Groß, Marcus; Rendtel, Ulrich

Working Paper

Kernel density estimation for heaped data

Diskussionsbeiträge, No. 2015/27

Provided in Cooperation with:
Free University Berlin, School of Business & Economics

Suggested Citation: Groß, Marcus; Rendtel, Ulrich (2015) : Kernel density estimation for heaped data, Diskussionsbeiträge, No. 2015/27, Freie Universität Berlin, Fachbereich Wirtschaftswissenschaft, Berlin

This Version is available at:
https://hdl.handle.net/10419/117333

Freie Universität Berlin

Kernel Density Estimation for Heaped Data

Marcus Groß
Ulrich Rendtel

School of Business & Economics
Discussion Paper

Economics

2015/27

Kernel Density Estimation for Heaped Data

Marcus Groß, Ulrich Rendtel *

Abstract

In self-reported data usually a phenomenon called 'heaping' occurs, i.e. survey participants round the values of their income, weight or height to some degree. Additionally, respondents may be more prone to round off or up due to social desirability. By ignoring the heaping process a severe bias in terms of spikes and bumps is introduced when applying kernel density methods naively to the rounded data. A generalized Stochastic Expectation Maximization (SEM) approach accounting for heaping with potentially asymmetric rounding behaviour in univariate kernel density estimation is presented in this work. The introduced methods are applied to survey data of the German Socio-Economic Panel and exhibit very good performance simulations.

Keywords: Heaping, Survey Data, Measurement error, Self-reported data, Kernel density estimation, Rounded data
Word count: 5310

1. INTRODUCTION

In survey data the researcher often encounters rounded values when the participants are asked to state metric variables such as income (Hanisch 2007; Czajka and Denmead 2008), household expenditures (Pudney 2008), body weight and height (Taylor et al. 2006), blood pressure (De Lusignan et al. 2004) or working hours (Otterbach and Sousa-Poza 2010). The rounding behaviour of self reported data is usually mixed, i.e. participants may round to multiples of 1, 2, 5, 10, 20, 50, 100.. or may report only two leading digits (Hanisch 2007) . This type of measurement error –when data are collected with various degrees of courseness– is called heaping. Heaping cannot be ignored because it is a well known fact (Heitjan and Rubin 1991; Schneeweiß and Komlos 2009), that if we naively use the self-reported values in

*Institute for Statistics and Econometrics, Freie Universität Berlin, Germany, marcus.gross@fu-berlin.de,ulrich.rendtel@fu-berlin.de

the estimation of a distribution, the estimates are biased. This is especially the case in (non-parametric) kernel density estimation where we observe bumps und spikes at the multiples of the rounding values. The standard methods of choosing the bandwidth are also not very useful in this setting. The Sheather-Jones estimate (Sheather and Jones 1991), which is mostly recommended in literature, produces often completely useless density estimates in self reported data. This is because a pilot estimate of the integral of the second derivative is employed to estimate the bandwidth. Due to the extremely multimodal nature of the heaped data, this plug-in estimate of the integrated second derivative is very large leading to very small bandwidths. Silverman's rule of thumb shows a better behaviour because it implicitly assumes a normal distribution for bandwidth selection but still gives not very satisfying results. Figure 1 shows two examples from a household survey, the German SocioEconomic Panel –'SOEP'– (Wagner et al. 2007) wave BC (2012): body weight of the female participants and monthly food and drink expenditures outside home.

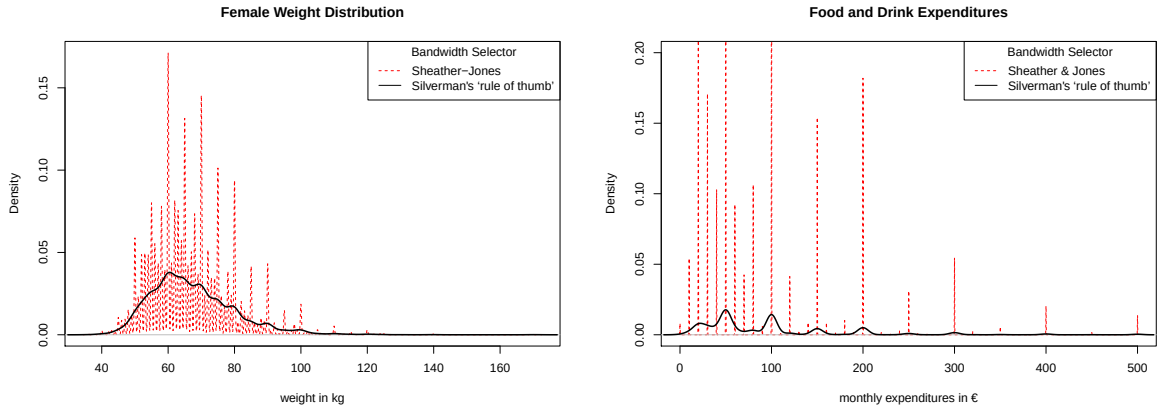

Figure 1: Kernel density estimator applied to self-reported female weight (left) and food and drink expenditures (right) taken from the Socio-Economic Panel 2012. The two popular bandwidth selectors ('Sheater-Jones' and Silverman's 'rule of thumb') show more or less severe spikes at the multiples of the rounding values.

Increasing the bandwidth thus far that the density estimate is sufficiently smooth leads to oversmoothing: the tails of the distribution get too heavy and important features of the distribution may be lost. Additionally, participants may be more prone to round up or down due to social desirability. For self-reported weight measurements with validation data, for example, respondents typically underreport their weight which can be (partially) explained by their tendency to round off (Rowland 1990; Shields et al. 2008; Merrill and Richardson 2009). This work proposes a non-parametric density estimation of self-reported measures in the presence of heaping. The primary goal of this work is to provide a method that reduces

the bias in kernel density estimation and estimates the parameters of the heaping process as well. To the authors best knowledge, this is the first general attempt for this type of problem. A measurement error model is employed to solve it. The model is initially formulated within a Bayesian framework whereby the resulting Gibbs-sampler was modified to a partly Bayesian Stochastic Expectation Maximization (SEM, Celeux et al. 1996) algorithm. Additionally, we show that under certain assumptions it is possible to identify and estimate a rounding direction bias (unequal probability of rounding up and down).

