Nonparametric Econometrics: A Primer

Nonparametric Econometrics: A Primer

Jeffrey S. Racine

Department of Economics
McMaster University
1280 Main Street West
Hamilton, Ontario
Canada L8S 4M4
racinej@mcmaster.ca

the essence of knowledge

Boston – Delft

Foundations and Trends® in Econometrics

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is J. S. Racine, Nonparametric Econometrics: A Primer, Foundations and Trends® in Econometrics, vol 3, no 1, pp 1–88, 2008

ISBN: 978-1-60198-110-3

Foundations and Trends® in Econometrics

Volume 3 Issue 1, 2008

Editorial Scope

Foundations and Trends® in Econometrics will publish survey and tutorial articles in the following topics:

- Identification
- Model Choice and Specification Analysis
- Non-linear Regression Models
- Simultaneous Equation Models
- Estimation Frameworks
- Biased Estimation
- Computational Problems
- Microeconometrics
- Treatment Modeling
- Discrete Choice Modeling
- Models for Count Data
- Duration Models
- Limited Dependent Variables
- Panel Data
- Dynamic Specification
- Inference and Causality
- Continuous Time Stochastic Models
- Modeling Non-linear Time Series
- Unit Roots
- Cointegration
- Latent Variable Models
- Qualitative Response Models
- Hypothesis Testing
- Interactions-based Models
- Duration Models
- Financial Econometrics
- Measurement Error in Survey Data
- Productivity Measurement and Analysis
- Semiparametric and Nonparametric Estimation
- Bootstrap Methods
- Nonstationary Time Series
- Robust Estimation

Information for Librarians

Foundations and Trends® in Econometrics, 2008, Volume 3, 4 issues. ISSN paper version 1551-3076. ISSN online version 1551-3084. Also available as a combined paper and online subscription.

Foundations and Trends® in
Econometrics
Vol. 3, No 1 (2008) 1–88

DOI: 10.1561/0800000009

Nonparametric Econometrics: A Primer

Jeffrey S. Racine

Department of Economics, McMaster University, 1280 Main Street West, Hamilton, Ontario, Canada L8S 4M4, racinej@mcmaster.ca

Abstract

This review is a primer for those who wish to familiarize themselves with nonparametric econometrics. Though the underlying theory for many of these methods can be daunting for some practitioners, this article will demonstrate how a range of nonparametric methods can in fact be deployed in a fairly straightforward manner. Rather than aiming for encyclopedic coverage of the field, we shall restrict attention to a set of touchstone topics while making liberal use of examples for illustrative purposes. We will emphasize settings in which the user may wish to model a dataset comprised of continuous, discrete, or categorical data (nominal or ordinal), or any combination thereof. We shall also consider recent developments in which some of the variables involved may in fact be irrelevant, which alters the behavior of the estimators and optimal bandwidths in a manner that deviates substantially from conventional approaches.

Contents

1

Introduction

Nonparametric methods are statistical techniques that do not require a researcher to specify functional forms for objects being estimated. Instead, the data itself informs the resulting model in a particular manner. In a regression framework this approach is known as “nonparametric regression” or “nonparametric smoothing.” The methods we survey are known as kernel[1] methods. Such methods are becoming increasingly popular for applied data analysis; they are best suited to situations involving large data sets for which the number of variables involved is manageable. These methods are often deployed after common parametric specifications are found to be unsuitable for the problem at hand, particularly when formal rejection of a parametric model based on specification tests yields no clues as to the direction in which to search for an improved parametric model. The appeal of nonparametric methods stems from the fact that they relax the parametric assumptions imposed on the data generating process and let the data determine an appropriate model.

[1] A “kernel” is simply a weighting function.

Nonparametric and semiparametric methods have attracted a great deal of attention from statisticians in the past few decades, as evidenced by the vast array of texts written by statisticians including Prakasa Rao (1983), Devroye and Györfi (1985), Silverman (1986), Scott (1992), Bickel et al. (1993), Wand and Jones (1995), Fan and Gijbels (1996), Simonoff (1996), Azzalini and Bowman (1997), Hart (1997), Efromovich (1999), Eubank (1999), Ruppert et al. (2003), Härdle et al. (2004), and Fan and Yao (2005). However, the number of texts tailored to the needs of applied econometricians is relatively scarce including, Härdle (1990), Horowitz (1998), Pagan and Ullah (1999), Yatchew (2003), and Li and Racine (2007a) being those of which we are currently aware.

