Data–dependent parameter choice for a conditional kernel density estimate.

Ann-Kathrin Bott* and Michael Kohler.
Fachbereich Mathematik, Technische Universität Darmstadt, Schlossgartenstr. 7, 64289 Darmstadt, Germany, email: abott@mathematik.tu-darmstadt.de, kohler@mathematik.tu-darmstadt.de

Abstract

In this paper we present a method to choose the parameters of a conditional kernel density estimate. This method is based on the combinatorial method for choosing the parameters of a density estimate. In contrast to most literature we consider conditional density estimation in L_1. We derive a theoretical result about the quality of the proposed method and illustrate the performance of the estimate for finite sample size by using simulated data.

Keywords: Conditional density estimation; L_1–error; bandwidth selection.

1. Introduction

One major problem in statistics is the estimation of a distribution from a given sample. The Lemma of Scheffé directly links this problem to density estimation in L_1. Therefore, a L_1–consistent density estimate enables a consistent estimation of the probability of all sets by the corresponding distribution estimate. Most density estimates depend on parameters. For instance, the histogram estimate depends on the partition of $\mathbb{R}^d$ and the Rosenblatt-Parzen kernel density estimate (cf., e.g., Rosenblatt (1956) and Parzen (1962)) depends on a bandwidth. Considering a finite sample the parameter choice is of great interest. Let $Z_1, \ldots, Z_n$ be an independent sample of an $\mathbb{R}^d$–valued random variable Z with density f. Moreover we assume that a class of density estimates $(f_{n,\theta})_{\theta \in \Theta}$ is given. Now we want to choose a parameter $\hat{\theta} \in \Theta$ such that

$$\int |f(x) - f_{n,\hat{\theta}}(x)| \, dx \approx \inf_{\theta \in \Theta} \int |f(x) - f_{n,\theta}(x)| \, dx.$$

Due to the fact that f is unknown, the L_1–error cannot be determined. This raises the question how to select parameters in order to minimize the L_1–error. Typically this question is considered in the literature in connection with the L_2–error, see, e.g., Stone (1984), Hall et al. (1991) and the literature cited therein. But much less is known concerning adaptation result in connection with the L_1–error. In this respect Devroye and Lugosi (1996) introduced the so called combinatorial method to choose the parameters of a density estimate in dependence of the given sample. At first, the sample is split into testing data $Z_1, \ldots, Z_m$ and learning data $Z_{m+1}, \ldots, Z_n$ for $0 < m \leq \lfloor n/2 \rfloor$. The learning data is used to define the estimate which is denoted by $f_{n-m,\theta}(\cdot) = f_{n-m,\theta}(\cdot\,, Z_{m+1}, \ldots, Z_n)$. The empirical distribution function of the testing data is defined as

$$\mu_m(A) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_A(Z_i) \quad (A \in \mathcal{B}_d).$$

The combinatorial method chooses the parameter $\hat{\theta} \in \Theta$ for which the expression

$$\Delta_\theta = \sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,\theta}(x) \, dx - \mu_m(A) \right| \tag{1}$$

is minimal, where $\mathcal{A}$ denotes the Yatracos class of subsets of $\mathbb{R}^d$, given by

$$\mathcal{A} = \left\{ \left\{ x \in \mathbb{R}^d : \, f_{n-m,\theta_1}(x) > f_{n-m,\theta_2}(x) \right\} \, : \theta_1, \theta_2 \in \Theta \right\}.$$

Devroye and Lugosi (1996) showed that the L_1–error of the resulting estimate $f_{n-m,\hat{\theta}}$ is linked to the L_1–error with the optimal parameter choice. If $\int f_{n-m,\theta}(x) \, dx = 1$ for all $\theta \in \Theta$, it holds

$$\int |f(x) - f_{n-m,\hat{\theta}}(x)| \, dx \leq 3 \cdot \inf_{\theta \in \Theta} \int |f(x) - f_{n-m,\theta}(x)| \, dx + 4\Delta + \frac{3}{n}, \tag{2}$$

where

$$\Delta = \sup_{A \in \mathcal{A}} \left| \int_A f(x)\,dx - \mu_m(A) \right|.$$

If the condition $\int f_{n-m,\theta}(x)\,dx = 1$ is not fulfilled for all $\theta \in \Theta$, the statement also holds but with factor "5" instead of "3". In addition, Devroye and Lugosi (1997a) derived upper bounds for $\mathbf{E}\{\Delta\}$ by combinatorial tools. For a suitable choice of m the last two summands are asymptotically negligible. Hence, the L_1–error can be bounded by a multiple of the L_1–error of the estimate with the optimal bandwidth. In Chapter 11 of Devroye and Lugosi (2001) concrete results for the classes of kernel density estimates are summarised. A comparison to other methods and simulation results are given in Devroye and Lugosi (1997b).

