

Studies in Big Data

Volume 37

Series editor

Janusz Kacprzyk, Polish Academy of Sciences, Warsaw, Poland
e-mail: kacprzyk@ibspan.waw.pl

The series “Studies in Big Data” (SBD) publishes new developments and advances in the various areas of Big Data- quickly and with a high quality. The intent is to cover the theory, research, development, and applications of Big Data, as embedded in the fields of engineering, computer science, physics, economics and life sciences. The books of the series refer to the analysis and understanding of large, complex, and/or distributed data sets generated from recent digital sources coming from sensors or other physical instruments as well as simulations, crowd sourcing, social networks or other internet transactions, such as emails or video click streams and other. The series contains monographs, lecture notes and edited volumes in Big Data spanning the areas of computational intelligence incl. neural networks, evolutionary computation, soft computing, fuzzy systems, as well as artificial intelligence, data mining, modern statistics and Operations research, as well as self-organizing systems. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution, which enable both wide and rapid dissemination of research output.

More information about this series at <http://www.springer.com/series/11970>

Artur Gramacki

Nonparametric Kernel Density Estimation and Its Computational Aspects

 Springer

Artur Gramacki
Institute of Control and Computation
Engineering
University of Zielona Góra
Zielona Góra
Poland

ISSN 2197-6503

ISSN 2197-6511 (electronic)

Studies in Big Data

ISBN 978-3-319-71687-9

ISBN 978-3-319-71688-6 (eBook)

<https://doi.org/10.1007/978-3-319-71688-6>

Library of Congress Control Number: 2017959157

© Springer International Publishing AG 2018

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature

The registered company is Springer International Publishing AG

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my brother Jarek

Foreword

In the rapidly developing field of processing and analyzing big data, new challenges come not only from a vast volume of data but also from their smoothing. There are many smoothing techniques known, but it seems that one of the most important ones is kernel smoothing. It belongs to a general class of nonparametric estimation of functions such as probability density functions or regression ones. Kernel-based estimators of such functions are very attractive for practitioners as they can uncover important patterns in the data while filtering noise and ignoring irrelevant details. Subsequently, the estimated function is smooth, and the level of smoothness can be typically controlled by a parameter known as a bandwidth.

The subject matter of this book is primarily kernel probability density function estimation. During the last few decades, intensive research has been conducted in this area, and it seems that from a theoretical point of view, nonparametric kernel density estimation has reached its maturity. Meanwhile, relatively less research has been devoted to computational problems regarding kernel density estimation and optimal bandwidth selection. These problems are very important in the context of the need of analyzing big datasets, both uni- and multidimensional.

A part of the book focuses on fundamental issues related to nonparametric density estimation to give the readers intuition and basic mathematical skills required to understand kernel smoothing. This part of the book is of tutorial value and can be perceived as a good starting point for readers unfamiliar with nonparametric techniques. The book is also meant for a more advanced audience, interested in recent developments related to very fast and accurate kernel density estimation as well as bandwidth selection.

A unified framework based on the fast Fourier transform is presented. Abundant experimental results included in the book confirm its practical usability as well as very good performance and accuracy. Additionally, some original preliminary research on using modern FPGA chips for bandwidth selection is presented. All the concepts described in this book are richly illustrated with both academic examples and real datasets. Also, a special chapter is devoted to interesting and important examples of practical usage of kernel-based smoothing techniques.

Concluding, the book can be strongly recommended to researchers and practitioners, both new and experienced not only in the field of data smoothing but also in wide understanding of processing and analyzing big data.

Zielona Góra, Poland
September 2017

Józef Korbicz
Professor,
Corresponding member of the Polish
Academy of Sciences

Preface

This book concerns the problem of data smoothing. There are many smoothing techniques, yet the kernel smoothing seems to be one of the most important and widely used ones. In this book, I focus on a well-known technique called kernel density estimation (KDE), which is an example of a nonparametric approach to data analysis.

During the last few decades, many books and papers devoted to this broad field have been published, so it seems that this area of knowledge is quite well understood and reached its maturity point. However, many (or even most) of the practical algorithms and solutions designed in the context of KDE are very time-consuming with quadratic computational complexity being a commonplace. This might be not problematic for situations, where datasets are not that big (at the level of hundreds of individual data points) but it already can be an obstacle for datasets containing thousands or more individual data points, especially in case of multivariate data. Progress in terms of theoretical results related to KDE does not go hand in hand with the development of fast and accurate algorithm for speeding up the required calculations in practical terms. In this sense, this book can be considered a valuable contribution to the field of KDE.

