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

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 Artur Gramacki
September 2017

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

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
$\boldsymbol{a}$	Vector of size d
c_i	Univariate binning grid counts
$\boldsymbol{c_i}$	Multivariate binning grid counts
$\mathsf{D} f$	First derivative (gradient) of $\boldsymbol{f}$
$\mathsf{D}^{\otimes r}$	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
$\hat{f}(\boldsymbol{x}; \mathbf{H})$	Multivariate kernel density estimate at point $\boldsymbol{x}$ and with bandwidth matrix $\mathbf{H}$
$\hat{f}_{-i}(\boldsymbol{x}; \mathbf{H})$	Multivariate leave-one-out kernel density estimate at point 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
$\boldsymbol{g_i}$	Multivariate binning grid points
$\bigtriangledown f$	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
$\boldsymbol{m}_{\mathbf{H}}(\boldsymbol{x})$	Mean shift operator
$\mathcal{N}(x; \mu, \sigma)$	Univariate normal distribution with mean μ and standard deviation σ at point x
$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$	Multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ at vector point $\boldsymbol{x}$
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 $\hat{f}(x_1, h), \hat{f}(x_2, h), \ldots, \hat{f}(x_n, h)$
V	Volume of the region $\mathcal{R}$
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

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
$X_1, X_2, \ldots, X_n$	Univariate random sample of size n (usually experimental data)
$\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$	Multivariate random sample of size n and dimension d (usually experimental data). $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})^T, i = 1, 2, \ldots, n$
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
$\phi_{\boldsymbol{\Sigma}}(\boldsymbol{x})$	Density of the multivariate normal distribution having mean zero and covariance matrix $\boldsymbol{\Sigma}$
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)
$\boldsymbol{\Psi_4}$	Matrix of fourth-order density derivative functionals of dimension $\frac{1}{2}d(d+1) \times \frac{1}{2}d(d+1)$
$\boldsymbol{\psi_4}$	Vector of fourth-order density derivative functionals of dimension d^4

List of Figures

List of Tables