

MULTIVARIATE NONPARAMETRIC REGRESSION AND VISUALIZATION



MULTIVARIATE NONPARAMETRIC REGRESSION AND VISUALIZATION

With R and Applications to Finance

Jussi Klemelä

 WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Copyright ©2014 by John Wiley & Sons, Inc. All rights reserved.

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 646-8600, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department with the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data:

Klemelä, Jussi.
Multivariate Nonparametric Regression and Visualization: with R and Applications to Finance /
Jussi Klemelä.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-470-38442-8 (cloth)
1. Smoothing (Statistics) 2. Estimation theory
QA278.K584 2014
519.5 - - dc22
Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To my parents



CONTENTS IN BRIEF

PART I METHODS OF REGRESSION AND CLASSIFICATION

1	Overview of Regression and Classification	3
2	Linear Methods and Extensions	77
3	Kernel Methods and Extensions	127
4	Semiparametric and Structural Models	229
5	Empirical Risk Minimization	241

PART II VISUALIZATION

6	Visualization of Data	277
7	Visualization of Functions	295



CONTENTS

Preface	xvii
Introduction	xix
I.1 Estimation of Functionals of Conditional Distributions	xx
I.2 Quantitative Finance	xxi
I.3 Visualization	xxi
I.4 Literature	xxiii
PART I METHODS OF REGRESSION AND CLASSIFICATION	
1 Overview of Regression and Classification	3
1.1 Regression	3
1.1.1 Random Design and Fixed Design	4
1.1.2 Mean Regression	5
1.1.3 Partial Effects and Derivative Estimation	8
1.1.4 Variance Regression	9
1.1.5 Covariance and Correlation Regression	13
1.1.6 Quantile Regression	14
1.1.7 Approximation of the Response Variable	18
1.1.8 Conditional Distribution and Density	21
	ix

1.1.9	Time Series Data	23
1.1.10	Stochastic Control	25
1.1.11	Instrumental Variables	26
1.2	Discrete Response Variable	29
1.2.1	Binary Response Models	29
1.2.2	Discrete Choice Models	31
1.2.3	Count Data	33
1.3	Parametric Family Regression	33
1.3.1	General Parametric Family	33
1.3.2	Exponential Family Regression	35
1.3.3	Copula Modeling	36
1.4	Classification	37
1.4.1	Bayes Risk	38
1.4.2	Methods of Classification	39
1.5	Applications in Quantitative Finance	42
1.5.1	Risk Management	42
1.5.2	Variance Trading	44
1.5.3	Portfolio Selection	45
1.5.4	Option Pricing and Hedging	50
1.6	Data Examples	52
1.6.1	Time Series of S&P 500 Returns	52
1.6.2	Vector Time Series of S&P 500 and Nasdaq-100 Returns	53
1.7	Data Transformations	53
1.7.1	Data Sphering	54
1.7.2	Copula Transformation	55
1.7.3	Transformations of the Response Variable	56
1.8	Central Limit Theorems	58
1.8.1	Independent Observations	58
1.8.2	Dependent Observations	58
1.8.3	Estimation of the Asymptotic Variance	60
1.9	Measuring the Performance of Estimators	61
1.9.1	Performance of Regression Function Estimators	61
1.9.2	Performance of Conditional Variance Estimators	66
1.9.3	Performance of Conditional Covariance Estimators	68
1.9.4	Performance of Quantile Function Estimators	69
1.9.5	Performance of Estimators of Expected Shortfall	71
1.9.6	Performance of Classifiers	72
1.10	Confidence Sets	73

1.10.1	Pointwise Confidence Intervals	73
1.10.2	Confidence Bands	75
1.11	Testing	75
2	Linear Methods and Extensions	77
2.1	Linear Regression	78
2.1.1	Least Squares Estimator	79
2.1.2	Generalized Method of Moments Estimator	81
2.1.3	Ridge Regression	84
2.1.4	Asymptotic Distributions for Linear Regression	87
2.1.5	Tests and Confidence Intervals for Linear Regression	90
2.1.6	Variable Selection	92
2.1.7	Applications of Linear Regression	94
2.2	Varying Coefficient Linear Regression	97
2.2.1	The Weighted Least Squares Estimator	97
2.2.2	Applications of Varying Coefficient Regression	98
2.3	Generalized Linear and Related Models	102
2.3.1	Generalized Linear Models	102
2.3.2	Binary Response Models	104
2.3.3	Growth Models	107
2.4	Series Estimators	107
2.4.1	Least Squares Series Estimator	107
2.4.2	Orthonormal Basis Estimator	108
2.4.3	Splines	110
2.5	Conditional Variance and ARCH Models	111
2.5.1	Least Squares Estimator	112
2.5.2	ARCH Model	113
2.6	Applications in Volatility and Quantile Estimation	116
2.6.1	Benchmarks for Quantile Estimation	116
2.6.2	Volatility and Quantiles with the LS Regression	118
2.6.3	Volatility with the Ridge Regression	121
2.6.4	Volatility and Quantiles with ARCH	122
2.7	Linear Classifiers	124
3	Kernel Methods and Extensions	127
3.1	Regressogram	129
3.2	Kernel Estimator	130
3.2.1	Definition of the Kernel Regression Estimator	130

