear Model	455
15.3.2	Selection of Nonlinear Additive Components	461
15.3.3	Estimating an Additive Model with a Known Link Function	463
15.4	Estimation of Partially Linear Varying Coefficient Models	466
15.4.1	Testing for Correct Parametric Regression Functional Form	471

15.4.2 A Consistent Test for an Additive Partially Linear Model	474
15.5 Other Series-Based Tests	479
15.6 Proofs	480
15.6.1 Proof of Theorem 15.1	480
15.6.2 Proof of Theorem 15.3	484
15.6.3 Proof of Theorem 15.6	488
15.6.4 Proof of Theorem 15.9	492
15.6.5 Proof of Theorem 15.10	497
15.7 Exercises	502
V Time Series, Simultaneous Equation, and Panel Data Models	503
16 Instrumental Variables and Efficient Estimation of Semiparametric Models	505
16.1 A Partially Linear Model with Endogenous Regressors in the Parametric Part	505
16.2 A Varying Coefficient Model with Endogenous Regressors in the Parametric Part	509
16.3 Ai and Chen's Efficient Estimator with Conditional Moment Restrictions	511
16.3.1 Estimation Procedures	511
16.3.2 Asymptotic Normality for $\hat{\theta}$	513
16.3.3 A Partially Linear Model with the Endogenous Regressors in the Nonparametric Part	515
16.4 Proof of Equation (16.16)	517
16.5 Exercises	520
17 Endogeneity in Nonparametric Regression Models	521
17.1 A Nonparametric Model	521
17.2 A Triangular Simultaneous Equation Model	522
17.3 Newey-Powell Series-Based Estimator	527
17.4 Hall and Horowitz's Kernel-Based Estimator	529
17.5 Darolles, Florens and Renault's Estimator	532
17.6 Exercises	533
18 Weakly Dependent Data	535
18.1 Density Estimation with Dependent Data	537

18.1.1	Uniform Almost Sure Rate of Convergence	541
18.2	Regression Models with Dependent Data	541
18.2.1	The Martingale Difference Error Case	541
18.2.2	The Autocorrelated Error Case	544
18.2.3	One-Step-Ahead Forecasting	546
18.2.4	d -Step-Ahead Forecasting	547
18.2.5	Estimation of Nonparametric Impulse Response Functions	548
18.3	Semiparametric Models with Dependent Data	551
18.3.1	A Partially Linear Model with Dependent Data	551
18.3.2	Additive Regression Models	552
18.3.3	Varying Coefficient Models with Dependent Data	553
18.4	Testing for Serial Correlation in Semiparametric Models	554
18.4.1	The Test Statistic and Its Asymptotic Distribution	554
18.4.2	Testing Zero First Order Serial Correlation	555
18.5	Model Specification Tests with Dependent Data	556
18.5.1	A Kernel Test for Correct Parametric Regression Functional Form	556
18.5.2	Nonparametric Significance Tests	557
18.6	Nonsmoothing Tests for Regression Functional Form	558
18.7	Testing Parametric Predictive Models	559
18.7.1	In-Sample Testing of Conditional CDFs	559
18.7.2	Out-of-Sample Testing of Conditional CDFs	562
18.8	Applications	564
18.8.1	Forecasting Short-Term Interest Rates	564
18.9	Nonparametric Estimation with Nonstationary Data	566
18.10	Proofs	567
18.10.1	Proof of Equation (18.9)	567
18.10.2	Proof of Theorem 18.2	569
18.11	Exercises	572
19	Panel Data Models	575
19.1	Nonparametric Estimation of Panel Data Models: Ignoring the Variance Structure	576
19.2	Wang's Efficient Nonparametric Panel Data Estimator	578
19.3	A Partially Linear Model with Random Effects	584
19.4	Nonparametric Panel Data Models with Fixed Effects	586

19.4.1	Error Variance Structure Is Known	587
19.4.2	The Error Variance Structure Is Unknown	590
19.5	A Partially Linear Model with Fixed Effects	592
19.6	Semiparametric Instrumental Variable Estimators	594
19.6.1	An Infeasible Estimator	594
19.6.2	The Choice of Instruments	595
19.6.3	A Feasible Estimator	597
19.7	Testing for Serial Correlation and for Individual Effects in Semiparametric Models	599
19.8	Series Estimation of Panel Data Models	602
19.8.1	Additive Effects	602
19.8.2	Alternative Formulation of Fixed Effects	604
19.9	Nonlinear Panel Data Models	606
19.9.1	Censored Panel Data Models	607
19.9.2	Discrete Choice Panel Data Models	614
19.10	Proofs	618
19.10.1	Proof of Theorem 19.1	618
19.10.2	Leading MSE Calculation of Wang's Estimator	621
19.11	Exercises	624
20	Topics in Applied Nonparametric Estimation	627
20.1	Nonparametric Methods in Continuous-Time Models	627
20.1.1	Nonparametric Estimation of Continuous-Time Models	627
20.1.2	Nonparametric Tests for Continuous-Time Models	632
20.1.3	Ait-Sahalia's Test	632
20.1.4	Hong and Li's Test	633
20.1.5	Proofs	636
20.2	Nonparametric Estimation of Average Treatment Effects	639
20.2.1	The Model	640
20.2.2	An Application: Assessing the Efficacy of Right Heart Catheterization	642
20.3	Nonparametric Estimation of Auction Models	645
20.3.1	Estimation of First Price Auction Models	645
20.3.2	Conditionally Independent Private Information Auctions	648
20.4	Copula-Based Semiparametric Estimation of Multivariate Distributions	651

20.4.1	Some Background on Copula Functions	651
20.4.2	Semiparametric Copula-Based Multivariate Distributions	652
20.4.3	A Two-Step Estimation Procedure	653
20.4.4	A One-Step Efficient Estimation Procedure	655
20.4.5	Testing Parametric Functional Forms of a Copula	657
20.5	A Semiparametric Transformation Model	659
20.6	Exercises	662
A	Background Statistical Concepts	663
1.1	Probability, Measure, and Measurable Space	663
1.2	Metric, Norm, and Functional Spaces	672
1.3	Limits and Modes of Convergence	680
1.3.1	Limit Supremum and Limit Infimum	680
1.3.2	Modes of Convergence	681
1.4	Inequalities, Laws of Large Numbers, and Central Limit Theorems	688
1.5	Exercises	694
	Bibliography	697
	Author Index	737
	Subject Index	744

Chapter 1

Density Estimation

The estimation of probability density functions (PDFs) and cumulative distribution functions (CDFs) are cornerstones of applied data analysis in the social sciences. Testing for the equality of two distributions (or moments thereof) is perhaps the most basic test in all of applied data analysis. Economists, for instance, devote a great deal of attention to the study of income distributions and how they vary across regions and over time. Though the PDF and CDF are often the objects of direct interest, their estimation also serves as an important building block for other objects being modeled such as a conditional mean (i.e., a “regression function”), which may be directly modeled using nonparametric or semiparametric methods (a conditional mean is a function of a conditional PDF, which is itself a ratio of unconditional PDFs). After mastering the principles underlying the nonparametric estimation of a PDF, the nonparametric estimation of the workhorse of applied data analysis, the conditional mean function considered in Chapter 2, progresses in a fairly straightforward manner. Careful study of the approaches developed in Chapter 1 will be most helpful for understanding material presented in later chapters.

We begin with the estimation of a univariate PDF in Sections 1.1 through 1.3, turn to the estimation of a univariate CDF in Sections 1.4 and 1.5, and then move on to the more general multivariate setting in Sections 1.6 through 1.8. Asymptotic normality, uniform rates of convergence, and bias reduction methods appear in Sections 1.9 through 1.12. Numerous illustrative applications appear in Section 1.13, while theoretical and applied exercises can be found in Section 1.14

We now proceed with a discussion of how to estimate the PDF

$f_X(x)$ of a random variable X . For notational simplicity we drop the subscript X and simply use $f(x)$ to denote the PDF of X . Some of the treatments of the kernel estimation of a PDF discussed in this chapter are drawn from the two excellent monographs by Silverman (1986) and Scott (1992).

1.1 Univariate Density Estimation

To best appreciate why one might consider using nonparametric methods to estimate a PDF, we begin with an illustrative example, the parametric estimation of a PDF.

Example 1.1. *Suppose X_1, X_2, \dots, X_n represent independent and identically distributed (i.i.d.) draws from a normal distribution with mean μ and variance σ^2 . We wish to estimate the normal PDF $f(x)$.*

By assumption, $f(x)$ has a known parametric functional form (i.e., univariate normal) given by $f(x) = (2\pi\sigma^2)^{-1/2} \exp[-\frac{1}{2}(x - \mu)^2/\sigma^2]$, where the mean $\mu = E(X)$ and variance $\sigma^2 = E[(X - E(X))^2] = \text{var}(X)$ are the only unknown parameters to be estimated. One could estimate μ and σ^2 by the method of maximum likelihood as follows. Under the i.i.d. assumption, the joint PDF of (X_1, \dots, X_n) is simply the product of the univariate PDFs, which may be written as

$$f(X_1, \dots, X_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i - \mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2}.$$

Conditional upon the observed sample and taking the logarithm, this gives us the log-likelihood function

$$\begin{aligned} \mathcal{L}(\mu, \sigma^2) &\equiv \ln f(X_1, \dots, X_n; \mu, \sigma^2) \\ &= -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln \sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2. \end{aligned}$$

The method of maximum likelihood proceeds by choosing those parameters that make it most likely that we observed the sample at hand given our distributional assumption. Thus, the likelihood function (or a monotonic transformation thereof, e.g., \ln) expresses the plausibility of different values of μ and σ^2 given the observed sample. We then maximize the likelihood function with respect to these two unknown parameters.