The first published paper in kernel estimation appeared in 1956 (Rosenblatt (1956)), and the idea was proposed in an USAF technical report as a means of liberating discriminant analysis from rigid parametric specifications (Fix and Hodges (1951)). Since then, the field has undergone exponential growth and has even become a fixture in undergraduate textbooks (see, e.g., Johnston and DiNardo (1997, Chap. 11)), which attests to the popularity of the methods among students and researchers alike.

Though kernel methods are popular, they are but one of many approaches toward the construction of flexible models. Approaches to flexible modeling include spline, nearest neighbor, neural network, and a variety of flexible series methods, to name but a few. In this article, however, we shall restrict attention to the class of nonparametric kernel methods, and will also touch on semiparametric kernel methods as well. We shall also focus on more practical aspects of the methods and direct the interested reader to Li and Racine (2007a) and the references listed above for details on the theoretical underpinnings in order to keep this review down to a manageable size.

It bears mentioning that there are two often heard complaints regarding the state of nonparametric kernel methods, namely, (1) the lack of software, and (2) the numerical burden associated with these methods. We are of course sympathetic to both complaints. The latter may unavoidable and simply be “the nature of the beast” as they say, though see *Computational Considerations* for a discussion of the issues. However, the former is changing and recent developments hold the promise for computational breakthroughs. Many statistical software packages now contain some elementary nonparametric methods (one-dimensional density estimation, one-dimensional regression) though they often use rule-of-thumb methods for bandwidth selection which, though computationally appealing, may not be robust choices in all applications. Recently, an R (R Development Core Team (2007)) package "np" has been created that provides an easy to use and open platform for kernel estimation, and we direct the interested reader to Hayfield and Racine (2007) for details. All examples in this review were generated using the np package, and code to replicate these results is available upon request.

References

Abramson, I. S. (1982), 'On bandwidth variation in kernel estimates — a square root law'. *The Annals of Statistics* **10**, 1217–1223.

Aitchison, J. and C. G. G. Aitken (1976), 'Multivariate binary discrimination by the kernel method'. *Biometrika* **63**(3), 413–420.

Azzalini, A. and A. W. Bowman (1997), *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-plus Illustrations*. New York: Oxford University Press.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*. Baltimore: Johns Hopkins University Press.

Bowman, A. W. (1984), 'An alternative method of cross-validation for the smoothing of density estimates'. *Biometrika* **71**, 353–360.

Bowman, A. W., P. Hall, and T. Prvan (1998), 'Bandwidth selection for the smoothing of distribution functions'. *Biometrika* **85**, 799–808.

Breiman, L., W. Meisel, and E. Purcell (1977), 'Variable kernel estimates of multivariate densities'. *Technometrics* **19**, 135–144.

Brian, R. S. (2007), *KernSmooth: Functions for Kernel Smoothing for Wand and Jones* (*1995*). R package version 2.22-21. URL: http://www.maths.unsw.edu.au/wand (original by Matt Wand. R port).

Cameron, A. C. and P. K. Trivedi (1998), *Regression Analysis of Count Data*. New York: Cambridge University Press.

Cantoni, E. and E. Ronchetti (2001), 'Resistant selection of the smoothing parameter for smoothing splines'. *Statistics and Computing* **11**, 141–146.

Cheng, M.-Y., P. Hall, and D. Titterington (1997), 'On the shrinkage of local linear curve estimators'. *Statistics and Computing* **7**, 11–17.

Čížek, P. and W. Härdle (2006), 'Robust estimation of dimension reduction space'. *Computational Statistics and Data Analysis* **51**, 545–555.

Cleveland, W. S. (1979), 'Robust locally weighted regression and smoothing scatterplots'. *Journal of the American Statistical Association* **74**, 829–836.

Croissant, Y. (2006), *Ecdat: Data Sets for Econometrics*. R package version 0.1-5. URL: http://www.r-project.org.