In this paper we deal with conditional density estimation. Here, one is interested in the conditional density of a random variable Y given a random vector X. This problem can be seen as generalization of regression. One is interested in the full density rather than in the expected value. In conditional density estimation it is usually assumed that a sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) is available. Already in Rosenblatt (1969) the kernel estimate of a conditional density was introduced. But it first received serious attention in Fan et al. (1996) and Hyndman et al. (1996). This estimate is defined by

$$\hat{f}_{Y|X}(y, x) = \frac{\sum_{i=1}^n K\left(\frac{\|x - X_i\|}{H_n}\right) \cdot K\left(\frac{|y - Y_i|}{h_n}\right)}{h_n \sum_{j=1}^n K\left(\frac{\|x - X_j\|}{H_n}\right)}. \tag{3}$$

The question now arises how to adaptively select the bandwidths. Literature only deals with methods concerning the L_2–error. Various attempts start with choosing the bandwidth h_n by referencing rules and afterwards select H_n by known methods for kernel regression estimate. Fan et al. (1996) choose h_n by the normal referencing rule of Silverman (1986) and H_n by the residual squares criterion (Fan and Gijbles (1996)). Also Bashtannyk and Hyndman (2001) first apply bandwidth rules based on a reference distribution to determine one of the bandwidths and then apply regression based bandwidth selectors to determine the second one. These methods use strong assumptions on the distributions and therefore work only well in a limited number of cases. Hall et al. (1999) proposed a bootstrap method, that works well for polynomial regression models. Also Bashtannyk and Hyndman (2001) considered this approach and extended the method. Fan and Yim (2004) proposed a method without restrictive assumptions. They choose the bandwidth by cross-validation. While the first mentioned ad-hoc methods can be efficiently calculated, they perform poorly on finite samples for most distributions. On the other hand the bootstrap method and cross-validation method are time-consuming but more reliable. Holmes et al. (2010) try to balance between both aspects and proposed a likelihood cross-validation method.

In this paper we derive and analyze a data dependent method to choose the bandwidths $h_n, H_n > 0$ of a conditional kernel estimate without any assumptions on the distribution of (X, Y). This method is motivated by the above mentioned combinatorial method of Devroye and Lugosi (1996). Since we do not estimate one single density, we transform Δ_θ such that the resulting adaptive estimate is an appropriate estimate of $f(\cdot, x)$ for $\mathbf{P}_X$-almost all $x \in \mathbb{R}^d$. The main difficulty here is that we estimate simultaneously $f(\cdot, X_i)$ for $i \in \{1, \ldots, n\}$, where for each i we have available only a sample of size one which we cannot split into learning and testing data.

Since we are interested in an estimation of the conditional distribution of Y given X, we measure the quality of the adaptive estimate by the L_1–error. More precisely, we consider the average L_1–error

$$\int \int |f_n(y, x) - f(y, x)|\,dy\,\mathbf{P}_X(dx),$$

and we show that the expected average L_1–error of our newly proposed adaptive estimate is (up to a term of order $\sqrt{\log(n)}/\sqrt{n}$) less than or equal to five times the expected L_1–error which we would get if we would be able to choose the bandwidth in an optimal way (which is never possible in an application).

Throughout the paper the following notation is used: The sets of natural numbers, integers, real numbers and positive real numbers including zero are denoted by $\mathbb{N}, \mathbb{Z}, \mathbb{R}$ and $\mathbb{R}_+$, respectively. $\mathcal{B}_d$ denotes the set of all Borel sets in $\mathbb{R}^d$ and $\mathbb{1}_B$ denotes the indicator function of the set B. $\|x\|$ is the Euclidean norm of a

vector $x \in \mathbb{R}^d$. For a real number z we denote by $\lfloor z \rfloor$ and $\lceil z \rceil$ the largest integer less than or equal to z and the smallest integer larger than or equal to z, respectively.
The outline of this paper is as follows: The main results are presented in Section 2. Section 3 illustrates the finite sample size behavior of our estimate by applying it to simulated data.