The necessary first order conditions for a maximization of the log-likelihood function are $\partial\mathcal{L}(\mu, \sigma^2)/\partial\mu = 0$ and $\partial\mathcal{L}(\mu, \sigma^2)/\partial\sigma^2 = 0$. Solving these first order conditions for the two unknown parameters μ and σ^2 yields

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \hat{\mu})^2.$$

$\hat{\mu}$ and $\hat{\sigma}^2$ above are the maximum likelihood estimators of μ and σ^2 , respectively, and the resulting estimator of $f(x)$ is

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp \left[-\frac{1}{2} \left(\frac{x - \hat{\mu}}{\hat{\sigma}} \right)^2 \right].$$

The “Achilles heel” of any parametric approach is of course the requirement that, prior to estimation, the analyst must specify the exact parametric functional form for the object being estimated. Upon reflection, the parametric approach is somewhat circular since we initially set out to estimate an unknown density but must first assume that the density is in fact known (up to a handful of unknown parameters, of course). Having based our estimate on the assumption that the density is a member of a known parametric family, we must then naturally confront the possibility that the parametric model is “mis-specified,” i.e., not consistent with the population from which the data was drawn. For instance, by assuming that X is drawn from a normally distributed population in the above example, we in fact impose a number of potentially quite restrictive assumptions: symmetry, unimodality, monotonically decreasing away from the mode and so on. If the true density were in fact asymmetric or possessed multiple modes, or was nonmonotonic away from the mode, then the presumption of distributional normality may provide a misleading characterization of the true density and could thereby produce erroneous estimates and lead to unsound inference.

At this juncture many readers will no doubt be pointing out that, having estimated a parametric PDF, one can always test whether the underlying distributional assumption is valid. We are, of course, completely sympathetic toward such arguments. Often, however, the rejection of a distributional assumption fails to provide any clear alternative. That is, we can reject the assumption of normality, but this rejection leaves us where we started, perhaps having ruled out but one of a large

number of candidate distributions. Against this backdrop, researchers might instead consider nonparametric approaches.

Nonparametric methods circumvent problems arising from the need to specify parametric functional forms prior to estimation. Rather than presume one knows the exact functional form of the object being estimated, one instead presumes that it satisfies some regularity conditions such as smoothness and differentiability. This does not, however, come without cost. By imposing less structure on the functional form of the PDF than do parametric methods, nonparametric methods require more data to achieve the same degree of precision as a *correctly specified* parametric model. Our primary focus in this text is on a class of estimators known as “nonparametric kernel estimators” (a “kernel function” is simply a weighting function), though in Chapters 14 and 15 we provide a treatment of alternative nonparametric methodologies including nearest neighbor and series methods.

Before proceeding to a formal theoretical analysis of nonparametric density estimation methods, we first consider a popular example of estimating the probability of a head on a toss of a coin which is closely related to the nonparametric estimation of a CDF. This in turn will lead us to the nonparametric estimation of a PDF.

Example 1.2. *Suppose we have a coin (perhaps an unfair one) and we want to estimate the probability of flipping the coin and having it land heads up. Let $p = P(H)$ denote the (unknown) population probability of obtaining a head. Taking a relative frequency approach, we would flip the coin n times, count the frequency of heads in n trials, and compute the relative frequency given by*

$$\hat{p} = \frac{1}{n} \{ \# \text{ of heads } \}, \quad (1.1)$$

which provides an estimate of p . The \hat{p} defined in (1.1) is often referred to as a “frequency estimator” of p , and it is also the maximum likelihood estimator of p (see Exercise 1.2). The estimator \hat{p} is, of course, fully nonparametric. Intuitively, one would expect that, if n is large, then \hat{p} should be “close” to p . Indeed, one can easily show that the mean squared error (MSE) of \hat{p} is given by (see Exercise 1.3)

$$\text{MSE}(\hat{p}) \stackrel{\text{def}}{=} \text{E} \left[(\hat{p} - p)^2 \right] = \frac{p(1-p)}{n},$$

so $\text{MSE}(\hat{p}) \rightarrow 0$ as $n \rightarrow \infty$, which is termed as \hat{p} converges to p in mean square error; see Appendix A for the definitions of various modes of convergence.

We now discuss how to obtain an estimator of the CDF of X , which we denote by $F(x)$. The CDF is defined as

$$F(x) = P[X \leq x].$$

With i.i.d. data X_1, \dots, X_n (i.e., random draws from the distribution $F(\cdot)$), one can estimate $F(x)$ by

$$F_n(x) = \frac{1}{n} \{ \# \text{ of } X_i\text{'s } \leq x \}. \quad (1.2)$$

Equation (1.2) has a nice intuitive interpretation. Going back to our coin-flip example, if a coin is such that the probability of obtaining a head when we flip it equals $F(x)$ ($F(x)$ is unknown), and if we treat the collection of data X_1, \dots, X_n as flipping a coin n times and we say that a head occurs on the i^{th} trial if $X_i \leq x$, then $P(H) = P(X_i \leq x) = F(x)$. The familiar frequency estimator of $P(H)$ is equal to the number of heads divided by the number of trials:

$$\hat{P}(H) = \frac{\# \text{ of heads}}{n} = \frac{1}{n} \{ \# \text{ of } X_i\text{'s } \leq x \} \equiv F_n(x). \quad (1.3)$$

Therefore, we call (1.2) a frequency estimator of $F(x)$. Just as before when estimating $P(H)$, we expect intuitively that as n gets large, $\hat{P}(H)$ should yield a more accurate estimate of $P(H)$. By the same reasoning, one would expect that as $n \rightarrow \infty$, $F_n(x)$ yields a more accurate estimate of $F(x)$. Indeed, one can easily show that $F_n(x) \rightarrow F(x)$ in MSE, which implies that $F_n(x)$ converges to $F(x)$ in probability and also in distribution as $n \rightarrow \infty$. In Appendix A we introduce the concepts of convergence in mean square error, convergence in probability, convergence in distribution, and almost sure convergence. It is well established that $F_n(x)$ indeed converges to $F(x)$ in each of these various senses. These concepts of convergence are necessary as it is easy to show that the ordinary limit of $F_n(x)$ does not exist, i.e., $\lim_{n \rightarrow \infty} F_n(x)$ does not exist (see Exercise 1.3, where the definition of an ordinary limit is provided). This example highlights the necessity of introducing new concepts of convergence modes such as convergence in mean square error and convergence in probability.

Now we take up the question of how to estimate a PDF $f(x)$ without making parametric presumptions about its functional form. From the

definition of $f(x)$ we have¹

$$f(x) = \frac{d}{dx}F(x). \quad (1.4)$$

From (1.2) and (1.4), an obvious estimator of $f(x)$ is²

$$\hat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}, \quad (1.5)$$

where h is a small positive increment.

By substituting (1.2) into (1.5), we obtain

$$\hat{f}(x) = \frac{1}{2nh} \{ \# \text{ of } X_1, \dots, X_n \text{ falling in the interval } [x-h, x+h] \}. \quad (1.6)$$

If we define a uniform kernel function given by

$$k(z) = \begin{cases} 1/2 & \text{if } |z| \leq 1 \\ 0 & \text{otherwise,} \end{cases} \quad (1.7)$$

then it is easy to see that $\hat{f}(x)$ given by (1.5) can also be expressed as

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n k\left(\frac{X_i - x}{h}\right). \quad (1.8)$$

Equation (1.8) is called a uniform kernel estimator because the kernel function $k(\cdot)$ defined in (1.7) corresponds to a uniform PDF. In general, we refer to $k(\cdot)$ as a kernel function and to h as a smoothing parameter (or, alternatively, a bandwidth or window width). Equation (1.8) is sometimes referred to as a “naïve” kernel estimator.

In fact one might use many other possible choices for the kernel function $k(\cdot)$ in this context. For example, one could use a standard normal kernel given by

$$k(v) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}v^2}, \quad -\infty < v < \infty. \quad (1.9)$$

This class of estimators can be found in the first published paper on kernel density estimation by Rosenblatt (1956), while Parzen (1962) established a number of properties associated with this class of estimators

¹We only consider the continuous X case in this chapter. We deal with the discrete X case in Chapters 3 and 4.

²Recall that the definition of the derivative of a function $g(x)$ is given by $dg(x)/dx = \lim_{h \rightarrow 0} \frac{g(x+h) - g(x)}{h}$, or, equivalently, $dg(x)/dx = \lim_{h \rightarrow 0} \frac{g(x+h) - g(x-h)}{2h}$.

and relaxed the nonnegativity assumption in order to obtain estimators which are more efficient. For this reason, this approach is sometimes referred to as “Rosenblatt-Parzen kernel density estimation.”

We will prove shortly that the kernel estimator $\hat{f}(x)$ defined in (1.8) constructed from any general nonnegative bounded kernel function $k(\cdot)$ that satisfies

$$\begin{aligned} (i) \quad & \int k(v) dv = 1 \\ (ii) \quad & k(v) = k(-v) \\ (iii) \quad & \int v^2 k(v) dv = \kappa_2 > 0 \end{aligned} \tag{1.10}$$

is a consistent estimator of $f(x)$. Note that the symmetry condition (ii) implies that $\int vk(v) dv = 0$. By consistency, we mean that $\hat{f}(x) \rightarrow f(x)$ in probability (convergence in probability is defined in Appendix A). Note that $k(\cdot)$ defined in (1.10) is a (symmetric) PDF. For recent work on kernel methods with asymmetric kernels, see Abadir and Lawford (2004).

To define various modes of convergence, we first introduce the concept of the “Euclidean norm” (“Euclidean length”) of a vector. Given a $q \times 1$ vector $x = (x_1, x_2, \dots, x_q)' \in \mathbb{R}^q$, we use $\|x\|$ to denote the Euclidean length of x , which is defined by

$$\|x\| = [x'x]^{1/2} \equiv \sqrt{x_1^2 + x_2^2 + \dots + x_q^2}.$$

When $q = 1$ (a scalar), $\|x\|$ is simply the absolute value of x .

In the appendix we discuss the notation $O(\cdot)$ (“big Oh”) and $o(\cdot)$ (“small Oh”). Let a_n be a nonstochastic sequence. We say that $a_n = O(n^\alpha)$ if $|a_n| \leq Cn^\alpha$ for all n sufficiently large, where α and $C (> 0)$ are constants. Similarly, we say that $a_n = o(n^\alpha)$ if $a_n/n^\alpha \rightarrow 0$ as $n \rightarrow \infty$. We are now ready to prove the MSE consistency of $\hat{f}(x)$.