The paper is organized as follows: Section 2.1 provides a literature overview on existing modeling approaches for heaped data. In Section 2.2 introduces a model for the rounding process respectively heaping. After a short introduction to kernel density estimation the measurement error model and it's computational implementation is presented in Section 3. Section 4 provides a simulation study and Section 5 demonstrates an application to self-reported data from the SOEP. A summary with an outlook concludes the article.

2. MODELING HEAPING IN SELF-REPORTED DATA

2.1 Heaping models in applications

Heaping occurs frequently in a variety of applications in quite different fields. Heitjan and Rubin (1990) modeled the heaping process as rounding with different interval length and used a complex imputation model in estimating the age of Tanzanian children. A similar approch was followed by Battistin et al. (2003) in household food expenditures. Wang and Heitjan (2008) proposed a model for heaped cigerate counts. A recent work of Crawford et al. (2014) formulated a general model for count data involving birth-death processes and applied this to the self-reported counts of the number of sex-partners. In addition, Bar and Lillard (2012) lately developed an approach for event time data by modeling the density by a mixture of two parametric distributions. However, in a very recent publication of the DIW (German Institute for Economic Research) dealing with self-reported data from the SOEP (Marcus et al. 2013), a modeling of the heaping process was discarded and a parametric density was naively (without any correction for the rounding process) applied to the reported data. This procedure was justified because of the more or less arbitrary assumptions on the heaping process the researcher has to rely on. The authors disagree with this assessment, because although we might not be able to reproduce the heaping pattern perfectly by the heaping model assumptions, the bias in the parameter estimates may be greatly reduced. Little work has been done in the context of heaping in non-parametric density estimation. One work of Camarda et al. (2008) deals with estimating age-at-death as well as body weight by assuming a smooth underlying density function modeled by B-

splines. However, it was assumed that the true unobserved value was the reported value itself or one of the two immediate neighbouring integers, which is clearly not suitable for other data such as monthly income. The method proposed in this article pursues a more general strategy applicable to a much wider variety of data.

2.2 A model for heaping

For this article the heaping process is modeled as follows: At first one has to assign rounding parameters **r** appropriate for the data. When looking at the SOEP female body weight example, for example, we observe that the most frequent end digit was **0** with 24.6% of the reported cases followed by **5** with 17.4%. Moreover, the respondents seem to prefer even over odd numbers. The end digits **2,4,6,8** are reported in 33.4% of the total cases while the end digits **1,3,7,9** only sum up to 24.5% (see Table 1 for details).

end digit	0	1	2	3	4	5	6	7	8	9
count	2990	569	1184	1006	883	2123	776	717	1229	691
%	24.6	4.7	9.7	8.3	7.3	17.4	6.4	5.9	10.1	5.7

Table 1: End digits of SOEP female body weight in kg.

Therefore, we may choose the rounding values $\boldsymbol{r} = (1, 2, 5, 10)$. In general, suitable potential rounding parameters are $\boldsymbol{r} = (.., 0.5, 1, 2, 5, 10, 20, ..)$ for variables with decimal numeral system (e.g. blood pressure, body weight,..), $\boldsymbol{r} = (1, 2, 3, 6, 12)$ for variables with duodecimal system (e.g. time in months, length in inches,..) and $\boldsymbol{r} = (1, 5, 10, 15, 30, 60)$ for the sexagesimal system (e.g. time in minutes). A probability vector $\boldsymbol{p} = (p_1, .., p_m)$ is assigned to the rounding values $\boldsymbol{r}$ denoting the probability of the respondent to report a value W_i $(i = 1, .., n)$ which is rounded by a value $R_i \in \{r_1, r_2, .., r_m\}$. For the moment $\boldsymbol{p}$ is assumed equal for all respondents and independent from the true, unobserved value X_i. This is a key assumption which is not always met and will be relaxed later. We then assume that the rounding is done correctly such that X_i lies within the interval $(W_i - 1/2R_i, W_i + 1/2R_i)$. As R_i is not uniform over the individuals, we have a heteroscedastic measurement error here.

The model for the heaping process described above may not fit very well to all kinds of data. Thus, we consider two extensions. As already mentioned the respondents may more likely round down than round up or vice versa. A first suggestion is to define a parameter $a \in (0, 1)$ allocating the probability of rounding down. However, when imposing the restriction $X_i \in (W_i - 1/2R_i, W_i + 1/2R_i)$ (rounding mathematically correct) it is not possible to choose the rounding direction independently from R_i and X_i. Consider the true value $X_i = 77.8$, rounding values $\boldsymbol{r} = (1, 10)$ and assume mathematically correct rounding

behaviour the respondent has to round up in any case regardless of his chosen rounding value R_i. We therefore introduce an alternative concept. We extend R_i such that it includes the rounding direction: $R_i \in \{-r_1, .., -r_m, +r_1, .., +r_m\}$ whereby negative values indicate a rounding up and positive values a rounding down. The rounding probabilities $\boldsymbol{p}$ are multiplied by a when rounding down ($R_i > 0$) and by $(1-a)$ when rounding up ($R_i < 0$) if the combination of R_i and X_i is compliant with the assumption of correct rounding (i.e. $X_i \in (W_i, W_i + 1/2R_i)$ for $R_i > 0$ and $X_i \in (W_i + 1/2R_i, W_i)$ for $R_i < 0$) and are set to 0 else. They are scaled afterwards such that the probabilities for all R_i sum up to 1. We give two numerical examples how the conditional probability distribution $\pi(R_i|X_i, \boldsymbol{p}, a)$ denoted as $\pi(R_i|\cdot)$ is modeled:

- First, consider $\boldsymbol{r} = (1, 10)$, $\boldsymbol{p} = (0.4, 0.6)$ and $a = 0.8$. The respondent's true value is $X = 12.6$. Note that $W_i = 12$ and $W_i = 20$ are not compatible with mathematical rounding. Possible reported values are $W_i = 13$ (rounding up by $R_i = 1$) and $W_i = 10$ (rounding down by $R_i = 10$). It follows that $(\pi(R_i = -1|\cdot), \pi(R_i = -10|\cdot), \pi(R_i = 1|\cdot), \pi(R_i = 10|\cdot)) \propto ((1 - 0.8) \cdot 0.4, 0, 0, 0.8 \cdot 0.6) = (0.08, 0, 0, 0.48)$. Consequently, $P(W_i = 13|\cdot) = 1/7$ and $P(W_i = 10|\cdot) = 6/7$.

- A little more complex example would be the following: Let $\boldsymbol{r} = (1, 2, 5, 10)$, $\boldsymbol{p} = (0.4, 0.3, 0.2, 0.1)$ and $a = 0.15$. For $X_i = 23.4$, possible reported values are $W_i = 23$ (rounding down by $R_i = 1$), $W_i = 24$ (rounding up by $R_i = 2$), $W_i = 25$ (rounding up by $R_i = 5$) and $W_i = 20$ (rounding down by $R_i = 10$). The conditional probabilities $(\pi(R_i = -1|\cdot), \pi(R_i = -2|\cdot), \pi(R_i = -5|\cdot), \pi(R_i = -10|\cdot), \pi(R_i = 1|\cdot), \pi(R_i = 2|\cdot), \pi(R_i = 5|\cdot), \pi(R_i = 10|\cdot))$ are proportional to $(0, 0.3 \cdot (1-0.15), 0.2 \cdot (1-0.15), 0, 0.4 \cdot 0.15, 0, 0, 0.1 \cdot 0.15) = (0, 0.255, 0.17, 0, 0.06, 0, 0, 0.015)$. Thus, $P(W_i = 23|\cdot) = 0.12$, $P(W_i = 24|\cdot) = 0.51$, $P(W_i = 25|\cdot) = 0.34$ and $P(W_i = 20|\cdot) = 0.03$.

In general, with direction parameter $a \in (0, 1)$ the conditional probability distribution of R_i given X_i, $\boldsymbol{p}$ and a is proportional to the following expression:

$$\pi(R_i = \pm r_j|X_i, \boldsymbol{p}, a) \propto a^{I(R_i>0)} \times (1-a)^{I(R_i<0)} \times p_1^{I(R_i=-r_1)} \times .. \times p_m^{I(R_i=+r_m)} \\ \times I(sgn(X_i \bmod (|R_i|) - \frac{1}{2}|R_i|) = -sgn(R_i))$$

The second line serves as a check whether the combination of X_i and R_i is compatible with the restriction of mathematically correct rounding.

The value a can be interpreted as the tendency to round off ($a > 0.5$) or to round up ($a < 0.5$). The reason to restrict to mathematically correct rounding is that it allows us

identify the rounding direction parameter a solely by the end digit pattern. In the simple example of a flat density, $a > 0.5$ and rounding values $\boldsymbol{r} = (1, 10)$ one would observe the end digits **1** to **4** less often than **6** to **9** (or the other way around for $a < 0.5$). This is because the respondent is only able to round down by $R_i = 10$ if $X_i \bmod 10 \in (0, 5)$ and round up by $r = 10$ if $X_i \bmod 10 \in (5, 10)$ with the result that for $a > 0.5$ most reported values W_i with end digit **0** correspond to a true value X_i with $X_i \bmod 10 \in (0, 5)$. The end digit pattern gets more complicated for more than two rounding values. In the SOEP female body weight example however, the left neighbours (**9**, **4**) of end digits **0** and **5**; show significant higher counts (691 to 569 and 883 to 776) than their right counterparts (**1**, **6**) indicating a tendency to round off.

A second extension allows for non-constant rounding probabilities. For example, the probability of a respondent with a true income of $X_i = 1600$ to choose $R_i = 1000$ (and round up to $W_i = 2000$) might be much lower than for someone earning 8600 (and report 9000). A natural choice would be to implement an ordered probit (or logit) model for the rounding probabilities $\boldsymbol{p}$ (as already done in Heitjan and Rubin 1990) with the logarithm of the true value as independent variable:

$$g_i = log(X_i)\beta + \epsilon_i \ , \ \epsilon_i \sim N(0, 1)$$

$\mathbf{g}$ denotes the latent continuous variable and we define $\boldsymbol{\tau} = (\tau_0, \tau_1, .., \tau_m)$ as threshold parameters with $\tau_0 = -\infty$ and $\tau_m = +\infty$. The value p_j $(j = 1, .., m)$ for respondent i is then defined as:

$$\begin{aligned} p_{ij} &= P(\tau_{j-1} < g_i \leq \tau_j) \\ &= \Phi(\tau_j - log(X_i)\beta) - \Phi(\tau_{j-1} - log(X_i)\beta) \end{aligned}$$

The rounding probabilities $\boldsymbol{p}$ may also depend on other characteristics of the respondents and can be introduced in the ordered probit regression formula as well. For $a = 0.5$ and $\beta = 0$ the extended model reduces to the standard rounding model.

