EXPLORING DATA WITH NON- AND SEMIPARAMETRIC MODELS

Marlene Müller
Fraunhofer ITWM, Germany
marlene.mueller@gmx.de

Today the use of exploratory and graphical techniques to analyze data is practically standard. With R (www.R-project.org) the appropriate software tools are available to everyone. We address in particular kernel density estimation and non- and semiparametric kernel regression techniques as methods that at one hand can help to explore data and on the other hand may assist in finding appropriate parametric models for fitting data. We discuss how to introduce the these methods in class and shows some examples using R.

INTRODUCTION

A basic knowledge of nonparametric and semiparametric estimation methods is more and more standard for people working in data analysis. With R (www.R-project.org) exists a comprehensive software and programming environment that provides already many non- and semiparametric estimators, but can also be used to implement new approaches. In the following we illustrate the schedule for a one-semester course in non- and semiparametric function estimation that intends to provide an overview on nonparametric function estimation for students in applied sciences.

Course topics

The course follows the textbook of Härdle, Müller, Sperlich and Werwatz (2004) and covers essentially the following three parts of topics:

- Density estimation: histogram, kernel density estimation, multidimensional kernels
- Nonparametric regression: kernel regression, more smoothers (regressogram, spline smoothing, k-nearest-neighbor regression), smoothing parameter selection
- Overview on semiparametric models: additive models, binary choice models, several semiparametric models (single index models, partial and generalized partial linear models, generalized additive models)

What students should understand?

The main focus of the course is to provide an understanding of the concepts rather than to precisely derive all theoretical properties of the estimators. In particular, the following conclusions should be possible to understand for the participants:

- Nonparametric function estimates have the capability to recover more features of the data than parametric function estimates.
- Nonparametric estimators however may have a smaller precision compared to parametric estimates given that the parametric functional form is the adequate one for the underlying data.
- The relevant steps to derive asymptotic bias, variance, MSE and MISE are discussed. This shows at which rate the convergence to the true underlying function holds and how to quantify the "price" to be paid for not forcing a parametric form of the function to estimate.
- The choice of the smoothing parameter is quite essential. Plug-in estimation and cross-validation as rather universal concepts are discussed.
- The choice of the kernel function in kernel density estimation and kernel regression has a rather small impact on the function estimator but may be essential for its computation and its smoothness when also function derivatives are of interest.
- Multidimensional nonparametric function estimation leads to the "curse of dimensionality". Moreover multidimensional estimates are more difficult in graphical presentation and interpretation.
- Different regression smoothers may provide a different complexity of computation. In particular, spline smoothing and k-nearest neighbour regression (in the one-dimensional case)

In C. Reading (Ed.), *Data and context in statistics education: Towards an evidence-based society. Proceedings of the Eighth International Conference on Teaching Statistics (ICOTS8, July, 2010), Ljubljana, Slovenia*. Voorburg, The Netherlands: International Statistical Institute. www.stat.auckland.ac.nz/~iase/publications.php [© 2010 ISI/IASE]

are compared to kernel regression methods as Nadaraya-Watson and local polynomial regression.

- Semiparametric regression models extend the concept of fully nonparametric regression. They allow us to consider partially known or parametric components and make it possible to combine categorical and continuous variables. Also, using structural assumptions on the model (e.g. additive component functions) provide an easy interpretation as well as estimation at the more precise one-dimensional rate of convergence.

Throughout the course R scripts are provided that can be altered or modified by the course participants to get a deeper insight into the methodology. In the following, we will present some of the course examples in more detail.

HISTOGRAM

To start the course we recall the histogram as a density estimator that most of the students are familiar with. It is emphasized that the histogram always provides a step function result and that different settings for the histogram bins may lead to different interpretation of the estimated distributions. A short derivation of the theoretical (asymptotic) properties is given so that the histogram can be compared with parametric and kernel density estimates in the following.

