

# Preface

This is Volume I of a proposed two volume text on the theory and practice of maximum likelihood estimation. It is intended for graduate students in statistics; applied, industrial, and engineering mathematics; and operations research, as well as for researchers and practitioners in the field. The present volume treats mostly nonparametric density estimation, whereas Volume II is concerned with indirect (or inverse) estimation problems. The material is divided into an introductory chapter and six parts:

- Parametric density estimation,
- Nonparametric density estimation,
- Convexity and optimization,
- Nonparametric least squares (well-posed and ill-posed),
- Generalized deconvolution problems with random sampling,
- Expectation-Maximization algorithms.

The first three parts constitute Volume I, and the others form Volume II. Each part consists of a number of theoretical chapters, in which the maximum (penalized) likelihood estimators for the problems under discussion are introduced and their asymptotic behavior thoroughly analyzed, as well as an “in action” chapter, in which computational issues are discussed and the small sample behavior of the estimators is demonstrated, using simulated and real data. The material is liberally sprinkled with exercises that cover modifications of the results presented and fill in missing details of the mathematical development, but sometimes deal with open problems. Computational projects are also indicated. We briefly discuss each part.

## *Parametric density estimation*

Here, we treat the standard asymptotic theory (consistency, best asymptotic normality) of parametric maximum likelihood estimators in the regular case. Computational issues, such as Newton’s method, the method of scores, and Expectation-Maximization (EM) algorithms are discussed. Additional topics include robustness, ridge regression, skewed heavy-tailed

densities (log-normal, Gamma, Weibull), and mixtures of normals. Simulation studies illustrating the small sample behavior of maximum likelihood estimators and the effects of misspecification of the parametric model are given in a separate chapter.

### *Nonparametric density estimation*

Under the usual nonparametric assumptions, we prove the almost sure bound of  $\mathcal{O}(n^{-2/5})$  on the  $L^1$  error of kernel density estimators, as well as the (best) asymptotic normality of an estimator of the entropy based on kernel estimators. The principal tools, discrete parameter submartingales and Devroye's exponential inequalities, are discussed in detail, but at an elementary level. The optimality of kernels is discussed (and later in the convexity part, optimal kernels of high order are computed).

In a chapter on maximum penalized likelihood estimation, we prove the a.s. bound of  $\mathcal{O}(n^{-2/5})$  for the mple using Good's first roughness penalization functional (the Good estimator). We also consider the roughness penalization of the log-density. Maximum smoothed likelihood estimation for log-concave densities is discussed briefly, as is minimum (smoothed) distance estimation.

In the chapter on monotone and unimodal densities, we analyze in detail the pool-adjacent-violators algorithm for computing the Grenander estimator, the solution to the monotone maximum likelihood estimation problem. We show that the kernel-estimator-made-monotone always has smaller error than the kernel estimator itself, for any reasonable measure of the global error (all  $L^p$  norms, Kullback-Leibler, Hellinger), and that the Grenander operator is a contraction in this sense. We also consider unimodal density estimation, with just about the same conclusion.

We discuss smoothing parameter selection (mostly) from the  $L^1$  perspective: Devroye's double kernel method and variations, the various plug-in methods, as well as a pilotless version of the Hall-Wand estimator. For some of these methods, we prove that the selected smoothing parameter  $H$  satisfies  $H \asymp n^{-1/5}$  almost surely. A discrepancy principle is discussed for kernel estimation and for the Good estimator.

Simulation studies comparing the various estimators are given in a separate chapter. We also compare various kernels with themselves and with the Good estimator. (It turns out that the GOOD estimator is remarkably GOOD!)

### *Convexity and optimization*

All of the estimators discussed, except the kernel estimator, are defined as solutions to convex minimization problems. In this part, we give a self-contained treatment of such problems, in particular the existence of the estimators is established. We also lay the groundwork for ill-posed indirect

estimation and for the convergence of EM and EM-like algorithms for such problems. Various inequalities used in other parts of the text, that derive from convexity, are either proved or stated as exercises.