This book is a result of my research in the area of numerical and computational problems related to KDE, an interest that has been developing since ca. 2010. It should be viewed primarily as a research monograph and is intended both for those new to such topics as well as for more experienced readers. The first few chapters present a background material, describing the fundamental concepts related to the nonparametric density estimation, kernel density estimation, and bandwidth selection methods. The presented material is richly illustrated by numerical examples, using both toy and real datasets. The following chapters are devoted to the presentation of our own research on fast computation of kernel density estimators and bandwidth selection. The developed methods are based on the fast Fourier transform (FFT) algorithm that relies on a preliminary data transformation known as data binning. Certain results obtained by me on utilizing field-programmable gate arrays (FPGA) in the context of fast bandwidth selection are also included. FPGA devices are a not so common choice in terms of implementing purely numerical algorithms.

The proposed implementation can be seen as a preliminary study of practical usability of such FPGA-based applications. The monograph ends with a chapter presenting a number of applications related to KDE. The following example applications are given: discriminant analysis, cluster analysis, kernel regression, multivariate statistical process control, and flow cytometry.

I wish to express my deepest gratitude to Prof. Józef Korbicz, Ph.D., D.Sc., a corresponding member of the Polish Academy of Sciences for his continuing support since 2000 to this day and for motivating me to work hard on the problems at hand. I would also like to express my sincere thanks to Prof. Andrzej Obuchowicz, Ph.D., D.Sc., for his help at a time of particular need in my life. I also extend my thanks to Prof. Dariusz Uciński, Ph.D., D.Sc., and to Marcin Mrugalski, Ph.D., D.Sc., for their continuing kindness. Marek Sawerwain, Ph.D., helped me a lot in all programming tasks related to FPGA, and I thank him very much for this. My warm thanks also go to my mother, to my wife Edyta, and to my children Ola and Kuba for their love, patience, and support. Finally, I would like to express my sincere gratitude to my brother Jarosław because I truly believe that writing this book would not have been possible without his powerful and continuous assistance and support, both in professional and in family life.

Zielona Góra, Poland
September 2017

Artur Gramacki

About this Book

This book describes computational problems related to kernel density estimation (KDE)—one of the most important and widely used data smoothing techniques. A very detailed description of novel FFT-based algorithms for both KDE computations and bandwidth selection is presented.

The theory of KDE appears to have matured and is now well developed and understood. However, there is not much progress observed in terms of performance improvements. This book is an attempt to remedy this.

The book primarily addresses researchers and advanced graduate or postgraduate students who are interested in KDE and its computational aspects. The book contains both some background and much more sophisticated material, hence also more experienced researchers in the KDE area may find it interesting.

The presented material is richly illustrated with many numerical examples using both artificial and real datasets. Also, a number of practical applications related to KDE are presented.

Contents

1	Introduction	1
1.1	Background	1
1.2	Contents	3
1.3	Topics Not Covered in This Book	5
2	Nonparametric Density Estimation	7
2.1	Introduction	7
2.2	Density Estimation and Histograms	7
2.3	Smoothing Histograms	10
2.4	A General Formulation of Nonparametric Density Estimation	12
2.5	Parzen Windows	14
2.6	k -nearest Neighbors	18
3	Kernel Density Estimation	25
3.1	Introduction	25
3.2	Univariate Kernels	26
3.3	Definitions	27
	3.3.1 Univariate Case	27
	3.3.2 Multivariate Case	35
3.4	Performance Criteria	42
	3.4.1 Univariate Case	42
	3.4.2 Multivariate Case	48
3.5	Adaptive Kernel Density Estimation	50
3.6	Kernel Density Estimation with Boundary Correction	54
3.7	Kernel Cumulative Distribution Function Estimation	56
3.8	Kernel Density Derivative Estimation	58
3.9	The Curse of Dimensionality	59
3.10	Computational Aspects	61