3.2.2	Comparison to the Regressogram	132
3.2.3	Gasser–Müller and Priestley–Chao Estimators	134
3.2.4	Moving Averages	134
3.2.5	Locally Stationary Data	136
3.2.6	Curse of Dimensionality	140
3.2.7	Smoothing Parameter Selection	140
3.2.8	Effective Sample Size	142
3.2.9	Kernel Estimator of Partial Derivatives	145
3.2.10	Confidence Intervals in Kernel Regression	146
3.3	Nearest-Neighbor Estimator	147
3.4	Classification with Local Averaging	148
3.4.1	Kernel Classification	148
3.4.2	Nearest-Neighbor Classification	149
3.5	Median Smoothing	151
3.6	Conditional Density Estimation	152
3.6.1	Kernel Estimator of Conditional Density	152
3.6.2	Histogram Estimator of Conditional Density	156
3.6.3	Nearest-Neighbor Estimator of Conditional Density	157
3.7	Conditional Distribution Function Estimation	158
3.7.1	Local Averaging Estimator	159
3.7.2	Time–Space Smoothing	159
3.8	Conditional Quantile Estimation	160
3.9	Conditional Variance Estimation	162
3.9.1	State–Space Smoothing and Variance Estimation	162
3.9.2	GARCH and Variance Estimation	163
3.9.3	Moving Averages and Variance Estimation	172
3.10	Conditional Covariance Estimation	176
3.10.1	State–Space Smoothing and Covariance Estimation	178
3.10.2	GARCH and Covariance Estimation	178
3.10.3	Moving Averages and Covariance Estimation	181
3.11	Applications in Risk Management	181
3.11.1	Volatility Estimation	182
3.11.2	Covariance and Correlation Estimation	193
3.11.3	Quantile Estimation	198
3.12	Applications in Portfolio Selection	205
3.12.1	Portfolio Selection Using Regression Functions	205
3.12.2	Portfolio Selection Using Classification	215
3.12.3	Portfolio Selection Using Markowitz Criterion	223

4	Semiparametric and Structural Models	229
4.1	Single-Index Model	230
4.1.1	Definition of the Single-Index Model	230
4.1.2	Estimators in the Single-Index Model	230
4.2	Additive Model	234
4.2.1	Definition of the Additive Model	234
4.2.2	Estimators in the Additive Model	235
4.3	Other Semiparametric Models	237
4.3.1	Partially Linear Model	237
4.3.2	Related Models	238
5	Empirical Risk Minimization	241
5.1	Empirical Risk	243
5.1.1	Conditional Expectation	243
5.1.2	Conditional Quantile	244
5.1.3	Conditional Density	245
5.2	Local Empirical Risk	247
5.2.1	Local Polynomial Estimators	247
5.2.2	Local Likelihood Estimators	255
5.3	Support Vector Machines	257
5.4	Stagewise Methods	259
5.4.1	Forward Stagewise Modeling	259
5.4.2	Stagewise Fitting of Additive Models	261
5.4.3	Projection Pursuit Regression	262
5.5	Adaptive Regressograms	264
5.5.1	Greedy Regressograms	264
5.5.2	CART	268
5.5.3	Dyadic CART	271
5.5.4	Bootstrap Aggregation	272

PART II VISUALIZATION

6	Visualization of Data	277
6.1	Scatter Plots	278
6.1.1	Two-Dimensional Scatter Plots	278
6.1.2	One-Dimensional Scatter Plots	278
6.1.3	Three- and Higher-Dimensional Scatter Plots	282
6.2	Histogram and Kernel Density Estimator	283