3. METHODS

3.1 Kernel density estimation

Kernel density estimation as a non-parametric approach for density estimation is an important tool in exploratory data analysis. Let $\boldsymbol{X} = (X_1, X_2, .., X_n)$ denote a sample of size n from a random variable with density f. The univariate kernel density estimate at point x is

given by:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where $K(\cdot)$ is kernel function and h denotes a bandwidth, which governs the smoothness of the density estimate. The kernel $K(\cdot)$ satisfies regularity conditions such as (a) $\int K(x)dx = 1$, (b) $\int xK(x)dx = 0$ and (c) $\int x^2K(x)dx < \infty$ (Scott 2009). The performance of a kernel density estimator is mainly affected by the particular choice of h (cf. Izenman 1991). Popular strategies to choose h are by minimizing the AMISE (Asymptotic Mean Integrated Squared Error) through plug-in- or cross-validation methods (cf. Izenman 1991 or Silverman 1986). Sheather (2004) gives a short overview in kernel density estimation, kernels and bandwidth choice methods. Unfortunately, the utilization of kernel density estimation methods with heaped data leads to severely biased estimates as already demonstrated in the introduction.

3.2 Model

The Bayesian approach to measurement error problems is to treat the unknown true values X_i as latent variables respectively parameters to be estimated (Carroll et al. 2010). Then the Likelihood can be split into two parts. We specify the following models: First, a measurement error model and second a model which assumes that all (latent) variables are observed. The distribution of $\boldsymbol{X}$ can be modeled parametrically (e.g. by a Gaussian with $\boldsymbol{\theta} = (\mu, \sigma)$) or by a non-parametric formulation. In the Bayesian case the latter alternative can be realized by a mixture of parametric distributions (Escobar and West 1995) or by kernel density estimation through likelihood cross-validation (Zhang et al. 2006) with $\theta = h$. Together with a hyperpriors for $\boldsymbol{p}$ and $\boldsymbol{\theta}$, the posterior distribution can be formulated using a hierarchical model (Carroll et al. 2010).

We start with the heaping (or measurement error) model without extensions:

$$\pi(\boldsymbol{X}, \boldsymbol{R}, \boldsymbol{\theta}, \boldsymbol{p}|\boldsymbol{W}) \propto \underbrace{\pi(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{R}) \times \pi(\boldsymbol{R}|\boldsymbol{p}) \times \pi(\boldsymbol{X}|\boldsymbol{\theta})}_{\text{Likelihood}} \times \underbrace{\pi(\boldsymbol{p})\pi(\boldsymbol{\theta})}_{\text{Priors}} \tag{2}$$

$$L(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{R}, \boldsymbol{\theta}, \boldsymbol{p}) = \prod_{i=1}^{n} \left(\underbrace{\pi(W_i|X_i, R_i) \times \pi(R_i|\boldsymbol{p})}_{\text{Measurement error model}} \right) \times \underbrace{\pi(\boldsymbol{X}|\boldsymbol{\theta})}_{\text{Observation model}} \tag{3}$$

The measurement error model consists of two parts. W_i only depends on X_i and R_i and we

can write it's distribution as a Dirac distribution:

$$\pi(W_i|X_i, R_i) = \begin{cases} 1 & \text{for } X_i \in (W_i - \frac{1}{2}R_i, W_i + \frac{1}{2}R_i) \\ 0 & \text{else} \end{cases},$$

By definition of our heaping model in Section 2.2, $\pi(R_i|\boldsymbol{p})$ follows a multinomial distribution. In order to implement the two extensions proposed in section 2.2 we have to introduce the parameters a, $\boldsymbol{\tau}$ (as threshold value for $\boldsymbol{p}$) as well as β into our likelihood respectively our measurement error model:

$$L(\boldsymbol{W}|\boldsymbol{X}, \boldsymbol{R}, \boldsymbol{\tau}, \boldsymbol{\theta}, a, \beta) = \left(\prod_{i=1}^{n} \underbrace{\pi(W_i|X_i, R_i) \times \pi(R_i|X_i, \boldsymbol{\tau}, a, \beta)}_{\text{Measurement error model}} \right) \times \underbrace{\pi(\boldsymbol{X}|\boldsymbol{\theta})}_{\text{Observation model}} \tag{4}$$

with

$$\pi(W_i|X_i, R_i) = \begin{cases} 1 & \text{for } R_i > 0 \text{ and } X_i \in [W_i, W_i + \frac{1}{2}R_i) \\ 1 & \text{for } R_i < 0 \text{ and } X_i \in (W_i - \frac{1}{2}R_i, W_i) \\ 0 & \text{else} \end{cases},$$

and (cf. section 2.2)

$$\begin{aligned} \pi(R_i = \pm r_j|X_i, \boldsymbol{\tau}, a, \beta) \propto\ & a^{I(R_i<0)} \times (1-a)^{I(R_i>0)} \\ & \times (\Phi(\tau_1 - log(X_i)\beta) - \Phi(\tau_0 - log(X_i)\beta))^{I(R_i=-r_1)} \\ & \times .. \\ & \times (\Phi(\tau_m - log(X_i)\beta) - \Phi(\tau_{m-1} - log(X_i)\beta))^{I(R_i=+r_m)} \\ & \times I(sgn(X_i \bmod |R_i| - \frac{1}{2}|R_i|) = -sgn(R_i)) \end{aligned}$$

After we have specified the measurement error model the distribution of $\boldsymbol{X}$ has to be specified. As this paper deals with kernel density estimation $\pi(\boldsymbol{X}|\boldsymbol{\theta})$ is defined by

$$\pi(\boldsymbol{X}|h) = \prod_{i=1}^{n} \hat{f}_{h,i}(X_i)$$

, where $\hat{f}_{h,i}(X_i)$ denotes the leave one out kernel density 'estimator' (Härdle and Scott 1992;

Zhang et al. 2006). [1]