As the above description makes clear, no attempt was made to be encyclopedic. The problems and estimators considered, and the general approach to the questions involved, are mostly determined by the interests and backgrounds of the authors and their friends. Some alternatives (with references) are briefly mentioned when appropriate, and the interested reader should follow these up, despite our editorial comments.

### *Why a new text?*

Why a new text on statistical estimation, and why did it take on the present form? Our interest in the field started in 1992 with maximum (smoothed) likelihood estimation for indirect estimation problems, more specifically, for nonparametric deconvolution based on independent, identically distributed data. At that time, the literature on this topic was limited, and did not appeal to us. Thus, the idea of writing a monograph on indirect estimation was born. It soon became clear that maximum penalized likelihood estimation (mple) for indirect problems is quite hard and that it was perhaps prudent to start with plain nonparametric density estimation. The standard mple method involved the roughness penalization proposed by GOOD (1971), but it turned out to have been in disuse. The method of choice was that of kernel estimation, so *that* had to be included. Here, there was already a huge and growing literature, of which we liked the often unquoted work of DEVROYE and coauthors, in particular, the seminal text, DEVROYE and GYÖRFI (1985). While attending the 1997 Symposium on Nonparametric Function Estimation in Montreal, the idea to include a chapter on maximum likelihood estimation for monotone and unimodal densities was born. With general nonparametric density estimation included, it seemed reasonable to include the general parametric case also. As far as mathematical background was concerned, a good deal of indirect estimation is naturally explained in the context of convex minimization problems. Somewhat surprisingly, even in the context of parametrics and (sub)martingales, convexity arguments are frequently and freely used. So it was decided to include a relatively elementary treatment of convexity and convex minimization problems.

During the preparation of this text it became clear that the main prerequisite for the text is an introductory course on finite-dimensional vector spaces (introductory functional analysis would be helpful, but is not necessary) and an acquaintance with probability theory, including the law of the iterated logarithm (no measure theory is required). Parts of the text have been used in various classes and seminars. Professors David Mason (University of Delaware) and Uwe Einmahl (Vrije Universiteit Brussel) have used

Chapter 2 in their courses. We have used the chapters on nonparametric density estimation and convexity at the University of Delaware. Early on, much of the material was tried out on students during informal seminars. We thank Chris Venaccio, Tim Loomer, Joe Collins, André Acosta, and Carmelita Perlitz for their patience. Paul Deheuvels, Luc Devroye, Alexander Goldenshluger, David Mason, and Andrei Zaitsev have read all or parts of the manuscript, and we thank them for their comments.

### *Acknowledgments*

We started this project as members of the DEPARTMENT OF MATHEMATICAL SCIENCES and finished it as members of FOOD AND RESOURCE ECONOMICS. It is our pleasant duty to thank John Nye, Dean of the College of Agriculture, and Bobby Gempesaw, then Chair of Food and Resource Economics (now Associate Provost) for their vision and (repeated and ultimately successful) efforts to rescue the statistics program at the University of Delaware, and to thank the members of Food and Resource Economics for the welcome they extended to us.

As with any intellectual endeavor, we were influenced by many people, from our teachers to anonymous referees, and we thank them all. However, four individuals must be explicitly mentioned. First of all, we wish to thank Zuhair Nashed of our old department for listening and helping us with both mathematical and departmental issues: In our new department, we can only say we miss him. Next, we wish to thank Paul Deheuvels for his interest and encouragement in our project, especially very early on, and for his painstaking review of the next-to-last version of the manuscript. A special thanks goes to Luc Devroye for his many suggestions and interest in the text and for urging us on to finish, but more importantly, for the many discussions at the Crab Trap, which helped shape our basic view of the field. Finally, it is our privilege to thank David Mason. His nitpicking critique is (now) greatly appreciated, as are some suggestions that helped shape the text. We must also thank him for single-handedly keeping the Statistics and Probability Seminar alive and for giving it such colorful names as the *Industrial Applied Statistics Seminar* or the *Probability, Informatics and Statistics Seminar*. Without his efforts, it is doubtful whether there would have been anything left to move to our new home.