Croissant, Y. and G. Millo (2007), *Plm: Linear Models for Panel Data*. R package version 0.2-2. URL: http://www.r-project.org.

Delgado, M. A. and W. G. Manteiga (2001), 'Significance testing in nonparametric regression based on the bootstrap'. *Annals of Statistics* **29**, 1469–1507.

Devroye, L. and L. Györfi (1985), *Nonparametric Density Estimation: The L^1 View*. New York: Wiley.

Donald, S. G. (1997), 'Inference concerning the number of factors in a multivariate nonparametric relationship'. *Econometrica* **65**, 103–131.

Draper, N. R. (1987), *Empirical Model-Building and Response Surfaces*. Wiley.

Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory and Applications*. New York: Springer Verlag.

Efron, B. (1982), *The Jackknife, the Bootstrap, and Other Resampling Plans*. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania 19103.

Elgammal, A., R. Duraiswami, and L. Davis (2003), 'The fast gauss transform for efficient kernel density evaluation with applications in computer vision'. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Epanechnikov, V. A. (1969), 'Nonparametric estimation of a multi-dimensional probability density'. *Theory of Applied Probability* **14**, 153–158.

Eubank, R. L. (1999), *Nonparametric Regression and Spline Smoothing*. Second edition. New York: Marcel Dekker.

Fan, J. (1992), 'Design-adaptive nonparametric regression'. *Journal of the American Statistical Association* **87**, 998–1004.

Fan, J. and I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*. London: Chapman and Hall.

Fan, J. and J. Jiang (2000), 'Variable bandwidth and one-step local m-estimator'. *Science in China* (*Series A*) **43**, 65–81.

Fan, J. and Q. W. Yao (2005), *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer.

Faraway, J. J. and M. Jhun (1990), 'Bootstrap choice of bandwidth for density estimation'. *Journal of the American Statistical Association* **85**, 1119–1122.

Fix, E. and J. L. Hodges (1951), 'Discriminatory analysis. nonparametric estimation: Consistency properties'. Technical Report 4, Randolph Field, Texas, USAF School of Aviation Medicine. Project No. 21–49–004.

Fox, J. (2002), *An R and S-PLUS Companion to Applied Regression*. Thousand Oaks, CA: Sage.

Geary, R. C. (1947), 'Testing for normality'. *Biometrika* **34**, 209–242.

Geisser, S. (1975), 'A predictive sample reuse method with application'. *Journal of the American Statistical Association* **70**, 320–328.

Gray, A. and A. W. Moore (2003), 'Very fast multivariate kernel density estimation via computational geometry'. Joint Statistical Meeting.

Greene, W. H. (2003), *Econometric Analysis*. Fifth edition. Upper Saddle River, NJ: Prentice Hall.

Greengard, L. (1988), *The Rapid Evaluation of Potential Fields in Particle Systems*. MIT Press.

Greengard, L. and J. Strain (1991), 'The fast gauss transform'. *Society for Industrial and Applied Mathematics: Journal of Science and Computation* **12**(1), 79–94.

Härdle, W. (1990), *Applied Nonparametric Regression*. New York: Cambridge University Press.

Härdle, W., M. Müller, S. Sperlich, and A. Werwatz (2004), *Nonparametric and Semiparametric Models*. Berlin: Springer Series in Statistics.

Härdle, W. and E. Mammen (1993), 'Comparing nonparametric versus parametric regression fits'. *Annals of Statistics* **21**, 1926–1947.

Hall, P., Q. Li, and J. S. Racine (2007), 'Nonparametric estimation of regression functions in the presence of irrelevant regressors'. *The Review of Economics and Statistics* **89**(4), 784–789.

Hall, P., J. S. Racine, and Q. Li (2004), 'Cross-validation and the estimation of conditional probability densities'. *Journal of the American Statistical Association* **99**(468), 1015–1026.

Hart, J. D. (1997), *Nonparametric Smoothing and Lack-of-Fit Tests*. New York: Springer Verlag.

Hastie, T. and R. Tibshirani (1993), 'Varying-coefficient models'. *Journal of the Royal Statistical Society Series B* **55**, 757–796.