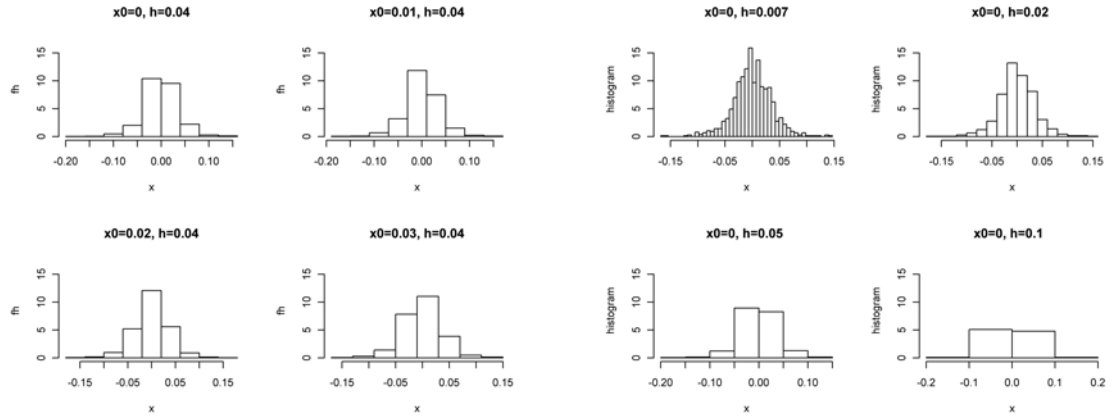

Figure 1. Different bin settings lead to different conclusions on the shape of the distribution (left panel: different origins indicate symmetric or skewed distribution, right panel: different bin widths show different smoothness of the estimated density)

DENSITY ESTIMATION

In a second chapter, kernel density estimation is then discussed in more detail. Here, the concept of local averaging is introduced and different kernel functions are introduced. A more detailed derivation of bias, variance and consequently MSE and MISE is possible here as most of the calculations can be based on Taylor expansions.

Definition of the kernel density estimator

Figure 2 shows the construction of the kernel density estimator by summing up over rescaled kernel functions. The R example script shows that the value of the density estimate corresponds to the concentration of the data points (marked in red) on the horizontal axis. The script can be changed by the course participants to obtain estimates for other sample sizes and bandwidths, for example.

Properties of nonparametric function estimators

The kernel density estimation is particularly suited to demonstrate the derivation of theoretical (asymptotic) properties of the estimator. In simulated data situations, the relevant terms can be also visualized. The left panel of Figure 3 shows the deviation between the true density and and the kernel density estimate. It can be seen that the bias effect is larger in regions where the true density has a higher curvature. The right panel of Figure 3 shows the dependence of bias and

variance of the estimator on the bandwidth h and consequently how to choose the optimal bandwidth by minimizing the mean squared error MSE.

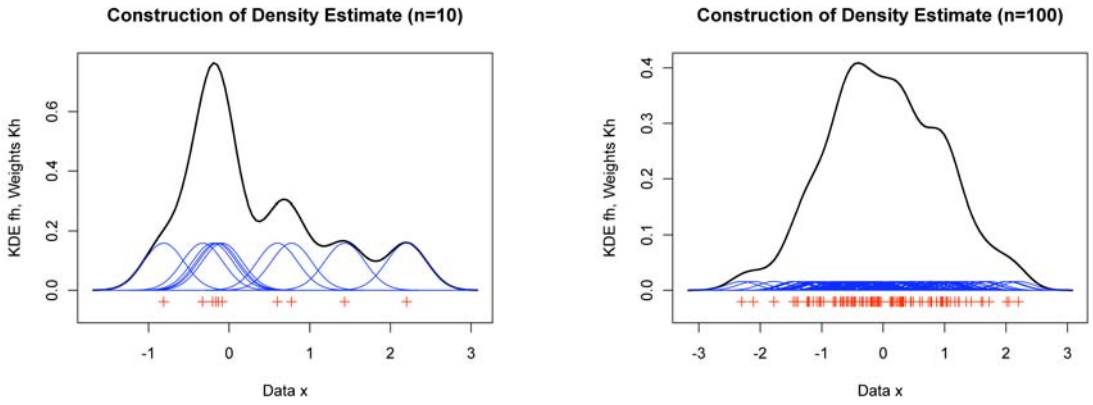

Figure 2. Visualization of the kernel density estimate construction principle: The density estimate is given by the sum of rescaled kernel functions (left panel: 10 and right panel: 100 observations)