Now we could place priors on h, a, $\boldsymbol{p}$ or $\boldsymbol{\tau}$, β and simply set up a Gibbs-sampler and sample alternately from the full conditional posteriors $\pi(X_i, R_i|\cdot)$, $\pi(\boldsymbol{p}|\cdot)$ or $\pi(a, \boldsymbol{\tau}, \beta|\cdot)$ and $\pi(h|\cdot)$. While this is completely feasible in theory one may face difficulties when applying computational intense methods like the likelihood cross-validation approach of Zhang et al. (2006) to rather large datasets as in our application example. However, thanks to the convenient hierarchical structure of our likelihood with the result that $\pi(h|\cdot)$ does not depend on W we propose to use a point estimate of h respectively the distribution of X within the Gibbs-Sampler to circumvent computational issues. As a consequence, the proposed estimator is a partly Bayesian method in the sense that the X_i as well as $\boldsymbol{p}$, a, β and $\boldsymbol{\tau}$ are treated as random variables but not $\boldsymbol{\theta}$. As already discussed in Groß et al. (2015) this approach is equal to a generalized Stochastic Expectation Maximization (SEM) algorithm (Celeux et al. 1996). This algorithm is strongly related to the Gibbs-sampler but usually converges much faster (Diebolt et al. 1994). In the context of non-parametric kernel density estimation, this approach enables us to use any bandwidth selection method from the rich variety available in literature and to avoid the computational intense likelihood cross-validation method. As discussed in the next section, Gibbs-sampler and Metropolis-Hastings steps are introduced into the S-step of the algorithm (cf. Diebolt et al. 1994).

3.3 Computational details

As argued in the previous subsection, we replace the full conditional distribution of h by the Sheater-Jones bandwidth selection or Silverman's rule of thumb and define the distribution of X_i given h by the kernel density 'estimator' defined in equation (1). We first consider the case without extensions for the joint full conditional distributions of X_i and R_i (given the rounded values W_i, the rounding parameters $\boldsymbol{p}$, and bandwidth h):

$$\pi\left(X_i, R_i|W_i, h, \boldsymbol{p}\right) \propto I(W_i - \frac{1}{2}R_i \leq X_i \leq W_i + \frac{1}{2}R_i) \times p_j \times \hat{f}_h(X_i),$$

Obviously, the full conditional distribution of X_i, R_i is the product of a uniform distribution on the interval with length R_i around W_i, the probability p_j of rounding to a certain degree of courseness r_j and the kernel density 'estimator' $\hat{f}_h(X_i)$ (equation 1). The conditional

[1] Note that the expression 'kernel density estimator' is ambiguous here as in this context it should be merely called 'kernel density'. However, as we think that a second definition of a kernel density f_h which would be equal to $\hat{f}_h$ could be even more confusing we quote the word 'estimator' when actually referring to a 'kernel density'.

distribution of $\boldsymbol{p}$ is the Dirichet distribution $Dir(\alpha)$:

$$\pi(\boldsymbol{p}|\boldsymbol{R}) \sim Dir(\#(\boldsymbol{R} = r_1), \ldots, \#(\boldsymbol{R} = r_m))$$

Next we consider the case of the two extensions of the heaping model. We could use a modified expression for the joint conditional distribution of X_i an R_i but no established distribution was found for the joint conditional distribution of $\boldsymbol{\tau}$, a and β. A Metropolis-Hastings step turned out to be computational cumbersome because of very slow convergence with the result that a Laplace normal approximation of the joint full conditional distribution $\pi(\boldsymbol{\tau}, a^*, \beta|\cdot)$ was utilized instead, where the parametrization $a^* = \Phi^{-1}(a)$ was used for the reason of computational convenience.

As a consequence a generalized SEM algorithm is proposed, sampling from the full conditional distributions of (X_i, R_i) as well as from (an approximation of) the full conditional distributions of $(\boldsymbol{\tau}, a, \beta)$ in the S-step(which replaces the E-step in the EM-algorithm) and a convenient point estimate for h as a surrogate of sampling from the full conditional distribution $\pi(h|\cdot)$ in the M-step. Our simulations show that the proposed algorithm works very well in terms of MSE and coverage intervals. The steps of the algorithm are described below:

1. Get a pilot estimate of f by setting h to a sufficiently *large* value such that no rounding spikes occur (e.g. $h = 2\ max(\boldsymbol{r})$). Set starting values for $\boldsymbol{\tau}$ to $\Phi^{-1}(0, 1/m, 2/m, \ldots, (m-1)/m, 1)$ and for a^*, β to 0.

2. Evaluate and save density estimate $\hat{f}_X$ on an equally-spaced fine grid G with gridwidth $\delta_G = \frac{min(\boldsymbol{r})}{k}$, whereby $1 < k \in \mathbb{N}$. In particular,
$G = \left\{\min(W_i) - \frac{1}{2}r_m, \min(W_i) - \frac{1}{2}r_m + \delta_G, \min(W_i) - \frac{1}{2}r_m + 2\delta_G, \ldots, \max(W_i) + \frac{1}{2}r_m\right\}$;
$i = 1, \ldots, n$.

3. Sample from $\pi(X_i, R_i|\cdot)$ by computing it for every combination of R_i and values $X_i \in G$; $i = 1, 2, \ldots, n$.

4. Sample from $\pi(\boldsymbol{p}|\boldsymbol{R})$ in case of the model without extensions or the joint full conditional $\pi(\boldsymbol{\tau}, a^*, \beta|\boldsymbol{X}, \boldsymbol{R})$ using a Laplace normal approximation (model with extensions).

5. Estimate the bandwidth h by Silverman's rule of thumb (or another bandwidth selection method) and recompute $\hat{f}_h$.