Hayfield, T. and J. S. Racine (2007), Np: *Nonparametric Kernel Smoothing Methods for Mixed Datatypes*. R package version 0.13-1.

Henderson, D., R. J. Carroll, and Q. Li (2006), 'Nonparametric estimation and testing of fixed effects panel data models'. Unpublished manuscript, Texas A & M University.

Hodges, J. L. and E. L. Lehmann (1956), 'The efficiency of some nonparametric competitors of the t-test'. *Annals of Mathematical Statistics* **27**, 324–335.

Hoerl, A. E. and R. W. Kennard (1970), 'Ridge regression: Biased estimation for nonorthogonal problems'. *Technometrics* **12**, 55–67.

Horowitz, J. L. (1998), *Semiparametric Methods in Econometrics*. New York: Springer-Verlag.

Horowitz, J. L. and W. Härdle (1994), 'Testing a parametric model against a semiparametric alternative'. *Econometric Theory* **10**, 821–848.

Horowitz, J. L. and V. G. Spokoiny (2001), 'An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative'. *Econometrica* **69**(3), 599–631.

Hristache, M., A. Juditsky, and V. Spokoiny (2001), 'Direct estimation of the index coefficient in a single-index model'. *Annals of Statistics* **29**, 595–623.

Hsiao, C., Q. Li, and J. S. Racine (2007), 'A consistent model specification test with mixed categorical and continuous data'. *Journal of Econometrics* **140**, 802–826.

Huber, P. J. (1964), 'Robust estimation of a location parameter'. *The Annals of Statistics* **35**, 73–101.

Hurvich, C. M., J. S. Simonoff, and C. L. Tsai (1998), 'Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion'. *Journal of the Royal Statistical Society Series B* **60**, 271–293.

Ichimura, H. (1993), 'Semiparametric least squares (SLS) and weighted SLS estimation of single-index models'. *Journal of Econometrics* **58**, 71–120.

Johnston, J. and J. DiNardo (1997), *Econometric Methods*. Fourth edition. McGraw-Hill.

Klein, R. W. and R. H. Spady (1993), 'An efficient semiparametric estimator for binary response models'. *Econometrica* **61**, 387–421.

Lavergne, P. and Q. Vuong (1996), 'Nonparametric selection of regressors: The nonnested case'. *Econometrica* **64**, 207–219.

Leung, D. (2005), 'Cross-validation in nonparametric regression with outliers'. *The Annals of Statistics* **33**, 2291–2310.

Li, Q. and J. S. Racine (2003), 'Nonparametric estimation of distributions with categorical and continuous data'. *Journal of Multivariate Analysis* **86**, 266–292.

Li, Q. and J. S. Racine (2004), 'Cross-validated local linear nonparametric regression'. *Statistica Sinica* **14**(2), 485–512.

Li, Q. and J. S. Racine (2007a), *Nonparametric Econometrics: Theory and Practice*. Princeton University Press.

Li, Q. and J. S. Racine (2007b), 'Smooth varying-coefficient nonparametric models for qualitative and quantitative data'. Unpublished manuscript, Department of Economics, Texas A&M University.

Li, Q. and J. S. Racine (forthcoming), 'Nonparametric estimation of conditional CDF and quantile functions with mixed categorical and continuous data'. *Journal of Business and Economic Statistics*.

Loader, C. R. (1999), 'Bandwidth selection: Classical or plug-in?'. *Annals of Statistics* **27**(2), 415–438.

Manski, C. F. (1988), 'Identification of binary response models'. *Journal of the American Statistical Association* **83**(403), 729–738.

Maronna, A., R. D. Martin, and V. J. Yohai (2006), *Robust Statistics: Theory and Methods*. Wiley.

Nadaraya, E. A. (1965), 'On nonparametric estimates of density functions and regression curves'. *Theory of Applied Probability* **10**, 186–190.

Pagan, A. and A. Ullah (1999), *Nonparametric Econometrics*. New York: Cambridge University Press.

Parzen, E. (1962), 'On estimation of a probability density function and mode'. *The Annals of Mathematical Statistics* **33**, 1065–1076.