Newark, Delaware  
April 13, 2001

Paul Eggermont, Vince LaRiccia

# Contents of Volume I

<b>Preface</b>	vii
<b>Notations, Acronyms, and Conventions</b>	xv
<b>1. Parametric and Nonparametric Estimation</b>	1
1. Introduction	1
2. Indirect problems, EM algorithms, kernel density estimation, and roughness penalization	8
3. Consistency of nonparametric estimators	13
4. The usual nonparametric assumptions	18
5. Parametric vs nonparametric Rates	20
6. Sieves and convexity	22
7. Additional notes and comments	26
<b>PART I: PARAMETRIC ESTIMATION</b>	
<b>2. Parametric Maximum Likelihood Estimation</b>	29
1. Introduction	29
2. Optimality of maximum likelihood estimators	37
3. Computing maximum likelihood estimators	49
4. The EM algorithm	53
5. Sensitivity to errors: M-estimators	63
6. Ridge regression	75
7. Right-skewed distributions with heavy tails	80
8. Additional comments	88
<b>3. Parametric Maximum Likelihood Estimation in Action</b>	91
1. Introduction	91
2. Best asymptotically normal estimators and small sample behavior	92
3. Mixtures of normals	96
4. Computing with the log-normal distribution	101
5. On choosing parametric families of distributions	104
6. Toward nonparametrics: mixtures revisited	113

## PART II : NONPARAMETRIC ESTIMATION

<b>4. Kernel Density Estimation</b>	119
1. Introduction	119
2. The expected $L^1$ error in kernel density estimation	130
3. Integration by parts tricks	136
4. Submartingales, exponential inequalities, and almost sure bounds for the $L^1$ error	139
5. Almost sure bounds for everything else	151
6. Nonparametric estimation of entropy	159
7. Optimal kernels	167
8. Asymptotic normality of the $L^1$ error	173
9. Additional comments	186
<b>5. Nonparametric Maximum Penalized Likelihood Estimation</b>	187
1. Introduction	187
2. Good's roughness penalization of root-densities	189
3. Roughness penalization of log-densities	202
4. Roughness penalization of bounded log-densities	207
5. Estimation under constraints	213
6. Additional notes and comments	218
<b>6. Monotone and Unimodal Densities</b>	221
1. Introduction	221
2. Monotone density estimation	225
3. Estimating smooth monotone densities	232
4. Algorithms and contractivity	234
5. Contractivity : the general case	244
6. Estimating smooth unimodal densities	250
7. Other unimodal density estimators	262
8. Afterthoughts : convex densities	265
9. Additional notes and comments	267
<b>7. Choosing the Smoothing Parameter</b>	271
1. Introduction	271
2. Least-squares cross-validation and plug-in methods	276
3. The double kernel method	283
4. Asymptotic plug-in methods	295
5. Away with pilot estimators !?	299
6. A discrepancy principle	306
7. The Good estimator	309
8. Additional notes and comments	316
<b>8. Nonparametric Density Estimation in Action</b>	319
1. Introduction	319
2. Finite-dimensional approximations	320
3. Smoothing parameter selection	323
4. Two data sets	329

5. Kernel selection	333
6. Unimodal density estimation	338
PART III: CONVEXITY	
<b>9. Convex Optimization in Finite- Dimensional Spaces</b>	347
1. Convex sets and convex functions	347
2. Convex minimization problems	357
3. Lagrange multipliers	361
4. Strict and strong convexity	370
5. Compactness arguments	373
6. Additional notes and comments	375
<b>10. Convex Optimization in Infinite- Dimensional Spaces</b>	377
1. Convex functions	377
2. Convex integrals	383
3. Strong convexity	387
4. Compactness arguments	390
5. Euler equations	395
6. Finitely many constraints	398
7. Additional notes and comments	402
<b>11. Convexity in Action</b>	405
1. Introduction	405
2. Optimal kernels	405
3. Direct nonparametric maximum roughness penalized likelihood density estimation	412
4. Existence of roughness penalized log-densities	417
5. Existence of log-concave estimators	423
6. Constrained minimum distance estimation	425
APPENDICES	
<b>1. Some Data Sets</b>	433
1. Introduction	433
2. The Old Faithful geyser data	433
3. The Buffalo snow fall data	434
4. The rubber abrasion data	434
5. Cloud seeding data	434
6. Texo oil field data	435
<b>2. The Fourier Transform</b>	437
1. Introduction	437
2. Smooth functions	437
3. Integrable functions	441
4. Square integrable functions	443
5. Some examples	446
6. The Wiener theorem for $L^1(\mathbb{R})$	456