Theorem 1.1. *Let X_1, \dots, X_n denote i.i.d. observations having a three-times differentiable PDF $f(x)$, and let $f^{(s)}(x)$ denote the s th order derivative of $f(x)$ ($s = 1, 2, 3$). Let x be an interior point in the support of X , and let $\hat{f}(x)$ be that defined in (1.8). Assume that the kernel function $k(\cdot)$ is bounded and satisfies (1.10). Also, as $n \rightarrow \infty$, $h \rightarrow 0$ and $nh \rightarrow \infty$, then*

$$\begin{aligned} \text{MSE} \left(\hat{f}(x) \right) &= \frac{h^4}{4} \left[\kappa_2 f^{(2)}(x) \right]^2 + \frac{\kappa f(x)}{nh} + o \left(h^4 + (nh)^{-1} \right) \\ &= O \left(h^4 + (nh)^{-1} \right), \end{aligned} \tag{1.11}$$

where $\kappa_2 = \int v^2 k(v) dv$ and $\kappa = \int k^2(v) dv$.

Proof of Theorem 1.1.

$$\begin{aligned} \text{MSE}(\hat{f}(x)) &\equiv \text{E} \left\{ \left[\hat{f}(x) - f(x) \right]^2 \right\} \\ &= \text{var}(\hat{f}(x)) + \left[\text{E}(\hat{f}(x)) - f(x) \right]^2 \\ &\equiv \text{var}(\hat{f}(x)) + \left[\text{bias}(\hat{f}(x)) \right]^2. \end{aligned}$$

We will evaluate the $\text{bias}(\hat{f}(x))$ and $\text{var}(\hat{f}(x))$ terms separately.

For the bias calculation we will need to use the Taylor expansion formula. For a univariate function $g(x)$ that is m times differentiable, we have

$$\begin{aligned} g(x) &= g(x_0) + g^{(1)}(x_0)(x - x_0) + \frac{1}{2!}g^{(2)}(x_0)(x - x_0)^2 + \\ &\quad \dots + \frac{1}{(m-1)!}g^{(m-1)}(x_0)(x - x_0)^{m-1} + \frac{1}{m!}g^{(m)}(\xi)(x - x_0)^m, \end{aligned}$$

where $g^{(s)}(x_0) = \left. \frac{\partial^s g(x)}{\partial x^s} \right|_{x=x_0}$, and ξ lies between x and x_0 .

The bias term is given by

$$\begin{aligned}
 \text{bias}(\hat{f}(x)) &= \mathbb{E} \left\{ \frac{1}{nh} \sum_{i=1}^n k \left(\frac{X_i - x}{h} \right) \right\} - f(x) \\
 &= h^{-1} \mathbb{E} \left[k \left(\frac{X_1 - x}{h} \right) \right] - f(x) \\
 &\quad \text{(by identical distribution)} \\
 &= h^{-1} \int f(x_1) k \left(\frac{x_1 - x}{h} \right) dx_1 - f(x) \\
 &= h^{-1} \int f(x + hv) k(v) h dv - f(x) \\
 &\quad \text{(change of variable, } x_1 - x = hv) \\
 &= \int \left\{ f(x) + f^{(1)}(x)hv + \frac{1}{2}f^{(2)}(x)h^2v^2 + O(h^3) \right\} k(v) dv \\
 &\quad - f(x) \\
 &= \left\{ f(x) + 0 + \frac{h^2}{2}f^{(2)}(x) \int v^2k(v) dv + O(h^3) \right\} - f(x) \\
 &\quad \text{by (1.10)} \\
 &= \frac{h^2}{2}f^{(2)}(x) \int v^2k(v) dv + O(h^3), \tag{1.12}
 \end{aligned}$$

where the $O(h^3)$ term comes from

$$(1/3!)h^3 \left| \int f^{(3)}(\tilde{x})v^3k(v) dv \right| \leq Ch^3 \int |v^3k(v)dv| = O(h^3),$$

where C is a positive constant, and where \tilde{x} lies between x and $x + hv$.

Note that in the above derivation we assume that $f(x)$ is three-times differentiable. We can weaken this condition to $f(x)$ being twice differentiable, resulting in $O(h^3)$ becomes $o(h^2)$, see Exercise 1.5)

$$\begin{aligned}
 \text{bias}(\hat{f}(x)) &= \mathbb{E}(\hat{f}(x)) - f(x) \\
 &= \frac{h^2}{2}f^{(2)}(x) \int v^2k(v) dv + o(h^2). \tag{1.13}
 \end{aligned}$$

Next we consider the variance term. Observe that

$$\begin{aligned}
 \text{var} \left(\hat{f}(x) \right) &= \text{var} \left[\frac{1}{nh} \sum_{i=1}^n k \left(\frac{X_i - x}{h} \right) \right] \\
 &= \frac{1}{n^2 h^2} \left\{ \sum_{i=1}^n \text{var} \left[k \left(\frac{X_i - x}{h} \right) \right] + 0 \right\} \\
 &\quad \text{(by independence)} \\
 &= \frac{1}{nh^2} \text{var} \left(k \left(\frac{X_1 - x}{h} \right) \right) \\
 &\quad \text{(by identical distribution)} \\
 &= \frac{1}{nh^2} \left\{ \text{E} \left[k^2 \left(\frac{X_1 - x}{h} \right) \right] - \left[\text{E} \left(k \left(\frac{X_1 - x}{h} \right) \right) \right]^2 \right\} \\
 &= \frac{1}{nh^2} \left\{ \int f(x_1) k^2 \left(\frac{x_1 - x}{h} \right) dx_1 \right. \\
 &\quad \left. - \left[\int f(x_1) k \left(\frac{x_1 - x}{h} \right) dx_1 \right]^2 \right\} \\
 &= \frac{1}{nh^2} \left\{ h \int f(x + hv) k^2(v) dv \right. \\
 &\quad \left. - \left[h \int f(x + hv) k(v) dv \right]^2 \right\} \\
 &= \frac{1}{nh^2} \left\{ h \int [f(x) + f^{(1)}(\xi)hv] k^2(v) dv - O(h^2) \right\} \\
 &= \frac{1}{nh} \left\{ f(x) \int k^2(v) dv + O \left(h \int |v| k^2(v) dv \right) - O(h) \right\} \\
 &= \frac{1}{nh} \{ \kappa f(x) + O(h) \}, \tag{1.14}
 \end{aligned}$$

where $\kappa = \int k^2(v) dv$.

Equations (1.12) and (1.14) complete the proof of Theorem 1.1. \square

Theorem 1.1 implies that (by Theorem A.7 of Appendix A)

$$\hat{f}(x) - f(x) = O_p \left(h^2 + (nh)^{-1/2} \right) = o_p(1).$$

By choosing $h = cn^{-1/\alpha}$ for some $c > 0$ and $\alpha > 1$, the conditions required for consistent estimation of $f(x)$, $h \rightarrow 0$ and $nh \rightarrow \infty$,

are clearly satisfied. The overriding question is what values of c and α should be used in practice. As can be seen, for a given sample size n , if h is small, the resulting estimator will have a small bias but a large variance. On the other hand, if h is large, then the resulting estimator will have a small variance but a large bias. To minimize $\text{MSE}(\hat{f}(x))$, one should balance the squared bias and the variance terms. The optimal choice of h (in the sense that $\text{MSE}(\hat{f}(x))$ is minimized) should satisfy $d\text{MSE}(\hat{f}(x))/dh = 0$. By using (1.11), it is easy to show that the optimal h that minimizes the leading term of $\text{MSE}(\hat{f}(x))$ is given by

$$h_{\text{opt}} = c(x)n^{-1/5}, \quad (1.15)$$

where $c(x) = \{\kappa f(x)/[\kappa_2 f^{(2)}(x)]^2\}^{1/5}$.

$\text{MSE}(\hat{f}(x))$ is clearly a “pointwise” property, and by using this as the basis for bandwidth selection we are obtaining a bandwidth that is optimal when estimating a density *at a point* x . Examining $c(x)$ in (1.15), we can see that a bandwidth which is optimal for estimation at a point x located in the tail of a distribution will differ from that which is optimal for estimation at a point located at, say, the mode. Suppose that we are interested not in tailoring the bandwidth to the pointwise estimation of $f(x)$ but instead in tailoring the bandwidth globally *for all points* x , that is, for all x in the support of $f(\cdot)$ (the support of x is defined as the set of points of x for which $f(x) > 0$, i.e., $\{x : f(x) > 0\}$). In this case we can choose h optimally by minimizing the “integrated MSE” (IMSE) of $\hat{f}(x)$. Using (1.11) we have

$$\begin{aligned} \text{IMSE}(\hat{f}) &\stackrel{\text{def}}{=} \int \text{E} [\hat{f}(x) - f(x)]^2 dx = \frac{1}{4} h^4 \kappa_2^2 \int [f^{(2)}(x)]^2 dx \\ &\quad + \frac{\kappa}{nh} + o(h^4 + (nh)^{-1}). \end{aligned} \quad (1.16)$$

Again letting h_{opt} denote the optimal smoothing parameter that minimizes the leading terms of (1.16), we use simple calculus to get

$$h_{\text{opt}} = c_0 n^{-1/5}, \quad (1.17)$$

where $c_0 = \kappa_2^{-2/5} \kappa^{1/5} \left\{ \int [f^{(2)}(x)]^2 dx \right\}^{-1/5} > 0$ is a positive constant. Note that if $f^{(2)}(x) = 0$ for (almost) all x , then c_0 is not well defined. For example, if X is, say, uniformly distributed over its support, then $f^{(s)}(x) = 0$ for all x and for all $s \geq 1$, and (1.17) is not defined in this case. It can be shown that in this case (i.e., when X is uniformly

distributed), h_{opt} will have a different rate of convergence equal to $n^{-1/3}$; see the related discussion in Section 1.3.1 and Exercise 1.16.

An interesting extension of the above results can be found in Zinde-Walsh (2005), who examines the asymptotic process for the kernel density estimator by means of generalized functions and generalized random processes and presents novel results for characterizing the behavior of kernel density estimators when the density does not exist, i.e., when the density does not exist as a locally summable function.

1.2 Univariate Bandwidth Selection: Rule-of-Thumb and Plug-In Methods

Equation (1.17) reveals that the optimal smoothing parameter depends on the integrated second derivative of the unknown density through c_0 . In practice, one might choose an initial “pilot value” of h to estimate $\int [f^{(2)}(x)]^2 dx$ nonparametrically, and then use this value to obtain h_{opt} using (1.17). Such approaches are known as “plug-in methods” for obvious reasons. One popular way of choosing the initial h , suggested by Silverman (1986), is to assume that $f(x)$ belongs to a parametric family of distributions, and then to compute h using (1.17). For example, if $f(x)$ is a normal PDF with variance σ^2 , then $\int [f^{(2)}(x)]^2 dx = 3/[8\pi^{1/2}\sigma^5]$. If a standard normal kernel is used, using (1.17), we get the pilot estimate

$$h_{\text{pilot}} = (4\pi)^{-1/10} \left[(3/8)\pi^{-1/2} \right]^{-1/5} \sigma n^{-1/5} \approx 1.06\sigma n^{-1/5}, \quad (1.18)$$

which is then plugged into $\int [\hat{f}^{(2)}(x)]^2 dx$, which then may be used to obtain h_{opt} using (1.17). A clearly undesirable property of the plug-in method is that it is not fully automatic because one needs to choose an initial value of h to estimate $\int [f^{(2)}(x)]^2 dx$ (see Marron, Jones and Sheather (1996) and also Loader (1999) for further discussion).