Prakasa Rao, B. L. S. (1983), *Nonparametric Functional Estimation*. Orlando, FL: Academic Press.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing. Vienna, Austria. (ISBN 3-900051-07-0. URL: http://www.R-project.org).

Racine, J. S. (1997), 'Consistent significance testing for nonparametric regression'. *Journal of Business and Economic Statistics* **15**(3), 369–379.

Racine, J. S. (2002), 'Parallel distributed kernel estimation'. *Computational Statistics and Data Analysis* **40**, 293–302.

Racine, J. S., J. D. Hart, and Q. Li (2006), 'Testing the significance of categorical predictor variables in nonparametric regression models'. *Econometric Reviews* **25**, 523–544.

Racine, J. S. and Q. Li (2004), 'Nonparametric estimation of regression functions with both categorical and continuous data'. *Journal of Econometrics* **119**(1), 99–130.

Racine, J. S. and L. Liu (2007), 'A partially linear kernel estimator for categorical data'. Unpublished manuscript, McMaster University.

Robinson, P. M. (1988), 'Root-n consistent semiparametric regression'. *Econometrica* **56**, 931–954.

Rosenblatt, M. (1956), 'Remarks on some nonparametric estimates of a density function'. *The Annals of Mathematical Statistics* **27**, 832–837.

Rudemo, M. (1982), 'Empirical choice of histograms and kernel density estimators'. *Scandinavian Journal of Statistics* **9**, 65–78.

Ruppert, D., R. J. Carroll, and M. P. Wand (2003), *Semiparametric Regression Modeling*. New York: Cambridge University Press.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995), 'An effective bandwidth selector for local least squares regression'. *Journal of the American Statistical Association* **90**, 1257–1270.

Scott, D. W. (1985), 'Averaged shifted histograms: Effective nonparametric density estimators in several dimensions'. *Annals of Statistics* **13**, 1024–1040.

Scott, D. W. (1992), *Multivariate Density Estimation: Theory, Practice, and Visualization*. New York: Wiley.

Scott, D. W. and S. J. Sheather (1985), 'Kernel density estimation with binned data'. *Communication in Statistics: Theory and Methods* **14**, 1353–1359.

Seifert, B. and T. Gasser (2000), 'Data adaptive ridging in local polynomial regression'. *Journal of Computational and Graphical Statistics* **9**, 338–360.

Sheather, S. J. and M. C. Jones (1991), 'A reliable data-based bandwidth selection method for kernel density estimation'. *Journal of the Royal Statistal Society, Series B* **53**, 683–690.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Simonoff, J. S. (1996), *Smoothing Methods in Statistics*. New York: Springer Series in Statistics.

Stone, C. J. (1974), 'Cross-validatory choice and assessment of statistical predictions (with discussion)'. *Journal of the Royal Statistical Society* **36**, 111–147.

Stone, C. J. (1977), 'Consistent nonparametric regression'. *Annals of Statistics* **5**, 595–645.

Venables, W. N. and B. D. Ripley (2002), *Modern Applied Statistics with S*. Fourth edition. New York: Springer.

Wand, M. P. and M. C. Jones (1995), *Kernel Smoothing*. London: Chapman and Hall.

Wang, F. T. and D. W. Scott (1994), 'The l_1 method for robust nonparametric regression'. *Journal of the American Statistical Association* **89**, 65–76.

Wang, M. C. and J. van Ryzin (1981), 'A class of smooth estimators for discrete distributions'. *Biometrika* **68**, 301–309.

Wang, N. (2003), 'Marginal nonparametric kernel regression accounting for within-subject correlation'. *Biometrika* **90**, 43–52.

Wang, N., R. J. Carroll, and X. Lin (2005), 'Efficient semiparametric marginal estimation for longitudinal/clustered data'. *Journal of the American Statistical Association* **100**, 147–157.

Watson, G. S. (1964), 'Smooth regression analysis'. *Sankhya* **26**(15), 359–372.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.

Wooldridge, J. M. (2003), *Introductory Econometrics*. Thompson South-Western.

Yatchew, A. J. (2003), *Semiparametric Regression for the Applied Econometrician*. New York: Cambridge University Press.