<b>3. Banach Spaces, Dual Spaces, and Compactness</b>	459
1. Banach spaces	459
2. Bounded linear operators	462
3. Compact operators	463
4. Dual spaces	469
5. Hilbert spaces	472
6. Compact Hermitian operators	478
7. Reproducing kernel Hilbert spaces	484
8. Closing comments	485
<b>References</b>	487
<b>Author Index</b>	499
<b>Subject Index</b>	505



# Notations, Acronyms, and Conventions

The numbering and referencing conventions are as follows. Items of importance, such as formulas, theorems, and exercises, are labeled as  $(Y.X)$ , with  $Y$  the current section number and  $X$  the current (consecutive) item number. The exceptions are tables and figures, which are independently labeled following the same system. A reference to Item  $(Y.X)$  is to the item with number  $X$  in section  $Y$  of the current chapter. References to items outside the current chapter take the form  $(Z.Y.X)$ , with  $Z$  the chapter number and  $X$  and  $Y$  as before. References to the literature take the standard form AUTHOR (YEAR) or AUTHOR#1 and AUTHOR#2 (YEAR), and so on. The references are arranged in alphabetical order by the first author, and by year.

We tried to limit our use of acronyms to some very standard ones, as in the following list.

iid	independent, identically distributed.
rv	random variable.
m(p)le	maximum (penalized) likelihood estimation (or estimator).
pdf	probability density function.
cdf	(cumulative) distribution function.
EM	Expectation–Maximization.
a.s.	almost surely or, equivalently, with probability 1.

Some of the standard notations throughout the text are as follows.

$\mathbb{1}(x \in A)$	The indicator function of the set $A$ .
$\mathbb{1}_A(x)$	Also the indicator function of the set $A$ .
$\mathbb{1}(X \leq x)$	The indicator function of the event $\{X \leq x\}$ .
$(x)_+$	The maximum of 0 and $x$ ( $x$ a real-valued expression).
$\asymp, \asymp_{\text{as}}$	Asymptotic equivalence, and the almost sure version. See Definition (1.3.6).
$=_{\text{as}}$	Almost sure equality.
$\leq_{\text{as}}$	Almost surely less than or equal. Likewise for $\geq_{\text{as}}$ .