6. Repeat steps 2-5 B (burn-in iterations) + N (additional iterations) times.

7. Discard the burn-in samples and get final estimate of f by averaging over the remaining samples. The samples of the measurement error parameters $\boldsymbol{p}$ or $\boldsymbol{\tau}$, a^* and β can be used to compute a point estimate by averaging as well as uncertainty intervals.

3.4 Computational Implementation in R

All computations were performed with *R* version 3.1.2 (R Core Team 2014). A package called *Kernelheaping* (Gross 2015) was made available on *CRAN* by the authors. It includes the full functionality as presented in this article and an additional example dataset concerning the hours per week of learning reported by students (taken from Utts and Heckard 2014). Kernel density and bandwidth estimation is done via the *density* function coming with the default installation of *R*. For non-negative data the boundary correction method introduced in Jones (1993) is utilized which is implemented in the *evmix* package (Scarrott and Hu 2014). For a sample size of $n = 5000$, 1000 iterations take about half an hour on a modern computer. The package also provides functions to perform convergence diagnostics and other convenience functions as well as functions to perform Monte-Carlo simulation studys.

4. SIMULATION STUDY

In this section we present results from a Monte-Carlo simulation study which we performed to evaluate the performance of the proposed kernel density estimator for heaped data in the previous section. The properties of the estimator are investigated and it's performance is compared to a simple $Naive$ kernel density estimator, which ignores the heaping process. The data is generated under different univariate distributions. Four scenarios, denoted by A-D, are considered. The sample size is always $n = 1000$. Under Scenario A we consider the heaping model without extensions. The data are generated by using a normal distrution,

$$X_A \sim N(0, 100),$$

with rounding values $\boldsymbol{r} = (1, 10, 100)$ and rounding probabilities $\boldsymbol{p} = (0.3, 0.4, 0.3)$.

In Scenario B we introduce a rounding bias with $a = 0.8$. Following the inspiring example of a weight distribution, the data are generated by a gamma distribution with shape α and scale θ with offset:

$$X_B \sim Ga(\alpha = 4, \theta = 8) + 45$$

The rounding values are $\boldsymbol{r} = (1, 2, 5, 10)$ with corresponding probabilities (which were arbitrary chosen) $\boldsymbol{p} = (0.1, 0.15, 0.4, 0.35)$.

In the third scenario the data follow a log-normal distribution with unequal rounding probabilities ($\beta = -1$) to model an income-like distribution,

$$X_C \sim logN(7, 0.6),$$

with rounding values $\boldsymbol{r} = (10, 20, 50, 100, 200, 500, 1000)$ and threshold values $\boldsymbol{\tau} = (-\infty, 6.33, 6.66, 7, 7.33, 7.66, 8, \infty)$. These threshold values coincide for rounding probabilities of $\boldsymbol{p} = (0.28, 0.12, 0.13, 0.13, 0.11, 0.09, 0.14)$ or $\boldsymbol{p} = (0.01, 0.02, 0.03, 0.05, 0.08, 0.11, 0.70)$ for $x = 1000$ or $x = 5000$.

A bimodal mixture of two normal distributions is considered in scenario D. With

$$X_{D_1} \sim N(40, 4) \text{ and } X_{D_2} \sim N(55, 6),$$

and mixture probabilities 0.4 (X_{D_1}) and 0.6 (X_{D_2}), an underlying heaping model with rounding bias $a = 0.2$ and unequal rounding probabilites ($\beta = -0.5$) with threshold values $\boldsymbol{\tau} = (-\infty, 1.84, 2.64, 3.05, \infty)$ is utilized in this case.

For each scenario we performed $n_{Sim} = 500$ simulation runs with B=100 burn-in iterations and N=500 additional iterations. We compare the following three estimators:

a) The *Naive* estimator, which naively applies the kernel density estimator to the heaped data

b) The *Corrected* estimator, that uses the algorithm presented in section 3.4 for kernel density estimation for heaped data

c) The *Oracle* estimator, that uses the original data (which are only available in simulations) for density estimation.

The Sheather-Jones estimator was used to for bandwidth estimation in each case. Figure 4 shows these three kernel density estimators as well as the true density from which the data is generated for a single simulation run of each scenario.

While the *Naive* estimator is very spiky and shows large deviations from the true density at the heaping points, the proposed *Corrected* density estimator is very close to the oracle estimator and represents the true density pretty well. In scenario D, we are able to recover the bimodal structure of the distribution, whereas with the *Naive* estimator this feature of the data gets lost.

Tables 1 shows the RMISE of of the three estimators for each scenario. While the *Naive* estimator exhibits a rather poor performance with a RMISE up to more than 10 times as