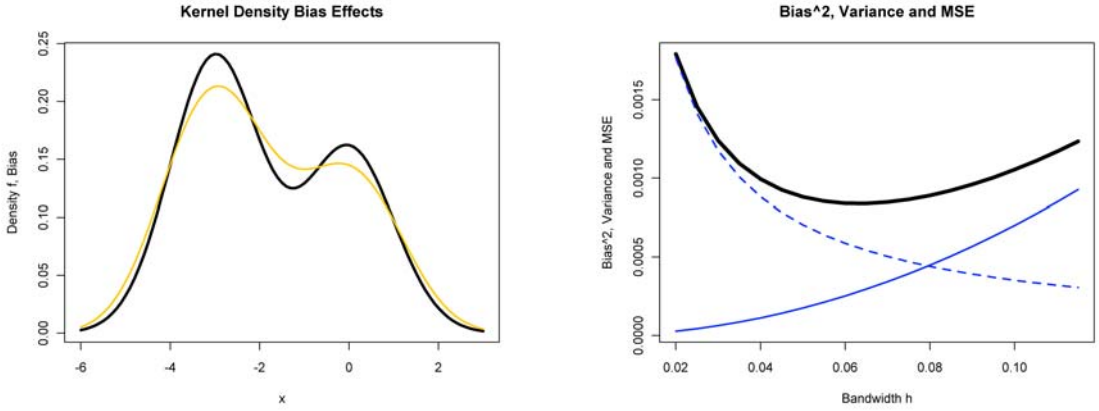

Figure 3. Illustration of the theoretical properties of the kernel density estimate (left panel: bias of the estimate and right panel: MSE as the sum of squared bias and variance)

Two- and multidimensional data

Kernel estimates have the property of a straightforward generalization to multidimensional data. This leads however to less precise estimates (compared to the one-dimensional case), the so-called the "curse of dimensionality". A further issue related to multidimensional data is their graphical representation and interpretation (Figure 4).

NONPARAMETRIC REGRESSION

In a third chapter, a variety of nonparametric regression approaches is discussed. The course starts with Nadaraya-Watson regression as this estimator is directly based on the estimation of one- and two-dimensional kernel densities. Local polynomial estimators can be then introduced as generalizing the local constant Nadaraya-Watson estimator. The main issue on this chapter of the course is however the comparison of the different methods. For that reason, also regressograms (as generalization of the histogram idea), spline smoothing and k-nearest neighbor regression are dealt with in some detail.

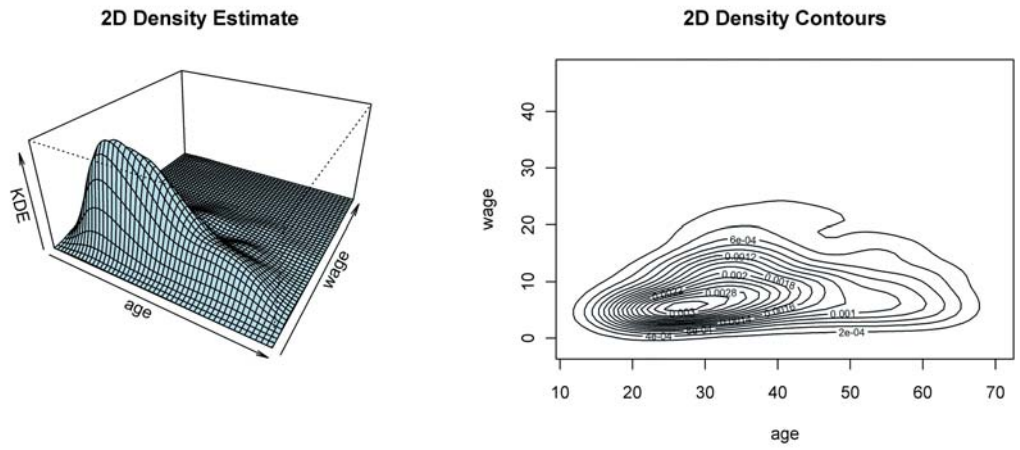

Figure 4. Graphical representation of 2D kernel density estimates
(left panel: perspective plot and right panel: contour lines)