2. Main Result

We assume that an independent and identically distributed sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ of an $\mathbb{R}^d \times \mathbb{R}$–valued random vector (X, Y) is available. We select simultaneously the bandwidths $h_n, H_n > 0$ of our estimate

$$f_n(y, x) = \frac{\sum_{i=1}^n K\left(\frac{\|x - X_i\|}{H_n}\right) K\left(\frac{|y - Y_i|}{h_n}\right)}{h_n \sum_{j=1}^n K\left(\frac{\|x - X_j\|}{H_n}\right)}$$

where $K(x) = \frac{1}{2} \cdot \mathbb{1}_{[-1,1]}(x)$ is the naive kernel. At first we choose a parameter set

$$\mathcal{P}_n \subseteq \left\{(h, H) \in \mathbb{R}^2 \middle| \, h \in [1/n,\, n],\, H > 0\right\}.$$

Now we split the data samples into two halves. The second half of the data $(X_{\lfloor n/2 \rfloor + 1}, Y_{\lfloor n/2 \rfloor + 1}), \ldots, (X_n, Y_n)$ is the so called learning data and is used to define our estimate:

$$\hat{f}_\theta(y, x) = \frac{\sum_{i=\lfloor n/2 \rfloor + 1}^n K\left(\frac{\|x - X_i\|}{H}\right) K\left(\frac{|y - Y_i|}{h}\right)}{h_n \sum_{j=\lfloor n/2 \rfloor + 1}^n K\left(\frac{\|x - X_j\|}{H}\right)}$$

with $\theta = (h, H)$. On the basis of the first half of the data (testing data) we evaluate our estimator and choose the parameters. Our goal is to select $\hat{\theta} \in \mathcal{P}_n$ such that the average L_1–error of the corresponding estimate $\hat{f}_{\hat{\theta}}$ is small. We select $\hat{\theta} = (\hat{h}, \hat{H})$ through minimizing

$$\Delta_\theta = \sup_{\theta_1, \theta_2 \in \mathcal{P}_n} \left| \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \int_{A_i(\theta_1, \theta_2)} \hat{f}_\theta(y, X_i)\, dy - \frac{1}{\lfloor n/2 \rfloor} \sum_{i=1}^{\lfloor n/2 \rfloor} \mathbb{1}_{A_i(\theta_1, \theta_2)}(Y_i) \right|,$$

where

$$A_i(\theta_1, \theta_2) = \left\{ y \in \mathbb{R} \,:\, \hat{f}_{\theta_1}(y, X_i) > \hat{f}_{\theta_2}(y, X_i) \right\}.$$

If the minimum does not exist, we choose $\hat{\theta} = (\hat{h}, \hat{H}) \in \mathcal{P}_n$ such that

$$\Delta_{\hat{\theta}} < \inf_{\theta \in \mathcal{P}_n} \Delta_\theta + \frac{1}{n}.$$

Δ_θ is motivated by (1). But here we consider the arithmetic mean of estimated L_1–errors. And since we estimate a whole class of densities, we need to regard a whole class of Yatracos sets, which are linked by the parameters $\theta_1, \theta_2 \in \mathcal{P}_n$.
The following theorem bounds the expected average L_1–error of this estimate by that of the estimate with optimal parameter choice.

Theorem 1 *Let $\hat{f}_{\hat{\theta}}$ be the above introduced estimate. It holds for all $n > 1$*

$$\begin{aligned}
&\mathbf{E}\left\{ \int \int |\hat{f}_{\hat{\theta}}(y, x) - f(y, x)|\, dy\, \mathbf{P}_X(dx) \right\} \\
\leq & 5 \cdot \inf_{\theta \in \mathcal{P}_n} \mathbf{E}\left\{ \int \int |\hat{f}_\theta(y, x) - f(y, x)|\, dy\, \mathbf{P}_X(dx) \right\} + \frac{2}{n} + 116 \sqrt{\frac{\log n}{\lfloor n/2 \rfloor}} + \frac{306}{\sqrt{\lfloor n/2 \rfloor \cdot \log n}}.
\end{aligned}$$

Proof. See Bott and Kohler (2015).

Remark 1. This theorem states that the expected average L_1–error of the proposed estimate lies close to five times the least possible. Here we have a factor of "5" instead of "3", since our estimate $\hat{f}_{\hat{\theta}}$ is (possibly) no density for all $x \in \mathbb{R}^d$.

Remark 2. Due to the splitting of the sample we compare the quality of our estimate to that of an estimate using also only half of the data. It is an open problem to show that

$$\inf_{\theta \in \mathcal{P}_n} \mathbf{E}\left\{ \int \int |\hat{f}_\theta(y,x) - f(y,x)| \, dy \, \mathbf{P}_X(dx) \right\}$$

with $\hat{f}_\theta$ using half of the data is not much larger than with $\hat{f}_\theta$ using all of the data. Devroye and Lugosi (1997b) addressed this problem in case of density estimation (c.f., Devroye and Lugosi (1997b) and Theorem 10.3 in Devroye and Lugosi (2001)) .