Often, practitioners will use (1.18) itself for the bandwidth. This is known as the “normal reference rule-of-thumb” approach since it is the optimal bandwidth for a particular family of distributions, in this case the normal family. Should the underlying distribution be “close” to a normal distribution, then this will provide good results, and for exploratory purposes it is certainly computationally attractive. In practice, σ is replaced by the sample standard deviation of $\{X_i\}_{i=1}^n$, while Silverman (1986, p. 47) advocates using a more robust measure

of spread which replaces σ with A , an “adaptive” measure of spread given by

$$A = \min(\text{standard deviation, interquartile range}/1.34).$$

We now turn our attention to a discussion of a number of fully automatic or “data-driven” methods for selecting h that are tailored to the sample at hand.

1.3 Univariate Bandwidth Selection: Cross-Validation Methods

In both theoretical and practical settings, nonparametric kernel estimation has been established as relatively insensitive to choice of kernel function. However, the same cannot be said for bandwidth selection. Different bandwidths can generate radically differing impressions of the underlying distribution. If kernel methods are used simply for “exploratory” purposes, then one might undersmooth the density by choosing a small value of h and let the eye do any remaining smoothing. Alternatively, one might choose a range of values for h and plot the resulting estimates. However, for sound analysis and inference, a principle having some known optimality properties must be adopted. One can think of choosing the bandwidth as being analogous to choosing the number of terms in a series approximation; the more terms one includes in the approximation, the more flexible the resulting model becomes, while the smaller the bandwidth of a kernel estimator, the more flexible it becomes. However, increasing flexibility (reducing potential bias) necessarily leads to increased variability (increasing potential variance). Seen in this light, one naturally appreciates how a number of methods discussed below are motivated by the need to balance the squared bias and variance of the resulting estimate.

1.3.1 Least Squares Cross-Validation

Least squares cross-validation is a fully automatic data-driven method of selecting the smoothing parameter h , originally proposed by Rudemo (1982), Stone (1984) and Bowman (1984) (see also Silverman (1986, pp. 48-51)). This method is based on the principle of selecting a bandwidth that minimizes the integrated squared error of the resulting estimate, that is, it provides an optimal bandwidth tailored to *all* x in the support of $f(x)$.

The integrated squared difference between \hat{f} and f is

$$\int [\hat{f}(x) - f(x)]^2 dx = \int \hat{f}(x)^2 dx - 2 \int \hat{f}(x)f(x) dx + \int f(x)^2 dx. \quad (1.19)$$

As the third term on the right-hand side of (1.19) is unrelated to h , choosing h to minimize (1.19) is therefore equivalent to minimizing

$$\int \hat{f}(x)^2 dx - 2 \int \hat{f}(x)f(x) dx \quad (1.20)$$

with respect to h . In the second term, $\int \hat{f}(x)f(x) dx$ can be written as $E_X[\hat{f}(X)]$, where $E_X(\cdot)$ denotes expectation with respect to X and not with respect to the random observations $\{X_j\}_{j=1}^n$ used for computing $\hat{f}(\cdot)$. Therefore, we may estimate $E_X[\hat{f}(X)]$ by $n^{-1} \sum_{i=1}^n \hat{f}_{-i}(X_i)$ (i.e., replacing E_X by its sample mean), where

$$\hat{f}_{-i}(X_i) = \frac{1}{(n-1)h} \sum_{j=1, j \neq i}^n k\left(\frac{X_i - X_j}{h}\right) \quad (1.21)$$

is the leave-one-out kernel estimator of $f(X_i)$.³ Finally, we estimate the first term $\int \hat{f}(x)^2 dx$ by

$$\begin{aligned} \int \hat{f}(x)^2 dx &= \frac{1}{n^2 h^2} \sum_{i=1}^n \sum_{j=1}^n \int k\left(\frac{X_i - x}{h}\right) k\left(\frac{X_j - x}{h}\right) dx \\ &= \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{k}\left(\frac{X_i - X_j}{h}\right), \end{aligned} \quad (1.22)$$

where $\bar{k}(v) = \int k(u)k(v-u) du$ is the twofold convolution kernel derived from $k(\cdot)$. If $k(v) = \exp(-v^2/2)/\sqrt{2\pi}$, a standard normal kernel, then $\bar{k}(v) = \exp(-v^2/4)/\sqrt{4\pi}$, a normal kernel (i.e., normal PDF) with mean zero and variance two, which follows since two independent $N(0, 1)$ random variables sum to a $N(0, 2)$ random variable.

³Here we emphasize that it is important to use the leave-one-out kernel estimator for computing $E_X(\cdot)$ above. This is because the expectations operator presumes that the X and the X_j 's are independent of one another. Without using the leave-one-out estimator, the cross-validation method will break down; see Exercise 1.6 (iii).

Least squares cross-validation therefore chooses h to minimize

$$CV_f(h) = \frac{1}{n^2 h} \sum_{i=1}^n \sum_{j=1}^n \bar{k} \left(\frac{X_i - X_j}{h} \right) - \frac{2}{n(n-1)h} \sum_{i=1}^n \sum_{j \neq i, j=1}^n k \left(\frac{X_i - X_j}{h} \right), \quad (1.23)$$

which is typically undertaken using numerical search algorithms.

It can be shown that the leading term of $CV_f(h)$ is CV_{f_0} given by (ignoring a term unrelated to h ; see Exercise 1.6)

$$CV_{f_0}(h) = B_1 h^4 + \frac{\kappa}{nh}, \quad (1.24)$$

where $B_1 = (\kappa_2^2/4) [\int [f^{(2)}(x)]^2 dx]$ ($\kappa_2 = \int v^2 k(v) dv$, $\kappa = \int k^2(v) dv$). Thus, as long as $f^{(2)}(x)$ does not vanish for (almost) all x , we have $B_1 > 0$.

Let h^0 denote the value of h that minimizes CV_{f_0} . Simple calculus shows that $h^0 = c_0 n^{-1/5}$ where

$$c_0 = [\kappa/(4B_1)]^{1/5} = \kappa^{1/5} \kappa_2^{-2/5} \left\{ \left[\int f^{(2)}(x) \right]^2 dx \right\}^{-1/5}.$$

A comparison of h^0 with h_{opt} in (1.17) reveals that the two are identical, i.e., $h^0 \equiv h_{\text{opt}}$. This arises because h_{opt} minimizes $\int E[\hat{f}(x) - f(x)]^2 dx$, while h^0 minimizes $E[CV_f(h)]$, the leading term of $CV_f(h)$. It can be easily seen that $E[CV_f(h)] + \int f(x)^2 dx$ is an alternative version of $\int E[\hat{f}(x) - f(x)]^2 dx$; hence, $E[CV_f(h)] + \int f(x)^2 dx$ also estimates $\int E[\hat{f}(x) - f(x)]^2 dx$. Given that $\int f(x)^2 dx$ is unrelated to h , one would expect that h^0 and h_{opt} should be the same.

Let \hat{h} denote the value of h that minimizes $CV_f(h)$. Given that $CV_f(h) = CV_{f_0} + (s.o.)$, where $(s.o.)$ denotes smaller order terms (than CV_{f_0}) and terms unrelated to h , it can be shown that $\hat{h} = h^0 + o_p(h^0)$, or, equivalently, that

$$\frac{\hat{h} - h^0}{h^0} \equiv \frac{\hat{h}}{h^0} - 1 \rightarrow 0 \text{ in probability.} \quad (1.25)$$

Intuitively, (1.25) is easy to understand because $CV_f(h) = CV_{f_0}(h) + (s.o.)$, thus asymptotically an h that minimizes $CV_f(h)$ should be

close to an h that minimizes $CV_{f_0}(h)$; therefore, we expect that \hat{h} and h^0 will be close to each other in the sense of (1.25). Härdle, Hall and Marron (1988) showed that $(\hat{h} - h^0)/h^0 = O_p(n^{-1/10})$, which indeed converges to zero (in probability) but at an extremely slow rate.

We again underscore the need to use the leave-one-out kernel estimator when constructing CV_f as given in (1.23). If instead one were to use the standard kernel estimator, least squares cross-validation will break down, yielding $\hat{h} = 0$. Exercise 1.6 shows that if one does not use the leave-one-out kernel estimator when estimating $f(X_i)$, then $h = 0$ minimizes the objective function, which of course violates the consistency condition that $nh \rightarrow \infty$ as $n \rightarrow \infty$.

Here we implicitly impose the restriction that $f^{(2)}(x)$ is not a zero function, which rules out the case for which $f(x)$ is a uniform PDF. In fact this condition can be relaxed. Stone (1984) showed that, as long as $f(x)$ is bounded, then the least squares cross-validation method will select h optimally in the sense that

$$\frac{\int [\hat{f}(x, \hat{h}) - f(x)]^2 dx}{\inf_h \int [\hat{f}(x, h) - f(x)]^2 dx} \rightarrow 1 \text{ almost surely,} \quad (1.26)$$

where $\hat{f}(x, \hat{h})$ denotes the kernel estimator of $f(x)$ with cross-validation selected \hat{h} , and $\hat{f}(x, h)$ is the kernel estimator with a generic h . Obviously, the ratio defined in (1.26) should be greater than or equal to one for any n . Therefore, Stone's (1984) result states that, asymptotically, cross-validated smoothing parameter selection is optimal in the sense of minimizing the estimation integrated square error. In Exercise 1.16 we further discuss the intuition underlying why $\hat{h} \rightarrow 0$ even when $f(x)$ is a uniform PDF.

1.3.2 Likelihood Cross-Validation

Likelihood cross-validation is another automatic data-driven method for selecting the smoothing parameter h . This approach yields a density estimate which has an entropy theoretic interpretation, since the estimate will be close to the actual density in a Kullback-Leibler sense. This approach was proposed by Duin (1976).