6.3	Dimension Reduction	284
6.3.1	Projection Pursuit	284
6.3.2	Multidimensional Scaling	286
6.4	Observations as Objects	288
6.4.1	Graphical Matrices	289
6.4.2	Parallel Coordinate Plots	290
6.4.3	Other Methods	293
7	Visualization of Functions	295
7.1	Slices	296
7.2	Partial Dependence Functions	298
7.3	Reconstruction of Sets	299
7.3.1	Estimation of Level Sets of a Function	300
7.3.2	Point Cloud Data	303
7.4	Level Set Trees	304
7.4.1	Definition and Illustrations	304
7.4.2	Calculation of Level Set Trees	308
7.4.3	Volume Function	313
7.4.4	Barycenter Plot	321
7.4.5	Level Set Trees in Regression Function Estimation	322
7.5	Unimodal Densities	325
7.5.1	Probability Content of Level Sets	327
7.5.2	Set Visualization	327
Appendix A:	R Tutorial	329
A.1	Data Visualization	329
A.1.1	QQ Plots	329
A.1.2	Tail Plots	330
A.1.3	Two-Dimensional Scatter Plots	330
A.1.4	Three-Dimensional Scatter Plots	331
A.2	Linear Regression	331
A.3	Kernel Regression	332
A.3.1	One-Dimensional Kernel Regression	332
A.3.2	Moving Averages	333
A.3.3	Two-Dimensional Kernel Regression	334
A.3.4	Three- and Higher-Dimensional Kernel Regression	336
A.3.5	Kernel Estimator of Derivatives	338
A.3.6	Combined State- and Time-Space Smoothing	340
A.4	Local Linear Regression	341

A.4.1	One-Dimensional Local Linear Regression	341
A.4.2	Two-Dimensional Local Linear Regression	342
A.4.3	Three- and Higher-Dimensional Local Linear Regression	343
A.4.4	Local Linear Derivative Estimation	343
A.5	Additive Models: Backfitting	344
A.6	Single-Index Regression	345
A.6.1	Estimating the Index	346
A.6.2	Estimating the Link Function	346
A.6.3	Plotting the Single-Index Regression Function	346
A.7	Forward Stagewise Modeling	347
A.7.1	Stagewise Fitting of Additive Models	347
A.7.2	Projection Pursuit Regression	348
A.8	Quantile Regression	349
A.8.1	Linear Quantile Regression	349
A.8.2	Kernel Quantile Regression	349
References		351
Author Index		361
Topic Index		365



PREFACE

The book is intended for students and researchers who want to learn to apply non-parametric and semiparametric methods and to use visualization tools related to these estimation methods. In particular, the book is intended for students and researchers in quantitative finance who want to apply statistical methods and for students and researchers of statistics who want to learn to apply statistical methods in quantitative finance. The book continues the themes of Klemelä (2009), which studied density estimation. The current book focuses on regression function estimation.

The book was written at the University of Oulu, Department of Mathematical Sciences. I wish to acknowledge the support provided by the University of Oulu and the Department of Mathematical Sciences.

The web page of the book is <http://cc.oulu.fi/~jklemela/regstruct/>.

JUSSI KLEMELÄ

*Oulu, Finland
October 2013*



INTRODUCTION

We study regression analysis and classification, as well as estimation of conditional variances, quantiles, densities, and distribution functions. The focus of the book is on nonparametric methods. Nonparametric methods are flexible and able to adapt to various kinds of data, but they can suffer from the curse of dimensionality and from the lack of interpretability. Semiparametric methods are often able to cope with quite high-dimensional data and they are often easier to interpret, but they are less flexible and their use may lead to modeling errors. In addition to terms “nonparametric estimator” and “semiparametric estimator”, we can use the term “structured estimator” to denote such estimators that arise, for example, in additive models. These estimators obey a structural restriction, whereas the term “semiparametric estimator” is used for estimators that have a parametric and a nonparametric component.

Nonparametric, semiparametric, and structured methods are well established and widely applied. There are, nevertheless, areas where a further work is useful. We have included three such areas in this book:

1. Estimation of several functionals of a conditional distribution; not only estimation of the conditional expectation but also estimation of the conditional variance and conditional quantiles.
2. Quantitative finance as an area of application for nonparametric and semiparametric methods.

3. Visualization tools in statistical learning.

I.1 ESTIMATION OF FUNCTIONALS OF CONDITIONAL DISTRIBUTIONS

One of the main topics of the book are the kernel methods. Kernel methods are easy to implement and computationally feasible, and their definition is intuitive. For example, a kernel regression estimator is a local average of the values of the response variable. Local averaging is a general regression method. In addition to the kernel estimator, examples of local averaging include the nearest-neighbor estimator, the regressogram, and the orthogonal series estimator.