Remark 3. By Theorem 1 a non-asymptotic upper bound of the expected average L_1–error is given. As we did not attempt to minimize the constants, the constants of the last two summands could potentially be much smaller.

3. Simulation

In this section we illustrate the performance of our estimator for finite sample size and a finite parameter set $\mathbf{h} \times \mathbf{H}$. We consider one example for three different sample sizes $n = 200, 500, 1000$. We compare the results to those of a conditional kernel estimate with cross-validated bandwidths like in Fan and Yim (2004). We evaluate the performance of both selection rules by the average L_1–error. The proposed estimate splits the data into learning and testing data. In Section 2 we assumed that $N = \lfloor n/2 \rfloor$ points were used to test the estimate and $n - N = \lceil n/2 \rceil$ data points to construct the estimate (new1). In addition we consider the proposed estimate with $N = \lfloor n/4 \rfloor$ testing data points and $n - N = \lceil 3n/4 \rceil$ learning data points (new2). To get an impression how small the average L_1–error could be under these circumstances, we compare our results to the estimate with n data points and the optimal parameter choice out of $\mathbf{h} \times \mathbf{H}$. In applications the underlying distribution is unknown and thus, this estimator is not applicable. In the implementation we approximate all integrals by Riemann sums. Here, $n = 200, 500, 1000$ independent copies of (X, Y) are sampled, where X is uniformly distributed on $[0, 2]$ and Y is exponentially distributed with a rate depending on the covariate. More precise,

$$Y \sim Exp(\lambda) \quad \text{with } \lambda = 0.25 + X \text{ and } X \sim \mathcal{U}[0, 2].$$

The bandwidths are selected out of finite parameter sets

$$h \in \mathbf{h} = \{0.08,\ 0.11,\ 0.16,\ 0.23,\ 0.33\, 0.48,\ 0.69\, 0.98,\ 1.40,\ 2.00\}$$
$$H \in \mathbf{H} = \{0.02,\ 0.03,\ 0.06,\ 0.09,\ 0.16,\ 0.26,\ 0.43,\ 0.72,\ 1.20,\ 2.00\}.$$

		opt	**new1**	**new2**	**CV**
n=200	m (sd) L_1–error	0.286 (0.024)	0.397 (0.054)	0.384 (0.063)	0.518 (0.112)
	m (sd) H	0.663 (0.116)	0.703 (0.286)	0.621 (0.418)	2.000 (0.000)
	m (sd) h	0.284 (0.062)	0.202 (0.083)	0.198 (0.102)	0.852 (0.748)
n=500	m (sd) L_1–error	0.234 (0.017)	0.306 (0.038)	0.288 (0.032)	0.492 (0.120)
	m (sd) H	0.504 (0.126)	0.576 (0.172)	0.518 (0.198)	2.000 (0.000)
	m (sd) h	0.232 (0.046)	0.166 (0.058)	0.157 (0.077)	0.813 (0.686)
n=1000	m (sd) L_1–error	0.201 (0.012)	0.265 (0.026)	0.246 (0.028)	0.488 (0.128)
	m (sd) H	0.449 (0.069)	0.500 (0.168)	0.448 (0.185)	2.000 (0.000)
	m (sd) h	0.181 (0.031)	0.140 (0.043)	0.129 (0.054)	0.794 (0.734)

Table 1: Summarised results.

Since the results of our simulation depend on randomly occurring data points, we repeat the whole procedure 100 times. The boxplots in Figure 1 report the average L_1–errors for all four estimates. Mean (m) and standard deviation (sd) of the average L_1–errors as well as mean and standard deviation of the chosen bandwidths are given in Table 1. Here both proposed estimates outperform considerably the cross-validated estimate for all sample sizes. The second version of our estimate (new2) achieves even slightly better results than the first version (new1). Even though the proposed estimates use less data than the cross-validated estimate, the mean bandwidths are smaller. For H the cross–validation always selects the maximal possible bandwidth.