Likelihood cross-validation chooses h to maximize the (leave-one-out) log likelihood function given by

$$\mathcal{L} = \ln L = \sum_{i=1}^n \ln \hat{f}_{-i}(X_i),$$

where $\hat{f}_{-i}(X_i)$ is the leave-one-out kernel estimator of $f(X_i)$ defined in (1.21). The main problem with likelihood cross-validation is that it is severely affected by the tail behavior of $f(x)$ and can lead to inconsistent results for fat tailed distributions when using popular kernel functions (see Hall (1987*a*, 1987*b*)). For this reason the likelihood cross-validation method has elicited little interest in the statistical literature.

However, the likelihood cross-validation method may work well for a range of standard distributions (i.e., thin tailed). We consider the performance of likelihood cross-validation in Section 1.3.3, when we compare the impact of different bandwidth selection methods on the resulting density estimate, and in Section 1.13, where we consider empirical applications.

1.3.3 An Illustration of Data-Driven Bandwidth Selection

Figure 1.1 presents kernel estimates constructed from $n = 500$ observations drawn from a simulated bimodal distribution. The second order Gaussian (normal) kernel was used throughout, and least squares cross-validation was used to select the bandwidth for the estimate appearing in the upper left plot of the figure, with $h_{\text{lscv}} = 0.19$. We also plot the estimate based on the normal reference rule-of-thumb ($h_{\text{ref}} = 0.34$) along with an undersmoothed estimate ($1/5 \times h_{\text{lscv}}$) and an oversmoothed estimate ($5 \times h_{\text{lscv}}$).⁴

Figure 1.1 reveals that least squares cross-validation appears to yield a reasonable density estimate for this data, while the reference rule-of-thumb is inappropriate as it oversmooths somewhat. Extreme oversmoothing can lead to a unimodal estimate which completely obscures the true bimodal nature of the underlying distribution. Also, undersmoothing leads to too many false modes. See Exercise 1.17 for an empirical application that investigates the effects of under- and over-smoothing on the resulting density estimate.

1.4 Univariate CDF Estimation

In Section 1.1 we introduced the empirical CDF estimator $F_n(x)$ given in (1.2), while Exercise 1.4 shows that it is a \sqrt{n} -consistent estimator

⁴Likelihood cross-validation yielded a bandwidth of $h_{\text{mlcv}} = 0.15$, which results in a density estimate virtually identical to that based upon least squares cross-validation for this dataset.

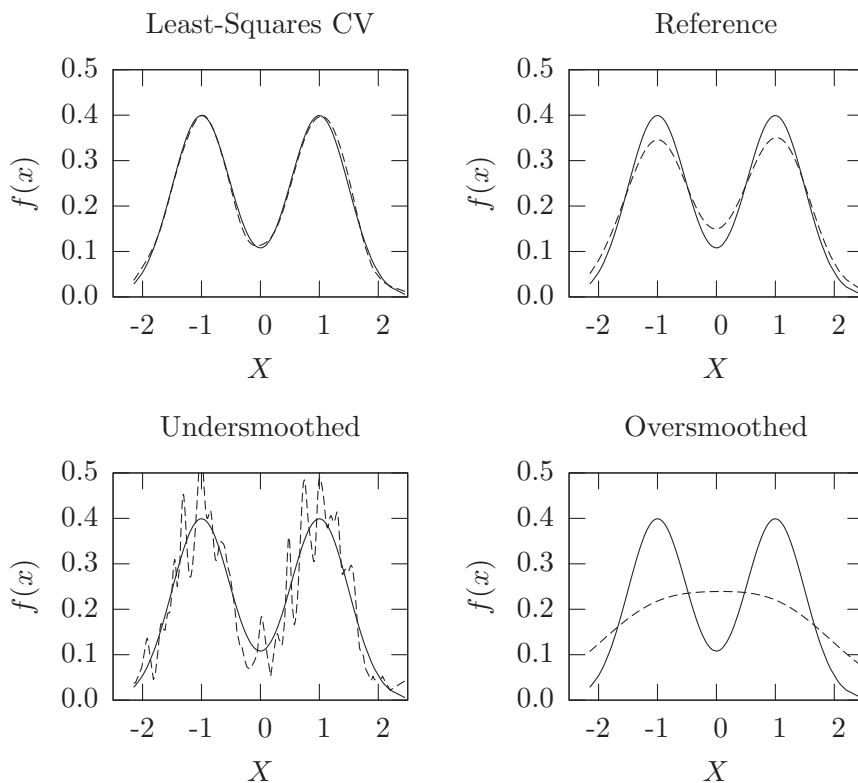


Figure 1.1: Univariate kernel estimates of a mixture of normals using least squares cross-validation, the normal reference rule-of-thumb, undersmoothing, and oversmoothing ($n = 500$). The correct parametric data generating process appears as the solid line, the kernel estimate as the dashed line.

of $F(x)$. However, this empirical CDF $F_n(x)$ is not smooth as it jumps by $1/n$ at each sample realization point. One can, however, obtain a smoothed estimate of $F(x)$ by integrating $\hat{f}(x)$. Define

$$\hat{F}(x) = \int_{-\infty}^x \hat{f}(v) dv = \frac{1}{n} \sum_{i=1}^n G\left(\frac{x - X_i}{h}\right), \quad (1.27)$$

where $G(x) = \int_{-\infty}^x k(v) dv$ is a CDF (which follows directly because $k(\cdot)$ is a PDF; see (1.10)). The next theorem provides the MSE of $\hat{F}(x)$.

Theorem 1.2. *Under conditions given in Bowman, Hall and Prvan (1998), in particular, assuming that $F(x)$ is twice continuously differentiable, $k(v) = dG(v)/dv$ is bounded, symmetric, and compactly supported, and that $d^2F(x)/dx^2$ is Hölder-continuous, $0 \leq h \leq Cn^{-\epsilon}$ for some $0 < \epsilon < \frac{1}{8}$, then as $n \rightarrow \infty$,*

$$\begin{aligned} \text{MSE}(\hat{F}) &= \mathbb{E} \left[\hat{F}(x) - F(x) \right]^2 \\ &= c_0(x)n^{-1} - c_1(x)hn^{-1} + c_2(x)h^4 + o(h^4 + hn^{-1}), \end{aligned}$$

where $c_0 = F(x)(1 - F(x))$, $c_1(x) = \alpha_0 f(x)$, $\alpha_0 = 2 \int vG(v)k(v) dv$, $f(x) = dF(x)/dx$, $c_2(x) = [(\kappa_2/2)F^{(2)}(x)]^2$, $\kappa_2 = \int v^2k(v) dv$, and where $F^{(s)}(x) = d^sF(x)/dx^s$ is the s th derivative of $F(x)$.

Proof. Note that $\mathbb{E} \left[\hat{F}(x) \right] = \mathbb{E} \left[G \left(\frac{x - X_i}{h} \right) \right]$. Then we have ($f = \int_{-\infty}^{\infty}$)

$$\begin{aligned} \mathbb{E} \left[G \left(\frac{x - X_i}{h} \right) \right] &= \int G \left(\frac{x - z}{h} \right) f(z) dz \\ &= h \int G(v) f(x - hv) dv = - \int G(v) dF(x - hv) \\ &= - [G(v)F(x - hv)] \Big|_{v=-\infty}^{v=\infty} + \int k(v)F(x - hv) dv \\ &= \int k(v) \left[F(x) - F^{(1)}(x)hv + (1/2)h^2F^{(2)}(x)v^2 \right] dv \\ &\quad + o(h^2) \\ &= F(x) + (1/2)\kappa_2h^2F^{(2)}(x) + o(h^2), \end{aligned} \tag{1.28}$$

where at the second equality above we used

$$- \int_{\infty}^{-\infty} [\dots] dv = \int_{-\infty}^{\infty} [\dots] dv.$$

Also note that we did not use a Taylor expansion in $\int G(v)F(x - hv) dv$ since $\int v^m G(v) dv = +\infty$ for any $m \geq 0$. We first used integration by parts to get $k(v)$, and then used the Taylor expansion since $\int v^m k(v) dv$ is usually finite. For example, if $k(v)$ has bounded support or $k(v)$ is a standard normal kernel function, then $\int v^m k(v) dv$ is finite for any $m \geq 0$.

Similarly,

$$\begin{aligned}
 \mathbb{E} \left[G^2 \left(\frac{x - X_i}{h} \right) \right] &= \int G^2 \left(\frac{x - z}{h} \right) f(z) dz = h \int G^2(v) f(x - hv) dv \\
 &= - \int G^2(v) dF(x - hv) \\
 &= 2 \int G(v) k(v) F(x - hv) dv \\
 &= 2 \int G(v) k(v) [F(x) - F^{(1)}(x)hv] dv + O(h^2) \\
 &= F(x) - \alpha_0 h f(x) + O(h^2),
 \end{aligned} \tag{1.29}$$

where $\alpha_0 = 2 \int v G(v) k(v) dv$, and where we have used the fact that

$$2 \int_{-\infty}^{\infty} G(v) k(v) dv = \int_{-\infty}^{\infty} dG^2(v) = G^2(\infty) - G^2(-\infty) = 1,$$

because $G(\cdot)$ is a (user-specified) CDF kernel function.