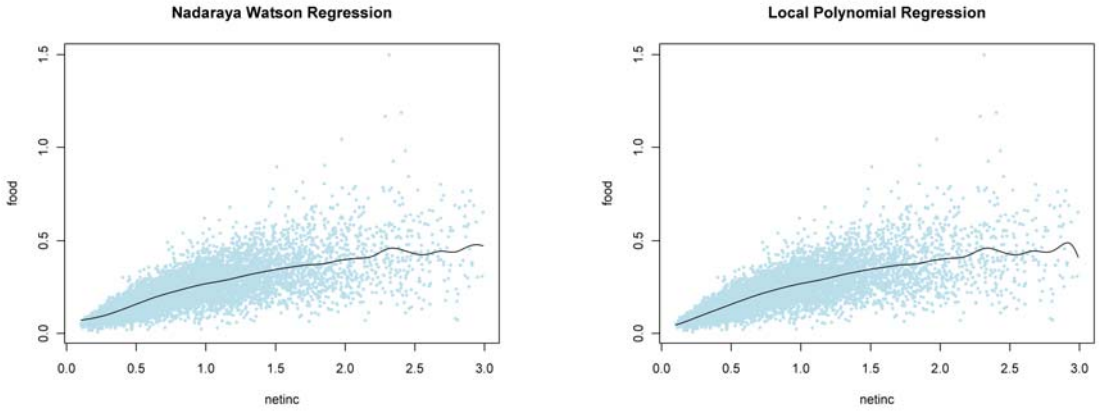

Figure 5. Comparison of Nadaraya-Watson regression (left panel)
versus local linear regression (right panel)

BINARY CHOICE MODELS AND SEMIPARAMETRIC MODELLING

The final part of the course is devoted to an overview on semiparametric models. In particular additive models and generalized additive models, single index models, partial linear and generalized partial linear models are discussed. This part of the course can only be meant as an introduction to the topics, to discuss some important applications and the essential modelling assumptions (for example: identification issues).

Binary choice models make it easy to understand why all these semiparametric models are useful to explore regression data in applications. The example we discuss here comes from field of credit scoring. The available data represent a (stratified) sample for credit applicants. The dependent variable Y is the binary observation of credit default (Y=1) or non-default (Y=0) whereas the explanatory variables describe loan characteristics (amount, maturity, purpose) and socio-economic chararacteristics of the credit applicants (age, employment and wealth variables, information on previous loans etc.)

Figure 6 shows a graphical visualization of the effects of the variables Age and Amount on the linear predictor in the classical logit fit

$$P(Y=1) = F(\beta_0+\beta_1 Age+\beta_2 Age^2+\beta_3 Amount+\beta_4 Amount^2+\ldots),$$

with F denoting the logistic cumulative distribution function given by $F(u)=1/\{1+\exp(-u)\}$. As an example for a semiparametric alternative, a generalized additive model using the logistic link

function is fitted available from the R package `mgcv`. Both additive component functions are here fitted by splines with smoothing parameters are automatically optimized within the `mgcv::gam` procedure. The fitted model is written as

$$P(Y=1) = F(\beta_0+m_1(Age)+m_2(Amount)+\ldots).$$

where $m_1(Age)$ and $m_2(Amount)$ denote two nonparametrically estimated additive component functions. Figure 7 shows the resulting estimates (together with confidence bands).