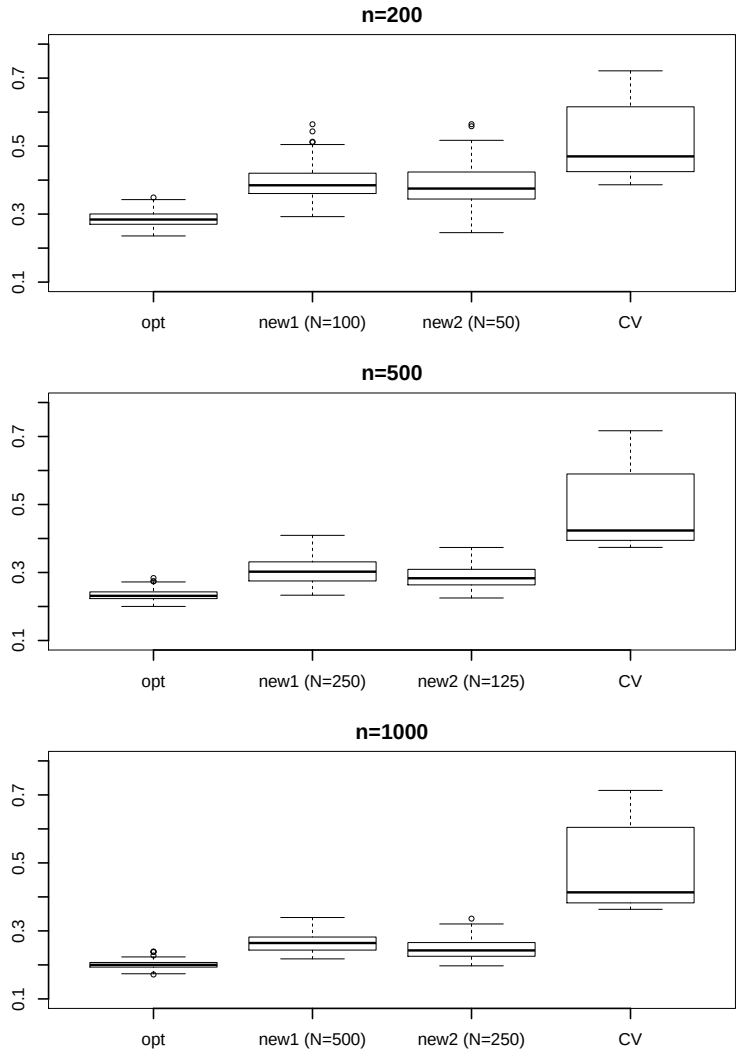

Figure 1: Boxplot of the average L_1–error

4. Acknowledgment
The authors would like to thank the German Research Foundation (DFG) for funding this project within the Collaborative Research Center 666.

References

Bashtannyk, D. M., & Hyndman, R. J. (2001). Bandwidth selection for kernel conditional density estimation.

Computational Statistics and Data Analysis, **36**, pp. 279–298.

Bott, A. and Kohler, M. (2015). Adaptive estimation of a conditional density. Submitted for publication.

Devroye, L. & Lugosi, G. (1996). A universally acceptable smoothing factor for kernel density estimation. Annals of Statistics, **24**, pp. 2499–2512.

Devroye, L. & Lugosi, G. (1997a). Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. Annals of Statistics, **25**, pp. 2626–2637.

Devroye, L. & Lugosi, G. (1997b). Universal smoothing factor selection in density estimation: theory and practice (with discussion). Test, **6**, pp. 223–320.

Devroye, L. & Lugosi, G. (2001). Combinatorial Methods in Density Estimation. Springer-Verlag, New York.

Fan, J. & Gijbels, I. (1996). Local Polynomial Modelling and its Applications. Monographs on Statistics and Applied Probability, Chapman & Hall, London.

Fan, J., Yao, Q. & Tong, H. (1996). Estimation of conditional densities and sensitivity measures in nonlinear dynamical systems. Biometrika, **83**, pp. 189–206.

Fan, J. & Yim, T. H. (2004). A crossvalidation method for estimating conditional densities. Biometrika, **91**, pp. 819–834.

Hall, P., Sheater, S. J., Jones, M. C. & Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. Biometrika, **78**, pp. 263–269.

Hall, P., Wolff, R.C.L. & Yao, Q. (1999). Methods for estimating a conditional distribution function. Journal of the American Statistical Association, **94**, pp. 154–163.

Holmes, M.P.,Gray, A.G. & Isbell Jr, C.L. (2010). Fast kernel conditional density estimation: A dual-tree Monte Carlo approach. Computational statistics & data analysis, **54**, pp. 1707–1718.

Hyndman, R. J., D.M. Bashtannyk & G.K. Grunwald (1996). Estimating and visualizing conditional densities. Journal of Computational and Graphical Statistics, **5**, pp. 315–336.

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

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

Rosenblatt, M. (1969). Conditional probability density and regression estimates. Multivariate Analysis II (Ed. P.R. Krishnaiah), Academic Press, New York pp. 25–31.

Scott, D. W. (1992). Multivariate density estimation: Theory, practice and visualization. John Wiley, New York.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. Monographs on Statistics and Probability, vol. 26. Chapman & Hall, London.

Stone, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. Annals of Statistics, **12**, pp. 1285–1297.