- 4 Bandwidth Selectors for Kernel Density Estimation 63**
- 4.1 Introduction 63
- 4.2 Univariate Bandwidth Selectors 64
 - 4.2.1 Univariate Rule-of-Thumb Selectors 64
 - 4.2.2 Univariate Plug-In Selectors 65
 - 4.2.3 Univariate Least Squares Cross Validation Selector 67
 - 4.2.4 Univariate Smoothed Cross Validation Selector 72
 - 4.2.5 Other Selectors 73
- 4.3 Multivariate Bandwidth Selectors 73
 - 4.3.1 Multivariate Rule-of-Thumb Selectors 74
 - 4.3.2 Multivariate Least Squares Cross Validation Selector 75
 - 4.3.3 Notation for Higher-Order Derivatives 76
 - 4.3.4 Multivariate Plug-In Selector 78
 - 4.3.5 Multivariate Smoothed Cross Validation Selector 79
 - 4.3.6 Simulations for a Bivariate Target Densities 80
- 4.4 Computational Aspects 80
- 5 FFT-Based Algorithms for Kernel Density Estimation and Bandwidth Selection 85**
- 5.1 Introduction 85
- 5.2 Data Binning 85
 - 5.2.1 Univariate Case 86
 - 5.2.2 Multivariate Case 88
- 5.3 The FFT-Based Algorithm for Density Estimation 90
 - 5.3.1 Introductory Notes and a Problem Demonstration 90
 - 5.3.2 Univariate Case 93
 - 5.3.3 Multivariate Case 100
- 5.4 The FFT-Based Algorithm for Bandwidth Selection 102
 - 5.4.1 Introductory Notes 102
 - 5.4.2 The Algorithm for the LSCV Selector 104
 - 5.4.3 Notes on the Algorithms for the PI and SCV Selectors 106
- 5.5 Experimental Results for Kernel Density Estimation 108
- 5.6 Experimental Results for Bandwidth Selection 110
 - 5.6.1 Accuracy, Synthetic Data 111
 - 5.6.2 Accuracy, Real Data 112
 - 5.6.3 Speed Comparisons 115
- 5.7 Concluding Remarks 118
- 6 FPGA-Based Implementation of a Bandwidth Selection Algorithm 119**
- 6.1 Introduction 119
- 6.2 High Level Synthesis 120
- 6.3 The Method Implemented and Data Preprocessing 121

- 6.4 FPGA-Based Implementation 124
 - 6.4.1 Implementation Preliminaries 124
 - 6.4.2 Implementation Details 126
 - 6.4.3 Results 127
- 6.5 Concluding Remarks 130
- 7 Selected Applications Related to Kernel Density Estimation 133**
 - 7.1 Introduction 133
 - 7.2 Kernel Discriminant Analysis 134
 - 7.3 Kernel Cluster Analysis 137
 - 7.4 Kernel Regression 141
 - 7.5 Multivariate Statistical Process Control 143
 - 7.6 Flow Cytometry 149
 - 7.6.1 Introduction 149
 - 7.6.2 Feature Significance 151
 - 7.6.3 A Gating Method 151
- 8 Conclusion and Further Research Directions 159**
- Bibliography 163**
- Index 173**

About the Author

Artur Gramacki is an Assistant Professor at the Institute of Control and Computation Engineering of the University of Zielona Góra, Poland. His main interests cover general exploratory data analysis, while recently he has focused on parametric and nonparametric statistics as well as kernel density estimation, especially its computational aspects. In his career, he has also been involved in many projects related to the design and implementation of commercial database systems, mainly using Oracle RDBMS. He is a keen supporter of the R Project for Statistical Computing, which he tries to use both in his research and teaching activities.

Abbreviations

AMISE	Asymptotic MISE
AMSE	Asymptotic Mean Squared Error
ASH	Averaged Shifted Histogram
BCV	Biased Cross-Validation
CDF	Cumulative Distribution Function
CV	Cross-Validation
DA	Discriminant Analysis
FFT	Fast Fourier Transform
FPGA	Field-Programmable Gate Arrays
GPU	Graphics Processing Units
HDL	Hardware Description Languages
HLS	High-Level Synthesis
ISB	Integrated Squared Bias
ISE	Integrated Squared Error
KCDE	Kernel CDF Estimation (or Estimator depending of a context)
KDA	Kernel Discriminant Analysis
KDE	Kernel Density Estimation (or Estimator depending of a context)
KNN	K-Nearest Neighbors
KNR	Kernel Nonparametric Regression
LSCV	Least-Squares Cross-Validation
MIAE	Mean Integrated Absolute Error
MISE	Mean Integrated Squared Error
MPI	Message Passing Interface
MS	Maximal Smoothing
MSE	Mean Squared Error
NS	Normal Scale
PI	Plug-in