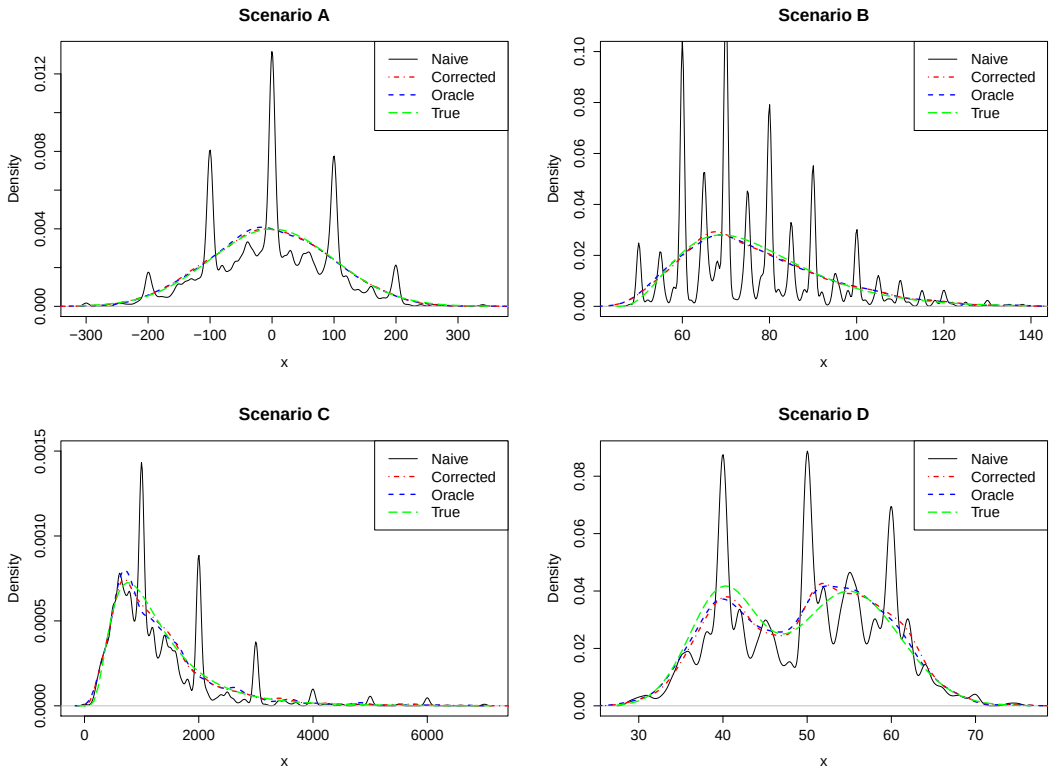

Figure 2: Graphical presentation of single simulation runs of scenarios A-D. The plots show kernel density estimators applied to heaped data (*Naive*, black solid line), applied to rounded data with correction algorithm (*Corrected*, red point-dotted line), applied to original data (*Oracle*, blue short-dashed line) and the true density function (*True*, green long-dotted line)

high as with the non-feasible *Oracle* estimator, the *Corrected* estimator leads to a negligable loss of some percent in RMISE. This slightly worse performance of the proposed estimator can be most likely assigned to the information loss induced by rounding.

Scenario	RMISE		
	Naive	*Corrected*	*Oracle*
A	0.0330 (0.0041)	0.0032 (0.0010)	0.0031 (0.0010)
B	0.2046 (0.0076)	0.1271 (0.0024)	0.1269 (0.0023)
C	0.0133 (0.0022)	0.0018 (0.0006)	0.0017 (0.0006)
D	0.2196 (0.0085)	0.0159 (0.0029)	0.0145 (0.0023)

Table 2: Root Mean Integrated Square Error (RMISE) for scenarios A-D for each estimator. Standard errors are given in parenthesis.

Besides trying to recover the true distribution one might be also interested in estimating the rounding parameters. We investigate some (frequentist) properties, namely the bias, standard deviation, Root Mean Square Error (RMSE) and the coverage rate of the 90% uncertainty intervals, of the estimates computed by the introduced algorithm. The results are shown in Tables 2-5.

	Parameter		
	p_1	p_2	p_3
True value	0.3	0.4	0.3
Bias	-0.0030	0.0018	0.0013
SD	0.0143	0.0168	0.0144
RMSE	0.0147	0.0169	0.0144
Coverage in %	88.6	87.0	93.8

Table 3: Scenario A: Bias, standard deviation, Root Mean Square Error and coverage rate of 90% uncertainty intervals for rounding parameters

Apparently, the algorithm is able to identify the rounding parameters very well. The coverage rates of the 90% uncertainty intervals are near to the nominal value as well. One may note that the threshold values have a rather large standard deviation, but this is due to the high correlation with β. The resulting rounding probabilities are pretty stable, though.

In general, the algorithm was very stable for the proposed starting values and showed very good and fast convergence. Depending on the application and heaping model only $B = 5$ to $B = 50$ burnin iterations were sufficient, but one should always consider trace plots of the MCMC-chains to ensure convergence. Trace plots for an application example can be found in the next section.

	Parameter				
	p_1	p_2	p_3	p_4	a
True value	0.1	0.15	0.4	0.35	0.8
Bias	0.0012	0.0029	-0.0046	-0.0005	-0.0079
SD	0.0164	0.0188	0.0344	0.0339	0.0501
RMSE	0.0165	0.0191	0.0348	0.0339	0.0507
Coverage in %	89.0	86.2	93.4	91.2	89.2

Table 4: Scenario B: Bias, standard deviation, Root Mean Square Error and coverage rate of 90% uncertainty intervals for rounding parameters

	Parameter						
	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	β
True value	6.33	6.66	7	7.33	7.66	8	-1
Bias	-0.0493	-0.0327	0.0165	0.0415	0.0410	-0.0307	-0.0353
SD	0.7564	0.8167	0.9243	0.8342	0.5915	0.6281	0.2412
RMSE	0.7580	0.8174	0.9244	0.8352	0.5929	0.6288	0.2431
Coverage in %	87.8	85.6	92.2	94.6	90.4	87.2	88.6

Table 5: Scenario C: Bias, standard deviation, Root Mean Square Error and coverage rate of 90% uncertainty intervals for rounding parameters

	Parameter				
	τ_1	τ_2	τ_3	a	β
True value	1.84	2.64	3.05	0.2	-0.5
Bias	-0.0545	-0.0561	0.0562	-0.0059	-0.0093
SD	1.1705	1.1532	1.1601	0.0524	0.1323
RMSE	1.1717	1.1545	1.1615	0.0527	0.1326
Coverage in %	89.0	92.8	87.0	91.4	90.8

Table 6: Scenario D: Bias, standard deviation, Root Mean Square Error and coverage rate of 90% uncertainty intervals for rounding parameters