From (1.28) we have $\text{bias}[\hat{F}(x)] = (1/2)\kappa_2 h^2 F^{(2)}(x) + o(h^2)$, and from (1.28) and (1.29) we have

$$\begin{aligned}
 \text{var} \left[\hat{F}(x) \right] &= n^{-1} \text{var} \left[G \left(\frac{x - X_i}{h} \right) \right] \\
 &= n^{-1} \left\{ \mathbb{E} \left[G^2 \left(\frac{x - X_i}{h} \right) \right] - \left[\mathbb{E} G \left(\frac{x - X_i}{h} \right) \right]^2 \right\} \\
 &= n^{-1} F(x) [1 - F(x)] - \alpha_0 f(x) h n^{-1} + o(h/n).
 \end{aligned}$$

Hence,

$$\begin{aligned}
 \mathbb{E} \left(\hat{F}(x) - F(x) \right)^2 &= \left[\text{bias} \left(\hat{F}(x) \right) \right]^2 + \text{var} \left[\hat{F}(x) \right] \\
 &= n^{-1} F(x) [1 - F(x)] + h^4 (\kappa_2/2)^2 \left[F^{(2)}(x) \right]^2 \\
 &\quad - \alpha_0 f(x) \frac{h}{n} + o(h^4 + n^{-1}h).
 \end{aligned} \tag{1.30}$$

This completes the proof of Theorem 1.2. □

From Theorem 1.2 we immediately obtain the following result on the IMSE of \hat{F} :

$$\begin{aligned} \text{IMSE}(\hat{F}) &= \int \text{E} \left[\hat{F}(x) - F(x) \right]^2 dx \\ &= C_0 n^{-1} - C_1 h n^{-1} + C_2 h^4 + o(h^4 + h n^{-1}), \end{aligned} \quad (1.31)$$

where $C_j = \int c_j(x) dx$ ($j = 0, 1, 2$). Letting h_0 denote the value of h that minimizes the leading term of IMSE, we obtain

$$h_0 = a_0 n^{-1/3},$$

where $a_0 = [C_1/(4C_2)]^{1/3}$, hence the optimal smoothing parameter for estimating univariate a CDF has a faster rate of convergence than the optimal smoothing parameter for estimating a univariate PDF ($n^{-1/3}$ versus $n^{-1/5}$). With $h \sim n^{-1/3}$, we have $h^2 = O(n^{-2/3}) = o(n^{-1/2})$. Hence, $\sqrt{n}[\hat{F}(x) - F(x)] \rightarrow N(0, F(x)[1 - F(x)])$ in distribution by the Liapunov central limit theorem (CLT); see Theorem A.5 in Appendix A for this and a range of other useful CLTs.

As is the case for nonparametric PDF estimation, nonparametric CDF estimation has widespread potential application though it is not nearly as widely used. For instance, it can be used to test stochastic dominance without imposing parametric assumptions on the underlying CDFs; see, e.g., Barrett and Donald (2003) and Linton, Whang and Maasoumi (2005).

1.5 Univariate CDF Bandwidth Selection: Cross-Validation Methods

Bowman et al. (1998) suggest choosing h for $\hat{F}(x)$ by minimizing the following cross-validation function:

$$CV_F(h) = \frac{1}{n} \sum_{i=1}^n \int \left\{ \mathbf{1}(X_i \leq x) - \hat{F}_{-i}(x) \right\}^2 dx, \quad (1.32)$$

where $\hat{F}_{-i}(x) = (n-1)^{-1} \sum_{j \neq i}^n G\left(\frac{x-X_j}{h}\right)$ is the leave-one-out estimator of $F(x)$.

Bowman et al. (1998) show that $CV_F = \text{E}[CV_F] + (s.o.)$ and that

(see Exercise 1.9)

$$\begin{aligned} E[CV_F(h)] &= \int F(1-F) dx + \frac{1}{n-1} \int F(1-F) dx - C_1 hn^{-1} \\ &\quad + C_2 h^4 + o(hn^{-1} + h^4). \end{aligned} \tag{1.33}$$

We observe that (1.33) has the same leading term as $\text{IMSE}(\hat{F})$ given in (1.31). Thus, asymptotically, selecting h via cross-validation leads to the same asymptotic optimality property for $\hat{F}(x)$ that would arise when using h_0 , the optimal deterministic smoothing parameter. If we let \hat{h} denote the cross-validated smoothing parameter, then it can be shown that $\hat{h}/h_0 \rightarrow 1$ in probability. Note that when using \hat{h} , the asymptotic distribution of $\hat{F}(x, \hat{h})$ is the same as $\hat{F}(x, h_0)$ (by using a stochastic equicontinuity argument as outlined in Appendix A), that is,

$$\sqrt{n} \left(\hat{F}(x) - F(x) \right) \xrightarrow{d} N(0, F(x)(1-F(x))), \tag{1.34}$$

where $\hat{F}(x)$ is defined in (1.27) with h replaced by \hat{h} . Note that no bias term appears in (1.34) since $\text{bias}(\hat{F}(x)) = O(h_0^2) = O(n^{-2/3}) = o(n^{-1/2})$, which was not the case for PDF estimation. Here the squared bias term has order smaller than the leading variance term of $O(n^{-1})$ (i.e., $\text{var}(\hat{F}(x)) = O(n^{-1})$).

We now turn our attention to a generalization of the univariate kernel estimators developed above, namely multivariate kernel estimators. Again, we consider only the continuous case in this chapter; we tackle discrete and mixed continuous and discrete data cases in Chapters 3 and 4.

1.6 Multivariate Density Estimation

Suppose that X_1, \dots, X_n constitute an i.i.d. q -vector ($X_i \in \mathbb{R}^q$, for some $q > 1$) having a common PDF $f(x) = f(x_1, x_2, \dots, x_q)$. Let X_{is} denote the s th component of X_i ($s = 1, \dots, q$). Using a “product kernel function” constructed from the product of univariate kernel functions, we estimate the PDF $f(x)$ by

$$\hat{f}(x) = \frac{1}{nh_1 \dots h_q} \sum_{i=1}^n K \left(\frac{X_i - x}{h} \right), \tag{1.35}$$

where $K\left(\frac{X_i-x}{h}\right) = k\left(\frac{X_{i1}-x_1}{h_1}\right) \times \cdots \times k\left(\frac{X_{iq}-x_q}{h_q}\right)$, and where $k(\cdot)$ is a univariate kernel function satisfying (1.10).

The proof of MSE consistency of $\hat{f}(x)$ is similar to the univariate case. In particular, one can show that

$$\text{bias}\left(\hat{f}(x)\right) = \frac{\kappa_2}{2} \sum_{s=1}^q h_s^2 f_{ss}(x) + O\left(\sum_{s=1}^q h_s^3\right), \quad (1.36)$$

where $f_{ss}(x)$ is the second order derivative of $f(x)$ with respect to x_s , $\kappa_2 = \int v^2 k(v) dv$, and one can also show that

$$\text{var}\left(\hat{f}(x)\right) = \frac{1}{nh_1 \dots h_q} \left[\kappa^q f(x) + O\left(\sum_{s=1}^q h_s^2\right) \right] = O\left(\frac{1}{nh_1 \dots h_q}\right), \quad (1.37)$$

where $\kappa = \int k^2(v) dv$. The proofs of (1.36) and (1.37), which are similar to the univariate X case, are left as an exercise (see Exercise 1.11).

Summarizing, we obtain the result

$$\begin{aligned} \text{MSE}\left(\hat{f}(x)\right) &= \left[\text{bias}\left(\hat{f}(x)\right)\right]^2 + \text{var}\left(\hat{f}(x)\right) \\ &= O\left(\left(\sum_{s=1}^q h_s^2\right)^2 + (nh_1 \dots h_q)^{-1}\right). \end{aligned}$$

Hence, if as $n \rightarrow \infty$, $\max_{1 \leq s \leq q} h_s \rightarrow 0$ and $nh_1 \dots h_q \rightarrow \infty$, then we have $\hat{f}(x) \rightarrow f(x)$ in MSE, which implies that $\hat{f}(x) \rightarrow f(x)$ in probability.

As we saw in the univariate case, the optimal smoothing parameters h_s should balance the squared bias and variance terms, i.e., $h_s^4 = O((nh_1 \dots h_q)^{-1})$ for all s . Thus, we have $h_s = c_s n^{-1/(q+4)}$ for some positive constant c_s ($s = 1, \dots, q$). The cross-validation methods discussed in Section 1.3 can be easily generalized to the multivariate data setting, and we can show that least squares cross-validation can optimally select the h_s 's in the sense outlined in Section 1.3 (see Section 1.8 below).

We briefly remark on the independence assumption invoked for the proofs presented above. Our assumption was that the data is independent across the i index. Note that no restrictions were placed on the s index for each component X_{is} ($s = 1, \dots, q$). The product kernel is used simply for convenience, and it certainly *does not* require that the X_{is} 's

are independent across the s index. In other words, the multivariate kernel density estimator (1.35) is capable of capturing general dependence among the different components of X_i . Furthermore, we shall relax the “independence across observations” assumption in Chapter 18, and will see that all of the results developed above carry over to the weakly dependent data setting.

1.7 Multivariate Bandwidth Selection: Rule-of-Thumb and Plug-In Methods

In Section 1.2 we discussed the use of the so-called normal reference rule-of-thumb and plug-in methods in a univariate setting. The generalization of the univariate normal reference rule-of-thumb to a multivariate setting is straightforward. Letting q be the dimension of X_i , one can choose $h_s = c_s X_{s, sd} n^{-1/(4+q)}$ for $s = 1, \dots, q$, where $X_{s, sd}$ is the sample standard deviation of $\{X_{is}\}_{i=1}^n$ and c_s is a positive constant. In practice one still faces the problem of how to choose c_s . The choice of $c_s = 1.06$ for all $s = 1, \dots, q$ is computationally attractive; however, this selection treats the different X_{is} 's symmetrically. In practice, should the joint PDF change rapidly in one dimension (say in x_1) but change slowly in another (say in x_2), then one should select a relatively small value of c_1 (hence a small h_1) and a relatively large value for c_2 (h_2). Unlike the cross-validation methods that we will discuss shortly, rule-of-thumb methods do not offer this flexibility.

For plug-in methods, on the other hand, the leading (squared) bias and variance terms of $\hat{f}(x)$ must be estimated, and then h_1, \dots, h_q must be chosen to minimize the leading MSE term of $\hat{f}(x)$. However, the leading MSE term of $\hat{f}(x)$ involves the unknown $f(x)$ and its partial derivative functions, and pilot bandwidths must be selected for *each* variable in order to estimate these unknown functions. How to best select the initial pilot smoothing parameters can be tricky in high-dimensional settings, and the plug-in methods are not widely used in applied settings to the best of our knowledge, nor would we counsel their use other than for exploratory data analysis.

1.8 Multivariate Bandwidth Selection: Cross-Validation Methods

1.8.1 Least Squares Cross-Validation

The univariate least squares cross-validation method discussed in Section 1.3.1 can be readily generalized to the multivariate density estimation setting. Replacing the univariate kernel function in (1.23) by a multivariate product kernel, the cross-validation objective function now becomes

$$CV_f(h_1, \dots, h_q) = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \bar{K}_h(X_i, X_j) - \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j \neq i, j=1}^n K_h(X_i, X_j), \quad (1.38)$$

where

$$K_h(X_i, X_j) = \prod_{s=1}^q h_s^{-1} k\left(\frac{X_{is} - X_{js}}{h_s}\right),$$

$$\bar{K}_h(X_i, X_j) = \prod_{s=1}^q h_s^{-1} \bar{k}\left(\frac{X_{is} - X_{js}}{h_s}\right),$$

and $\bar{k}(v)$ is the twofold convolution kernel based upon $k(\cdot)$, where $k(\cdot)$ is a univariate kernel function satisfying (1.10).