ROT	Rules-of-Thumb
RTL	Register-Transfer Level
SCV	Smoothed Cross-Validation
UCV	Unbiased Cross-Validation

Notation

\otimes	Kronecker product operator
\odot	Element-wise multiplication
$\mathbb{1}_{\{x \in A\}}$	Indicator function, that is $\mathbb{1}_{\{x \in A\}} = 1$ for $x \in A$ and $\mathbb{1}_{\{x \in A\}} = 0$ for $x \notin A$
\mathbf{A}	$d \times d$ matrix
\mathbf{a}	Vector of size d
c_i	Univariate binning grid counts
\mathbf{c}_i	Multivariate binning grid counts
Df	First derivative (gradient) of f
$D^{\otimes r}$	r -th Kronecker power of the operator \mathbf{D}
d	Problem dimensionality
F	Unknown cumulative distribution function
\hat{F}	Kernel cumulative distribution function estimate
\mathcal{F}	Fourier transform operator
\mathcal{F}^{-1}	Inverse Fourier transform operator
f	Unknown density function
$f^{(r)}$	Unknown density derivative function of order r
$\hat{f}^{(r)}$	Kernel density function estimate
\tilde{f}	Kernel density function estimate after binning employed
$\hat{f}^{(r)}$	Kernel density derivative function estimate of order r
$\hat{f}(x; h)$	Univariate kernel density estimate at point x and with bandwidth scalar h
$\hat{f}_{-i}(x; h)$	Univariate leave-one-out kernel density estimate at point x and with bandwidth scalar h
$\hat{f}(\mathbf{x}; \mathbf{H})$	Multivariate kernel density estimate at point \mathbf{x} and with bandwidth matrix \mathbf{H}
$\hat{f}_{-i}(\mathbf{x}; \mathbf{H})$	Multivariate leave-one-out kernel density estimate at point \mathbf{x} and with bandwidth matrix \mathbf{H}

$f_1 * f_2$	Convolution of functions f_1 and f_2 , that is $f_1 * f_2(x) = \int f_1(u)f_2(x - u) du$
G	$d \times d$ pilot bandwidth matrix which is symmetric and positive definite
\mathcal{G}	Set of classes in Discriminant Analysis
g	Pilot bandwidth
g_i	Univariate binning grid points
\mathbf{g}_i	Multivariate binning grid points
Hf	Hessian operator of f
\mathcal{H}_r	r -th order Hermite polynomial
\mathbf{H}	$d \times d$ bandwidth matrix (smoothing matrix) which is symmetric and positive definite
h	Bandwidth scalar (smoothing parameter)
$\mathbf{H}_{\mathcal{F}}$	Class of symmetric, positive definite $d \times d$ matrices
$\mathbf{H}_{\mathcal{D}}$	Subclass of diagonal positive definite $d \times d$ matrices
$\mathbf{H}_{\mathcal{S}}$	Subclass of a positive constant times the identity matrix
I	Antiderivative of the kernel function K
K	Unscaled kernel function
\mathcal{K}	Symmetric univariate kernel, this symbol was used in Sect. 3.3.2 to differentiate univariate (\mathcal{K}) and multivariate (\mathcal{K}) kernels
K^{P}	Product kernel
K^{R}	Radially symmetric kernel
$K * K$	Convolution of kernel K with itself, that is $K * K(x) = \int K(u)K(x - u) du$
$K_{\mathbf{H}}$	Scaled kernel function with bandwidth matrix \mathbf{H}
K_h	Scaled kernel function with bandwidth scalar h
L	Pilot kernel (the first meaning of the symbol)
L	Values which is $L < M - 1$ (the second meaning of the symbol)
M	Number of grid points
$\mathbf{m}_{\mathbf{H}}(\mathbf{x})$	Mean shift operator
$\mathcal{N}(x; \mu, \sigma)$	Univariate normal distribution with mean μ and standard deviation σ at point x
$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ at vector point \mathbf{x}
n	Random sample size (usually size of experimental data)
$R(g) = \int g(x)^2 dx$	(g is a real-valued univariate function)
T	Geometric mean of the values of KDE of all data points, that is $\hat{f}(x_1, h), \hat{f}(x_2, h), \dots, \hat{f}(x_n, h)$
V	Volume of the region \mathcal{R}
$\text{vec}\mathbf{A}$	The vectorization operator, it is a transformation that converts a matrix \mathbf{A} into a column vector by stacking the columns of this matrix on top of one another