We cover linear regression and generalized linear models. These models can be seen as starting points to many semiparametric and structured regression models. For example, the single index model, the additive model, and the varying coefficient linear regression model can be seen as generalizations of the linear regression model or the generalized linear model.

Empirical risk minimization is a general approach to statistical estimation. The methods of empirical risk minimization can be used in regression function estimation, in classification, in quantile regression, and in the estimation of other functionals of the conditional distribution. The method of local empirical risk minimization is a method which can be seen as a generalization of the kernel regression.

A regular regressogram is a special case of local averaging, but the empirical choice of the partition leads to a rich class of estimators. The choice of the partition is made using empirical risk minimization. In the one- and two-dimensional cases a regressogram is usually less efficient than the kernel estimator, but in high-dimensional cases a regressogram can be useful. For example, a method to select the partition of a regressogram can be seen as a method of variable selection, if the chosen partition is such that it can be defined using only a subset of the variables. The estimators that are defined as a solution of an optimization problem, like the minimizers of an empirical risk, need typically be calculated with numerical methods. Stagewise algorithms can also be taken as a definition of an estimator, even without giving an explicit minimization problem which they solve.

A regression function is defined as the conditional expectation of the distribution of a response variable. The conditional expectation is useful in making predictions as well as in finding causal relationships. We cover also the estimation of the conditional variance and conditional quantiles. These are needed to give a more complete view of the conditional distribution. Also, the estimation of the conditional variance and conditional quantiles is needed in risk management, which is an important area of quantitative finance. The conditional variance can be estimated by estimating the conditional expectation of the squared random variable, whereas a conditional quantile is a special case of the conditional median. In the time series setting the standard approaches for estimating the conditional variance are the ARCH and GARCH modeling, but we discuss nonparametric alternatives. The GARCH estimator is close

to a moving average, whereas the ARCH estimator is related to linear state space modeling.

In classification we are not interested in the estimation of functionals of a distribution, but the aim is to construct classification rules. However, most of the regression function estimation methods have a counterpart in classification.

I.2 QUANTITATIVE FINANCE

Risk management, portfolio selection, and option pricing can be identified as three important areas of quantitative finance. Parametric statistical methods have been dominating the statistical research in quantitative finance. In risk management, probability distributions have been modeled with the Pareto distribution or with distributions derived from the extreme value theory. In portfolio selection the multivariate normal model has been used together with the Markowitz theory of portfolio selection. In option pricing the Black-Scholes model of stock prices has been widely applied. The Black-Scholes model has also been extended to more general parametric models for the process of stock prices.

In risk management the p -quantile of a loss distribution has a direct interpretation as such threshold that the probability of the loss exceeding the threshold is less than p . Thus estimation of conditional quantiles is directly relevant for risk management. Unconditional quantile estimators do not take into account all available information, and thus in risk management it is useful to estimate conditional quantiles. The estimation of the conditional variance can be applied in the estimation of a conditional quantile, because in location-scale families the variance determines the quantiles. The estimation of conditional variance can be extended to the estimation of the conditional covariance or the conditional correlation.

We apply nonparametric regression function estimation in portfolio selection. The portfolio is selected either with the maximization of a conditional expected utility or with the maximization of a Markowitz criterion. When the collection of allowed portfolio weights is a finite set, then also classification can be used in portfolio selection. The squared returns are much easier to predict than the returns themselves, and thus in quantitative finance the focus has been in the prediction of volatility. However, it can be shown that despite the weak predictability of the returns, portfolio selection can profit from statistical prediction.

Option pricing can be formulated as a problem of stochastic control. We do not study the statistics of option pricing in detail, but give a basic framework for solving some option pricing problems nonparametrically.

I.3 VISUALIZATION

Statistical visualization is often considered as a visualization of the raw data. The visualization of the raw data can be a part of the exploratory data analysis, a first step to model building, and a tool to generate hypotheses about the data-generating mechanism. However, we put emphasis on a different approach to visualization.

In this approach, visualization tools are associated with statistical estimators or inference procedures. For example, we estimate first a regression function and then try to visualize and describe the properties of this regression function estimate. The distinction between the visualization of the raw data and the visualization of the estimator is not clear when nonparametric function estimation is used. In fact, nonparametric function estimation can be seen as a part of exploratory data analysis.