Exercise 1.12 shows that the leading term of $CV_f(h_1, \dots, h_q)$ is given by (ignoring a term unrelated to the h_s 's)

$$CV_{f0}(h_1, \dots, h_q) = \int \left[\sum_{s=1}^q B_s(x) h_s^2 \right]^2 dx + \frac{\kappa^q}{n h_1 \dots h_q}, \quad (1.39)$$

where $B_s(x) = (\kappa_2/2) f_{ss}(x)$.

Defining a_s via $h_s = a_s n^{-1/(q+4)}$ ($s = 1, \dots, q$), we have

$$CV_{f0}(h_1, \dots, h_q) = n^{-4/(q+4)} \chi_f(a_1, \dots, a_q), \quad (1.40)$$

where

$$\chi_f(a_1, \dots, a_q) = \int \left[\sum_{s=1}^q B_s(x) a_s^2 \right]^2 dx + \frac{\kappa^q}{a_1 \dots a_q}. \quad (1.41)$$

Let the a_s^0 's be the values of the a_s 's that minimize $\chi_f(a_1, \dots, a_q)$. Under the same conditions used in the univariate case and, in addition, assuming that $f_{ss}(x)$ is not a zero function for all s , Li and Zhou (2005) show that each a_s^0 is uniquely defined, positive, and finite (see Exercise 1.10). Let h_1^0, \dots, h_q^0 denote the values of h_1, \dots, h_q that minimize CV_{f_0} . Then from (1.40) we know that $h_s^0 = a_s^0 n^{-1/(q+4)} = O(n^{-1/(q+4)})$.

Exercise 1.12 shows that CV_{f_0} is also the leading term of $E[CV_f]$. Therefore, the nonstochastic smoothing parameters h_s^0 can be interpreted as optimal smoothing parameters that minimize the leading term of the IMSE.

Let $\hat{h}_1, \dots, \hat{h}_q$ denote the values of h_1, \dots, h_q that minimize CV_f . Using the fact that $CV_f = CV_{f_0} + (s.o.)$, we can show that $\hat{h}_s = h_s^0 + o_p(h_s^0)$. Thus, we have

$$\frac{\hat{h}_s - h_s^0}{h_s^0} = \frac{\hat{h}_s}{h_s^0} - 1 \rightarrow 0 \quad \text{in probability, for } s = 1, \dots, q. \quad (1.42)$$

Therefore, smoothing parameters selected via cross-validation have the same asymptotic optimality properties as the nonstochastic optimal smoothing parameters.

Note that if $f_{ss}(x) = 0$ almost everywhere (a.e.) for some s , then $B_s = 0$ and the above result does not hold. Stone (1984) shows that the cross-validation method still selects h_1, \dots, h_q optimally in the sense that the integrated estimation square error is minimized; see also Ouyang et al. (2006) for a more detailed discussion of this case.

1.8.2 Likelihood Cross-Validation

Likelihood cross-validation for multivariate models follows directly via (multivariate) maximization of the likelihood function outlined in Section 1.3.2, hence we do not go into further details here. However, we do point out that, though straightforward to implement, it suffers from the same defects outlined for the univariate case in the presence of fat tail distributions (i.e., it has a tendency to oversmooth in such situations).

1.9 Asymptotic Normality of Density Estimators

In this section we show that $\hat{f}(x)$ has an asymptotic normal distribution. The most popular CLT is the Lindeberg-Levy CLT given in

Theorem A.3 of Appendix A, which states that $n^{1/2}[n^{-1}\sum_{i=1}^n Z_i] \rightarrow N(0, \sigma^2)$ in distribution, provided that Z_i is i.i.d. $(0, \sigma^2)$. Though the Lindeberg-Levy CLT can be used to derive the asymptotic distribution of various semiparametric estimators discussed in Chapters 7, 8, and 9, it cannot be used to derive the asymptotic distribution of $\hat{f}(x)$. This is because $\hat{f}(x) = n^{-1}\sum_i Z_{i,n}$, where the summand $Z_{i,n} = K_h(X_i, x)$ depends on n (since $h = h(n)$). We shall make use of the Liapunov CLT given in Theorem A.5 of Appendix A

Theorem 1.3. *Let X_1, \dots, X_n be i.i.d. q -vectors with its PDF $f(\cdot)$ having three-times bounded continuous derivatives. Let x be an interior point of the support of X . If, as $n \rightarrow \infty$, $h_s \rightarrow 0$ for all $s = 1, \dots, q$, $nh_1 \dots h_q \rightarrow \infty$, and $(nh_1 \dots h_q) \sum_{s=1}^q h_s^6 \rightarrow 0$, then*

$$\sqrt{nh_1 \dots h_q} \left[\hat{f}(x) - f(x) - \frac{\kappa_2}{2} \sum_{s=1}^q h_s^2 f_{ss}(x) \right] \xrightarrow{d} N(0, \kappa^q f(x)).$$

Proof. Using (1.36) and (1.37), one can easily show that

$$\sqrt{nh_1 \dots h_q} \left[\hat{f}(x) - f(x) - \frac{\kappa_2}{2} \sum_{s=1}^q h_s^2 f_{ss}(x) \right]$$

has asymptotic mean zero and asymptotic variance $\kappa^q f(x)$, i.e.,

$$\begin{aligned} & \sqrt{nh_1 \dots h_q} \left[\hat{f}(x) - f(x) - \frac{\kappa_2}{2} \sum_{s=1}^q h_s^2 f_{ss}(x) \right] \\ &= \sqrt{nh_1 \dots h_q} \left[\hat{f}(x) - \mathbb{E}(\hat{f}(x)) \right] \\ & \quad + \sqrt{nh_1 \dots h_q} \left[\mathbb{E}(\hat{f}(x)) - f(x) - \frac{\kappa_2}{2} \sum_{s=1}^q h_s^2 f_{ss}(x) \right] \\ &= \sqrt{nh_1 \dots h_q} \left[\hat{f}(x) - \mathbb{E}(\hat{f}(x)) \right] \\ & \quad + O\left(\sqrt{nh_1 \dots h_q} \sum_{s=1}^q h_s^3 \right) \quad (\text{by (1.36)}) \\ &= \sum_{i=1}^n (nh_1 \dots h_q)^{-1/2} \\ & \quad \times \left[K\left(\frac{X_i - x}{h}\right) - \mathbb{E}\left(K\left(\frac{X_i - x}{h}\right)\right) \right] + o(1) \\ &\equiv \sum_{i=1}^n Z_{n,i} + o(1) \xrightarrow{d} N(0, \kappa^q f(x)), \end{aligned}$$

by Liapunov's CLT, provided we can verify that Liapunov's CLT condition (A.21) holds, where

$$Z_{n,i} = (nh_1 \dots h_q)^{-1/2} \left[K \left(\frac{X_i - x}{h} \right) - E \left(K \left(\frac{X_i - x}{h} \right) \right) \right]$$

and

$$\sum_{i=1}^n \sigma_{n,i}^2 \stackrel{\text{def}}{=} \sum_{i=1}^n \text{var}(Z_{n,i}) = \kappa^q f(x) + o(1)$$

by (1.37). Pagan and Ullah (1999, p. 40) show that (A.21) holds under the condition given in Theorem 1.3. The condition that $\int k(v)^{2+\delta} dv < \infty$ for some $\delta > 0$ used in Pagan and Ullah is implied by our assumption that $k(v)$ is nonnegative and bounded, and that $\int k(v) dv = 1$, because $\int k(v)^{2+\delta} dv \leq C \int k(v) dv = C$ is finite, where $C = \sup_{v \in \mathbb{R}^q} k(v)^{1+\delta}$. \square

1.10 Uniform Rates of Convergence

Up to now we have demonstrated only the case of pointwise and IMSE consistency (which implies consistency in probability). In this section we generalize pointwise consistency in order to obtain a stronger “uniform consistency” result. We will prove that nonparametric kernel estimators are uniformly almost surely consistent and derive their uniform almost sure rate of convergence. Almost sure convergence implies convergence in probability; however, the converse is not true, i.e., convergence in probability may not imply convergence almost surely; see Serfling (1980) for specific examples.

We have already established pointwise consistency for an interior point in the support of X . However, it turns out that popular kernel functions such as (1.9) may not lead to consistent estimation of $f(x)$ when x is at the boundary of its support, hence we need to exclude the boundary ranges when considering the uniform convergence rate. This highlights an important aspect of kernel estimation in general, and a number of kernel estimators introduced in later sections are motivated by the desire to mitigate such “boundary effects.” We first show that when x is at (or near) the boundary of its support, $\hat{f}(x)$ may not be a consistent estimator of $f(x)$.

Consider the case where X is univariate having bounded support. For simplicity we assume that $X \in [0, 1]$. The pointwise consistency result $\hat{f}(x) - f(x) = o_p(1)$ obtained earlier requires that x lie in the

interior of its support. Exercise 1.13 shows that, for x at the boundary of its support, $\text{MSE}(\hat{f}(x))$ may not be $o(1)$. Therefore, some modifications may be needed to consistently estimate $f(x)$ for x at the boundary of its support. Typical modifications include the use of boundary kernels or data reflection (see Gasser and Müller (1979), Hall and Wehrly (1991), and Scott (1992, pp. 148–149)). By way of example, consider the case where x lies on its lowermost boundary, i.e., $x = 0$, hence $\hat{f}(0) = (nh)^{-1} \sum_{i=1}^n K((X_i - 0)/h)$. Exercise 1.13 shows that for this case, $E[\hat{f}(0)] = f(0)/2 + O(h)$. Therefore, $\text{bias}[\hat{f}(0)] = E[\hat{f}(0)] - f(0) = -f(0)/2 + O(h)$, which will not converge to zero if $f(0) \neq 0$ (when $f(0) > 0$).