$\text{vech}\mathbf{A}$	The half-vectorization operator (or vector half operator), it is defined only for a symmetric matrix \mathbf{A} by vectorizing only the lower triangular part of \mathbf{A}
X	the univariate data vector
X_1, X_2, \dots, X_n	Univariate random sample of size n (usually experimental data)
$\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$	Multivariate random sample of size n and dimension d (usually experimental data). $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{id})^T, i = 1, 2, \dots, n$
x	Univariate data point
$\lceil x \rceil$	Ceiling(x), least integer greater than or equal to x
$\lfloor x \rfloor$	Floor(x), greatest integer less than or equal to x
δ_k	Mesh size (used in binning)
λ	Largest eigenvalue of \mathbf{H}
$\mu_l(g) = \int x^l g(x) dx$	(l -th central moment of g , g is a real-valued univariate function)
π_k	Prior probability of the class k in Discriminant Analysis
$\hat{\pi}_k$	Sample proportion of the class k in Discriminant Analysis
$\phi(x)$	Density of the standard univariate normal distribution having mean zero and variance equal to one
$\phi_\sigma(x)$	Density of the univariate normal distribution having mean zero and variance σ
$\phi_\Sigma(\mathbf{x})$	Density of the multivariate normal distribution having mean zero and covariance matrix Σ
Ψ_r	Integrated density derivative functional Ψ_r $= \int f^{(r)}(x)f(x) dx$ (r an even integer, $\Psi_r = 0$ if r is odd)
Ψ_4	Matrix of fourth-order density derivative functionals of dimension $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$
ψ_4	Vector of fourth-order density derivative functionals of dimension d^4

List of Figures

Fig. 1.1	Probability density function of the normal (or Gaussian) distribution	2
Fig. 2.1	A sample histogram for a toy univariate dataset of seven data points	8
Fig. 2.2	Five histograms with different bin origins and constant bin width	9
Fig. 2.3	ASH density estimates of a trimodal mixture of Gaussians	11
Fig. 2.4	Hypercubes in one, two and three dimensions	12
Fig. 2.5	A visualization of the idea of the Parzen windows and the k -nearest neighbors techniques	14
Fig. 2.6	Window functions in one and two dimensions	15
Fig. 2.7	Construction of the Parzen windows estimator for $n = 4$ as a sum of boxes centered at the observations.	16
Fig. 2.8	Parzen window estimates of the mixture of two Gaussians.	17
Fig. 2.9	The distance between the estimation point x and its k -th closest neighbors.	18
Fig. 2.10	Four toy datasets (of size 1, 3, 4 and 5) used to demonstrate the main idea behind KNN estimators	20
Fig. 2.11	KNN estimates of a mixture of two Gaussians	21
Fig. 2.12	KNN estimates of a mixture of two Gaussians	22
Fig. 2.13	Two-dimensional toy example of a KNN ($k = 5$) classifier	23
Fig. 2.14	KNN classifiers on a two-class Gaussian mixture data	24
Fig. 3.1	Selected kernel shapes	28
Fig. 3.2	A toy example demonstrating the idea of the kernel density estimation with Gaussian kernels	30
Fig. 3.3	Three kernel density estimates with different bandwidths	32
Fig. 3.4	Six different kernel types and theirs KDEs.	33