$\preceq$	For symmetric matrices $A, B \in \mathbb{R}^{n \times n}$ , we write $A \preceq B$ if $C = B - A$ is semi-positive definite, i.e., $x^T C x \geq 0$ for all $x \in \mathbb{R}^n$ .
$\mathcal{O}_P, o_P$	See Definition (1.3.4), and (1.3.5).
Opie-One	An Opie-One estimator $\theta_n$ for $\theta_o$ is such that $\sqrt{n}(\theta_n - \theta_o) = \mathcal{O}_P(1).$
$\int_{\mathbb{R}} f dG$ $\varphi * \psi$	This is shorthand for $\int_{\mathbb{R}} f(x) dG(x)$ . Similarly for $\int_{\mathbb{R}} f g$ . The convolution of two functions, defined as $\varphi * \psi(x) = \int_{\mathbb{R}} \varphi(x - y) \psi(y) dy, \quad x \in \mathbb{R}.$
$\varphi * d\Psi$	The convolution of a function $\varphi$ and a distribution $\Psi$ , defined as $\varphi * d\Psi(x) = \int_{\mathbb{R}} \varphi(x - y) d\Psi(y), \quad x \in \mathbb{R}.$
$F_n$ $A_h$	The empirical distribution function of the data $X_1, \dots, X_n$ . Typically, $A_h(x) = h^{-1} A(h^{-1}x)$ , and similarly for other capital symbols. For lowercase symbols, this does not usually apply.
$(A^m)_h$	Following the previous convention, $(A^m)_h(x) = h^{-1} \{ A(h^{-1}x) \}^m.$
$A_h * dF_n$	The kernel estimator with kernel $A$ , written as the convolution of $A_h$ and the empirical distribution function.
$A_h dF_n$	A boundary kernel estimator; so it is not quite a convolution of a kernel $A$ and the empirical distribution function.
$f_h$	This is typically the large sample asymptotic estimator under consideration. For kernel estimation, it is $A_h * f$ or $A_h * f_o$ , where $f$ or $f_o$ is the “true” density.
$f^{nh}$	The estimator based on $X_1, X_2, \dots, X_n$ . In kernel estimation, it equals $A_h * dF_n$ . But it also denotes the maximum penalized likelihood estimator, the GOOD (1971) mple, the log-penalization, and the monotone and unimodal estimator. The context <i>should</i> make it clear.
$g_\kappa$	The kernel obtained with fractional integration by parts. See § 4.3.
$T$	The one-sided exponential kernel or pdf, which is given by $T(x) = \exp(- x ) \mathbb{1}(x \geq 0)$ . But also the histogram operator, see § 8.2, in particular, (8.2.3).
$\mathfrak{B}$	The two-sided exponential kernel, $\mathfrak{B}(x) = \frac{1}{2} \exp(- x )$ .
$\phi$	The normal kernel or density $\phi(x) = (2\pi)^{-1/2} \exp(-\frac{1}{2}x^2)$ .
$\phi_\sigma$	The scaled normal $\phi_\sigma(x) = \sigma^{-1} \phi(\sigma^{-1}x)$ . Sometimes we also use :

$\phi(\cdot; \mu, \sigma)$	$\phi(x; \mu, \sigma) = \phi_\sigma(x - \mu)$ .
$\ \cdot\ _p$	The $L^p(\Omega)$ norm, $1 \leq p \leq \infty$ . See (1.3.8).
$L^p(\Omega)$	Space of (equivalence classes of measurable) functions $f$ on $\Omega \subset \mathbb{R}$ with $\ f\ _p < \infty$ .
$W^{m,p}(\Omega)$	Sobolev space of all functions in $L^p(\Omega)$ with $m$ -th derivative in $L^p(\Omega)$ also.
$H(\varphi, \psi)$	The Hellinger distance. See (1.3.18).
$KL(\varphi, \psi)$	The Kullback-Leibler distance. See (1.3.19).
$PHI(\varphi, \psi)$	The Pearson's $\varphi^2$ distance. See (1.3.20).
$a \vee b$	The function with values $[a \vee b](x) = \max(a(x), b(x))$ .
$a \wedge b$	The function with values $[a \wedge b](x) = \min(a(x), b(x))$ .
$LCM(F)$	The least concave majorant of the distribution $F$ .
$lcm(f)$	The function, continuous from the left, which is equal almost everywhere to the derivative of $LCM(F)$ , where $F$ is the distribution corresponding to the density $f$ .
$\mathbb{R}_+, \mathbb{R}_{++}$	$\mathbb{R}_{++} = (0, \infty)$ (0 not included), $\mathbb{R}_+ = \mathbb{R}_{++} \cup \{0\}$ (somewhat pedantic, but ...).
$\delta f(x; h)$	The Gateaux variation of $f$ at $x$ in the direction $h$ . For differentiable functions, it is just $h f'(x)$ . See § 10.1.
$\text{var}_n(A; h)$	See (7.3.27).
$\text{var}_o(A; h)$	See (7.4.3).
$\text{var}_o(A; h; x)$	See (7.4.11).

Maximum Penalized Likelihood Estimation

Volume I: Density Estimation

Eggermont, P.P.B.; LaRiccia, V.N.

2001, XVIII, 512 p., Hardcover

ISBN: 978-0-387-95268-0