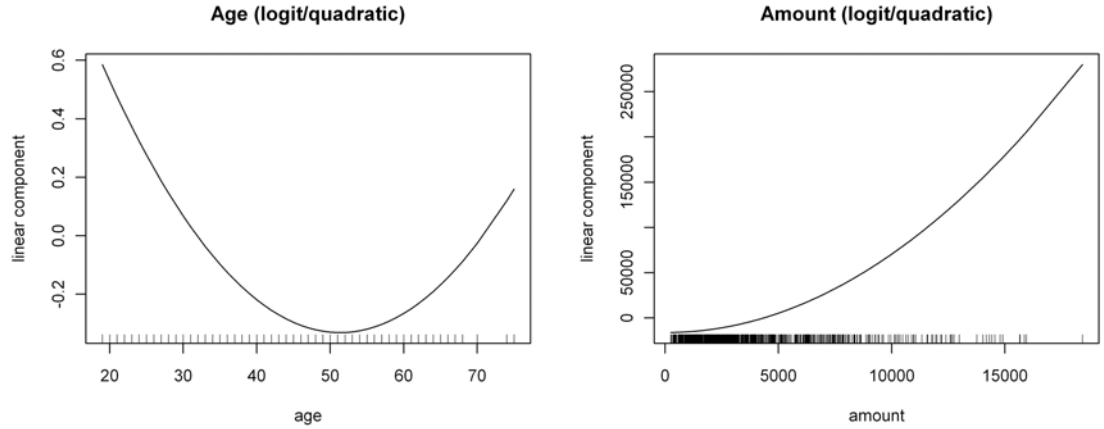

Figure 6. The effect of the variables Age and Amount in the linear predictor of the logit credit scoring model (both variables included with linear and quadratic terms)

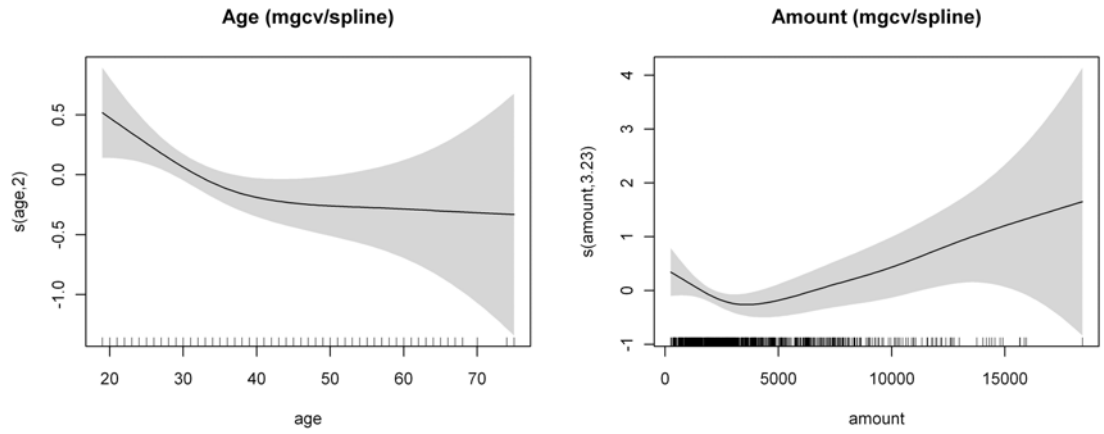

Figure 7. The effect of the variables Age and Amount in the predictor of the semiparametric logit credit scoring model (both variables included as spline fitted components)

CONCLUSION

We propose a one-semester course that introduces to the most important concepts of nonparametric function estimation. The intention of this course is that students of applied sciences can use these techniques (which are freely available in R, see www.R-project.org) to explore their data in a more flexible way and are also able to assess these methods for their applicability. The course is complemented by R scripts which can be individually altered or modified in order to see more features of the proposed estimators as well as to apply the methods on other data sets. The course follows in large parts the contents given in Härdle, Müller, Sperlich and Werwatz (2004). For additional relevant and complementary resources we refer to the following list of references.

REFERENCES

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

Härdle, W. (1990). *Applied Nonparametric Regression. Econometric Society Monographs No. 19.* Cambridge University Press.

Härdle, W. (1991). *Smoothing Techniques, With Implementations in S.* New York: Springer.

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

Hastie, T. J., & Tibshirani, R. J. (1990). Generalized Additive Models. London: Chapman and Hall.

Müller, M. (2009). *R material for "Nonparametric and Semiparametric Models"*. Online: www.marlenemueller.de/nspm.html.

Pagan, A., & Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge University Press.

R Development Core Team (2010). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. Online: www.R-project.org.

Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

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

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis. Vol. 26 of Monographs on Statistics and Applied Probability*. London: Chapman and Hall.

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

Wood, S. N. (2006). *Generalized Additive Models: An Introduction with R.* London: Chapman and Hall.

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