Fig. 3.5	Six different kernel types and their KDEs.	34
Fig. 3.6	A toy example demonstrates the idea of the kernel density estimation with Gaussian kernels	37
Fig. 3.7	An illustrative example of contour-type and perspective-type presentations of a bivariate KDE	38
Fig. 3.8	An illustrative example of 3D KDE combining both contour-type and scatterplot-type plots	38
Fig. 3.9	Contours from product (left column) and radially symmetric (right column) kernels with equal bandwidths in each direction	40
Fig. 3.10	A toy example demonstrating the idea of the linear transformation.	42
Fig. 3.11	Scatterplots of two datasets of size $n = 5000$, where the whitening transformation is not enough to get spherically-symmetric distributions	43
Fig. 3.12	Plot of the AMISE, MISE, integrated squared bias and integrated variance versus h	48
Fig. 3.13	A demonstration of the balloon KDE for a toy five points example.	52
Fig. 3.14	A demonstration of the sample point KDE for a toy six points example	53
Fig. 3.15	A demonstration of the sample point KDE.	54
Fig. 3.16	A demonstration of the KDE with boundary correction	55
Fig. 3.17	A demonstration of KDE with boundary correction	56
Fig. 3.18	A toy example demonstrating the idea of the cumulative distribution function estimation with Epanechnikov kernels	58
Fig. 3.19	Volume of a unit hypersphere as a function of dimensionality	61
Fig. 4.1	LSCV(h) versus h for a number of samples	70
Fig. 4.2	$h_{\text{MISE}} - h$ for LSCV and PI(DPI) methods and different n	71
Fig. 4.3	Contour plots for 12 bivariate target densities (normal mixtures)	81
Fig. 4.4	Boxplots of the ISE error for the KDE using the three bandwidth selectors (PI, SCV, LSCV/UCV).	82
Fig. 5.1	The idea of simple and linear binning	86
Fig. 5.2	Visualization of the univariate linear binning	87
Fig. 5.3	The ideas behind simple and linear binning, bivariate case.	88
Fig. 5.4	Visualization of the bivariate linear binning	89
Fig. 5.5	An illustrative example of data binning	91
Fig. 5.6	Density estimates for a sample dataset with FFT and without it, for both unconstrained and constrained bandwidth matrices	92
Fig. 5.7	Original vectors \mathbf{c} and \mathbf{k} (two upper axes) and the same vectors after zero-padding.	97

Fig. 5.8	Demonstration of how the binning affects the accuracy of the univariate KDE.	99
Fig. 5.9	Demonstration of behavior of Wand's original algorithm used for the task of bandwidth selection.	103
Fig. 5.10	Visualization of kernels	104
Fig. 5.11	Boxplots of the ISE errors	112
Fig. 5.12	The effect of the binning and FFT procedures	114
Fig. 5.13	Speed comparison results for the <i>fft-M</i> , <i>fft-L</i> and <i>direct</i> implementations	116
Fig. 6.1	Flowchart of the PLUGIN algorithm with optional data preprocessing (z-score standardization)	123
Fig. 6.2	Performance and scalability of different PLUGIN algorithm implementations	130
Fig. 6.3	Three fundamental methods of the <i>for</i> loop implementation used in the $\hat{\Psi}_4(g_4)$ calculation	131
Fig. 7.1	A toy example of the univariate KDA	135
Fig. 7.2	A toy example of the bivariate KDA	136
Fig. 7.3	A toy example of the bivariate mean shift clustering algorithm	139
Fig. 7.4	A simple demonstration of the efficiency differences between the mean shift clustering and the gradient descent clustering	140
Fig. 7.5	A simple example of the nonparametric regression.	143
Fig. 7.6	A simple example of the three main types of Shewhart control charts	144
Fig. 7.7	Six example datasets generated using the copula approach.	146
Fig. 7.8	Shewhart control charts and corresponding UCLs for two different non-normally distributed datasets (generated using the copula approach)	147
Fig. 7.9	Boxplots showing the UCLs calculated using KDE-based approach	148
Fig. 7.10	A simplified diagram showing the main components of a typical flow cytometer device	150
Fig. 7.11	A toy example of the bivariate feature significance determination of the curvature of a sample kernel density estimate	152
Fig. 7.12	A sample flow cytometry dataset with and without excessive boundary points	153
Fig. 7.13	The scatterplot of the original (untransformed) flow cytometry dataset	155
Fig. 7.14	The scatterplot of the flow cytometry dataset after the biexponential transformation.	156

Fig. 7.15	The intermediate results of the gating algorithm.	157
Fig. 7.16	A simple comparison of the gating results obtained using the <i>flowClust</i> package and the nonparametric method presented in this section.	158

List of Tables

Table 3.1	Popular univariate kernel types.	27
Table 5.1	Speed comparisons of the FFT-based, sequential non-FFT and vectorized non-FFT versions.	110
Table 5.2	Values of P_1 and P_2 calculated using (5.42) for selected grid and sample sizes	117
Table 5.3	Values of L_k calculated using (5.29) for some selected values of the grid and sample sizes.	118
Table 6.1	Resource usage for three different FPGA implementations of the PLUGIN algorithms compared with the CPU and GPU implementations.	127
Table 6.2	Execution times (in sec.) for the three different FPGA implementations of the PLUGIN algorithm and for the CPU and GPU implementations.	128
Table 6.3	Speedups for the three different FPGA implementations of the PLUGIN algorithm and for the CPU and GPU implementations	128
Table 6.4	Accuracy (relative error) for the three different FPGA implementations of the PLUGIN algorithm.	129