The SiZer is an example of a tool that combines visualization and inference, see Chaudhuri & Marron (1999). This methodology combines formal testing for the existence of modes with the SiZer maps to find out whether a mode of a density estimate of a regression function estimate is really there.

Semiparametric function estimates are often easier to visualize than nonparametric function estimates. For example, in a single index model the regression function estimate is a composition of a linear function and a univariate function. Thus in a single index model we need only to visualize the coefficients of the linear function and a one-dimensional function. The ease of visualization gives motivation to study semiparametric methods.

CART, as presented in Breiman, Friedman, Olshen & Stone (1984), is an example of an estimation method whose popularity is not only due to its statistical properties but also because it is defined in terms of a binary tree that gives directly a visualization of the estimator. Even when it is possible to find estimators with better statistical properties than CART, the possibility to visualization gives motivation to use CART.

Visualization of nonparametric function estimates, such as kernel estimates, is challenging. For the visualization of completely nonparametric estimates, we can use level set tree-based methods, as presented in Klemelä (2009). Level set tree-based methods have found interest also in topological data analysis and in scientific visualization, and these methods have their origin in the concept of a Reeb graph, defined originally in Reeb (1946).

In density estimation we are often interested in the mode structure of the density, defined as the number of local extremes, the largeness of the local extremes, and the location of the local extremes. The local extremes of a density function are related to the areas of concentration of the probability mass. In regression function estimation we are also interested in the mode structure. The local maxima of a regression function are related to the regions of the space of the explanatory variables where the response variable takes the largest values. The antimode structure is equally important to describe. The antimode structure means the number of local minima, the size of the local minima, and the location of the local minima. The local minima of a regression function are related to the areas of the space of the explanatory variables where the response variable takes the smallest values.

The mode structure of a regression function does not give complete information about the properties of the regression function. In regression analysis we are interested in the effects of the explanatory variables on the response variable and in the interaction between the explanatory variables. The effect of an explanatory variable can be formalized with the concept of a partial effect. The partial effect of an explanatory variable is the partial derivative of the regression function with respect to this variable. Nearly constant partial effects indicate that the regression function is

close to a linear function, since the partial derivatives of a linear function are constants. The local maxima of a partial effect correspond to the areas in the space of the explanatory variables where the increase of the expected value of the response variable, resulting from an increase of the value of the explanatory variable, is the largest. We can use level set trees of partial effects to visualize the mode structure and the antimode structure of the partial effects, and thus to visualize the effects and the interactions of the explanatory variables.

I.4 LITERATURE

We mention some of the books that have been used in the preparation of this book. Härdle (1990) covers nonparametric regression with an emphasis on kernel regression, discussing smoothing parameter selection, giving confidence bands, and providing various econometric examples. Hastie, Tibshirani & Friedman (2001) describe high-dimensional linear and nonlinear classification and regression methods, giving many examples from biometry and machine learning. Györfi, Kohler, Krzyzak & Walk (2002) cover asymptotic theory of kernel regression, nearest-neighbor regression, empirical risk minimization, and orthogonal series methods, and they also include a treatment of time series prediction. Ruppert, Wand & Carroll (2003) view nonparametric regression as an extension of parametric regression and treat them together. Härdle, Müller, Sperlich & Werwatz (2004) explain single index models, generalized partial linear models, additive models, and several nonparametric regression function estimators, giving econometric examples. Wooldridge (2005) provides an asymptotic theory of linear regression, including instrumental variables and panel data. Fan & Yao (2005) study nonlinear time series and use nonparametric function estimation in time series prediction and explanation. Wasserman (2005) provides information on nonparametric regression and density estimation with confidence intervals and bootstrap confidence intervals. Horowitz (2009) covers semiparametric models and discusses the identifiability and asymptotic distributions. Spokoiny (2010) introduces local parametric methods into nonparametric estimation.

Bouchaud & Potters (2003) have developed nonparametric techniques for financial analysis. Franke, Härdle & Hafner (2004) discuss statistical analysis of financial markets, with emphasis being on the parametric methods. Ruppert (2004) is a textbook suitable for statistics students interested in quantitative finance, and this book discusses statistical tools related to classical financial models. Malevergne & Sornette (2005) have analyzed financial data with nonparametric methods. Li & Racine (2007) consider various non- and semiparametric regression models presenting asymptotic distribution theory and the theory of smoothing parameter selection, directing towards econometric applications.