In the literature, various boundary kernels are proposed to overcome the boundary (bias) problem. For example, a simple boundary corrected kernel is given by (assuming that $X \in [0, 1]$)

$$k_h(x, y) = \begin{cases} h^{-1}k\left(\frac{y-x}{h}\right) / \int_{-x/h}^{\infty} k(v) dv & \text{if } x \in [0, h) \\ h^{-1}k\left(\frac{y-x}{h}\right) & \text{if } x \in [h, 1-h] \\ h^{-1}k\left(\frac{y-x}{h}\right) / \int_{-\infty}^{(1-x)/h} k(v) dv & \text{if } x \in (1-h, 1], \end{cases} \quad (1.43)$$

where $k(\cdot)$ is a second order kernel satisfying (1.10). Now, we estimate $f(x)$ by

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^n k_h(x, X_i), \quad (1.44)$$

where $k_h(x, X_i)$ is defined in (1.43). Exercise 1.14 shows that the above boundary corrected kernel successfully overcomes the boundary problem.

We now establish the uniform almost sure convergence rate of $\hat{f}(x) - f(x)$ for $x \in \mathcal{S}$, where \mathcal{S} is a bounded set excluding the boundary range of the support of X . In the above example, when the support of x is $[0, 1]$, we can choose $\mathcal{S} = [\epsilon, 1 - \epsilon]$ for arbitrarily *small* positive ϵ ($0 < \epsilon < 1/2$). We assume that $f(x)$ is bounded below by a positive constant on \mathcal{S} .

Theorem 1.4. *Under smoothness conditions on $f(\cdot)$ given in Masry (1996b), and also assuming that $\inf_{x \in \mathcal{S}} f(x) \geq \delta > 0$, we have*

$$\sup_{x \in \mathcal{S}} |\hat{f}(x) - f(x)| = O\left(\frac{(\ln(n))^{1/2}}{(nh_1 \dots h_q)^{1/2}} + \sum_{s=1}^q h_s^2\right) \text{ almost surely.}$$

A detailed proof of Theorem 1.4 is given in Section 1.12.

Since almost sure convergence implies convergence in probability, the uniform rate also holds in probability, i.e., under the same conditions as in Theorem 1.4, we have

$$\sup_{x \in \mathcal{S}} \left| \hat{f}(x) - f(x) \right| = O_p \left(\frac{(\ln(n))^{1/2}}{(nh_1 \dots h_q)^{1/2}} + \sum_{s=1}^q h_s^2 \right).$$

Using the results of (1.36) and (1.37), we can establish the following uniform MSE rate.

Theorem 1.5. *Assuming that $f(x)$ is twice differentiable with bounded second derivatives, then we have*

$$\sup_{x \in \mathcal{S}} \mathbb{E} \left\{ \left[\hat{f}(x) - f(x) \right]^2 \right\} = O \left(\sum_{s=1}^q h_s^4 + (nh_1 \dots h_q)^{-1} \right).$$

Proof. This follows from (1.36) and (1.37), by noting that $\sup_{x \in \mathcal{S}} f(x)$ and $\sup_{x \in \mathcal{S}} |f_{ss}(x)|$ are both finite ($s = 1, \dots, q$). \square

Note that although convergence in MSE implies convergence in probability, one cannot derive the uniform convergence rate in probability from Theorem 1.5. This is because

$$\mathbb{E} \left\{ \sup_{x \in \mathcal{S}} \left[\hat{f}(x) - f(x) \right]^2 \right\} \neq \sup_{x \in \mathcal{S}} \mathbb{E} \left[\hat{f}(x) - f(x) \right]^2,$$

and

$$\mathbb{P} \left[\sup_{x \in \mathcal{S}} \left| \hat{f}(x) - f(x) \right| > \epsilon \right] \neq \sup_{x \in \mathcal{S}} \mathbb{P} \left[\left| \hat{f}(x) - f(x) \right| > \epsilon \right].$$

The sup and the $\mathbb{E}(\cdot)$ or the $\mathbb{P}(\cdot)$ operators do not commute with one another.

Cheng (1997) proposes alternative (local linear) density estimators that achieve automatic boundary corrections and enjoy some typical optimality properties. Cheng also suggests a data-based bandwidth selector (in the spirit of plug-in rules), and demonstrates that the bandwidth selector is very efficient regardless of whether there are non-smooth boundaries in the support of the density.

(continued...)

Subject Index

- absolutely continuous, 677
- additive model, 283
- additive partially linear model, 297
- almost everywhere (a.e.), 667
- applications
 - adolescent growth, 44, 92, 202
 - Boston housing, 200
 - conditionally independent private information auctions, 648
 - continuous time models, 627
 - dining out, 145
 - direct marketing, 277
 - extramarital affairs, 172
 - female labor force participation, 175
 - first price auction models, 645
 - growth convergence clubs, 385
 - inflation forecasting, 93
 - interest rate forecasting, 564
 - Italian income, 45, 206
 - job prestige, 92
 - labor productivity, 177
 - OECD growth rates, 178, 207
 - old faithful geyser, 44
 - political corruption, 171
 - right-heart catheterization, 642
 - strike volume, 147
 - unemployment and city size, 43
 - value at risk, 203
 - wage inequality, 41
- average treatment effects, 639
- backfitting, 283
- bandwidth, *see*
 - smoothing parameter
- big $O(\cdot)$, 684
- big $O_p(\cdot)$, 685
- bootstrap, 360, 365
 - block, 563
 - i.i.d., 378
 - number of replications, 360
 - stationary, 558
 - wild, 289, 308, 357
- Borel-Cantelli lemma, 689
- Borel measurable function, 668
- Borel measurable set, 667
- boundary correction, 80
- boundary effects, 30
- Brownian motion, 678
- censored model
 - nonparametric, 343, 345, 346
 - parametric, 332
 - semiparametric, 335, 337
- central limit theorem
 - Degenerate U-statistics, 692
 - Hilbert-valued, 694
 - Liapunov, 689
 - Lindeberg-Feller, 688
 - Lindeberg-Levy, 688
- central limit theorem (CLT), 23
- characteristic function, 671
- cointegration, 564
- confusion matrix, 279
- convergence
 - almost everywhere, 682
 - almost surely, 682
 - in r^{th} mean, 682
 - in distribution, 682
 - in probability, 682
 - weak, 687
- copula, 651

- Cramer-Wold theorem, 689
cumulative distribution function (CDF), 3, 7
 cross-validation, 23
 frequency, 7
 nonsmooth, 182
 smooth, 20, 184
curse of dimensionality, xvii
- density estimation
 least squares cross-validation bandwidth selection, 15, 27
 likelihood cross-validation bandwidth selection, 18, 28
 plug-in bandwidth selection, 14, 26
 rule-of-thumb bandwidth selection, 14, 26
- Dirac delta function, 679
- empirical distribution function, 19
- fixed effects, 586
Fourier series, 512
frequency method, 6, 115
- Gaussian process, 678
generalized method of moments (GMM), 512
- hazard function, 198
Hilbert space, 674
hypothesis testing
 conditional parametric density function, 402
 conditional parametric distributions, 382
 correct parametric functional form, 355, 365, 398
 equality of density functions, 362, 401
 independence, 378
 omitted variables, 370
 parametric density function, 380
 parametric single index model, 369
 serial dependence, 404
 significance, 375
 significance test, 401
- inequality
 Cauchy, 690
 Cauchy-Schwarz, *see* Cauchy
 Chebychev, 690
 Hölder, 690
 Markov, 689
 triangle, 481
- instrumental variable, 506
integrated mean squared error (IMSE), 13
integrated square error (ISE), 157
- Kaplan-Meier estimator, 338
- kernel
 Aitchison and Aitken, 167
 Bartlett, 405
 convolution, 16
 Daniell, 405
 Epanechnikov, 35
 Gaussian, 34
 higher order, 33
 Parzen, 405
 triangular, 400
 uniform, 8
- Khinchin's law of large numbers, 688
- knots, 446
Kullback-Leibler, 382
- latent variable, 316
law of iterated expectations, 690
Lebesgue-Stieltjes integral, 670
Lebesgue measure, 666
link function, 250, 295, 463
Lipschitz function, 672
local average, 64
local constant estimator, 60
 AIC_c bandwidth selection, 72

- irrelevant regressors and bandwidth selection, 73
- least squares cross-validation, 69
- plug-in bandwidths, 66
- rule-of-thumb bandwidths, 66
- local linear estimator, 79
 - least squares cross-validation, 83
- local polynomial estimator, 85
- location-scale model, 346

- maximum likelihood estimation, 4
- mean squared error (MSE), 6
- measure, 666
- measurement error, 92
- minimax, 532
- MINPIN, 230
- mixing, 535
 - α -mixing, 535
 - β -mixing, 535
 - ϕ -mixing, 535
 - ρ -mixing, 535
 - Mixingale, 536
 - strong, *see* α -mixing

- naïve kernel estimator, 8
- Nadaraya-Watson estimator, *see* local constant estimator
- nearest neighbor, 416
- neural network, 547
- nonlinear-differencing, 606, 614
- nonstationary data, 566
- normal rule-of-thumb, 14

- oracle estimator, 287
- orthonormal basis, 674

- panel data, 575
- Parseval's equality, 675
- partially linear model, 222
- partial derivative estimator, 80
- pilot bandwidth, 14
- Pitman local alternatives, 400
- poolable, 578
- power series, 512
- probability density function (PDF), 3

- product-limit estimator, *see* Kaplan-Meier estimator
- product kernel function, 24
 - discrete data, 126
 - mixed data, 137

- quantile regression, 189

- random effects, 578
- Riemann-Stieltjes integral, 669
- Riemann integral, 668
- Rosenblatt-Parzen estimator, 9

- selectivity model
 - parametric, 316
 - semiparametric, 317, 318, 320
- semiparametric efficiency bound, 234, 267
- sieves, 610
- sigma-field, 664
- single index model, 249
- small $o(\cdot)$, 684
- small $o_p(\cdot)$, 685
- smoothing parameter, 8
- smooth coefficient model, 301
- Sobolev norm, 675
- spectrum, 404
- spline, 512
- spline function, 446
- stochastic equicontinuity, 686
- survival function, 198

- time-differencing, 606, 607
- Tobit
 - type-2, 316
 - type-3, 320
- transformation model, 659
- trimming, 254, 256, 260, 266, 359

- U-statistic, 691
- U-statistic H-decomposition, 691

- wavelet, 428
- weakly dependent, 535
- weakly exogenous, 506
- weighted integrated mean squared error (WIMSE), 67
- window width, *see* smoothing parameter