5. APPLICATION

Now we examine the two self-reported data examples of the SOEP 2012 already presented in the introduction. In our first example we have body weight data of $n = 12168$ German women. The sample mean is 67.23 kg and the standard deviation amounts to 12.66 kg. We expect different probabilities for the rounding values depending on the actual weight. In particular, when looking at the data 51.7% of the respondents with reported weight above 90 kg report an end-digit of 0 or 5 while this is only the case for 40.3% of the group with reported weight lower than 90 kg. Additionally, we like to investigate a possible rounding bias. Therefore, the heaping model with both extensions is utilized. For bandwidth estimation we used the Sheater-Jones estimate as well as Silverman's rule of thumb. The algorithm was executed with $B = 500$ burn-in samples and $N = 2000$ additional samples and with rounding values $\boldsymbol{r} = (1, 2, 5, 10)$. The resulting densities of both the *Corrected* and the *Naive* estimator are shown in Figure 3. Though the algorithm produces a considerably smoother density estimate as the *Naive* method, but it is still very wiggly for the Sheather-Jones bandwidth selector. The authors attribute this to the fact that the imposed heaping model does not capture the actual heaping process completely. The rule of thumb bandwidth generates much smoother density estimates. However, the *Naive* estimator exhibits small bumps at the multiples of the rounding values compared to the *Corrected* estimator.

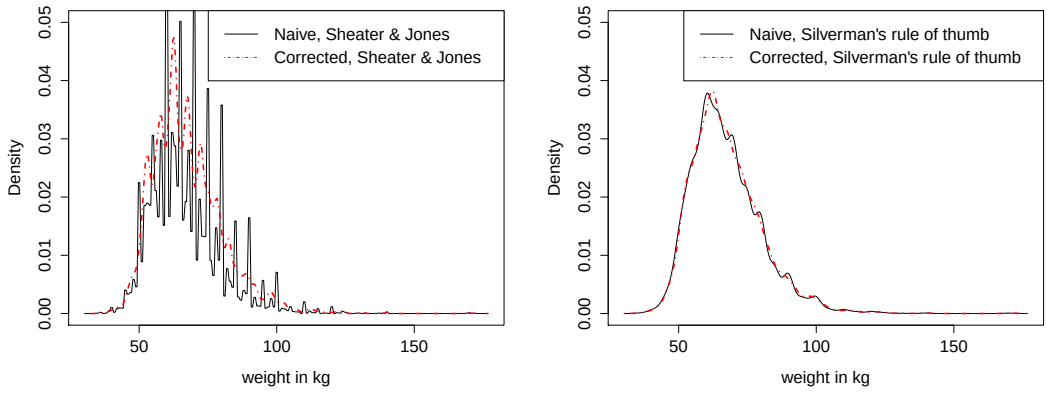

Figure 3: Kernel density estimation of self-reported female body weight for *Naive* and *Corrected* method for different bandwidth choices

Table 7 states the rounding parameter estimates. The threshold values $\boldsymbol{\tau}$ and the slope parameter of the ordered probit β suggest rounding probabilities of $\boldsymbol{p} = (0.653, 0.099, 0.222, 0.026)$ for the rounding values $\boldsymbol{r} = (1, 2, 5, 10)$ at the sample mean. The point estimate of the rounding bias a is 0.76 which means –as one could expect– that the

survey respondents are much more likely to round off than to round up. As a consequence, the mean of the imputed weights X_i is more than 200 g higher now (67.45 kg). The lower border of the 95% uncertainty interval for a is considerably above 0.5. However, to further approve this result we ran the algorithm on a different survey data sample on weight, namely the German General Social Survey 2008 ('ALLBUS', Wasmer et al. 2007). In this survey $n = 1451$ women reported their body weight and the rounding bias was estimated to $a = 0.694$ (with 95% uncertainty interval [0.503,0.853]), which is very similar to the estimate on the SOEP data. Men, as a remark, were less prone to biased rounding with point estimation values of ($a = 0.596$ for SOEP and $a = 0.569$ for ALLBUS).

Parameter	Mean	SD	95% uncertainty interval
τ_1	8.050	0.544	[6.996 9.140]
τ_2	8.337	0.541	[7.299, 9.419]
τ_3	9.604	0.551	[8.543 10.667]
a	0.760	0.027	[0.706, 0.805]
β	-1.826	0.126	[-2.076, -1.583]

Table 7: SOEP female body weight: Mean, standard deviation and 95% coverage intervals for rounding parameters

In the second example, households were asked to state their monthly food and drink expenditure outside home. The $n = 6096$ respondents stated a mean expenditure of 92.42€ with a standard deviation of 78.07€ . The algorithm was applied with rounding values $\boldsymbol{r} = (1, 2, 5, 10, 20, 50, 100)$. The heaping model with the ordinal probit model extension for non-constant rounding probabilities was utilized here, as the data suggest strong dependence of rounding behaviour on the magnitude of the expenditures. All reported values above 180€ are divisible by 10, while at least 6.7% of the reported values below 100€ are not. Figure 4 displays the resulting density estimates for different bandwidth choices. Again, for the Sheather-Jones bandwidth selector, the algorithm produces a markedly improved density estimate which is still quite rough nevertheless. For Silverman's rule of thumb, the estimate is conveniently smooth but shows a bimodal structure that may not be genuine to the underlying true expenditures. To produce a sufficiently smooth estimate, the authors suggest to manually tune the bandwidth. A bandwidth of 1.5 times the rule of thumb generates a smooth unimodal density estimate, while the *Naive* approach is still very spiky (to obtain a comparable smooth estimate, a bandwidth of 4 times the rule of thumb was necessary leading to a flatter density estimate).

The summary statistics for the rounding parameters $\boldsymbol{\tau}$ and β can be found in Table 8. The negative value of β indicates that higher rounding values ($\boldsymbol{r} = (1, 2, 5, 10, 20, 50, 100)$)

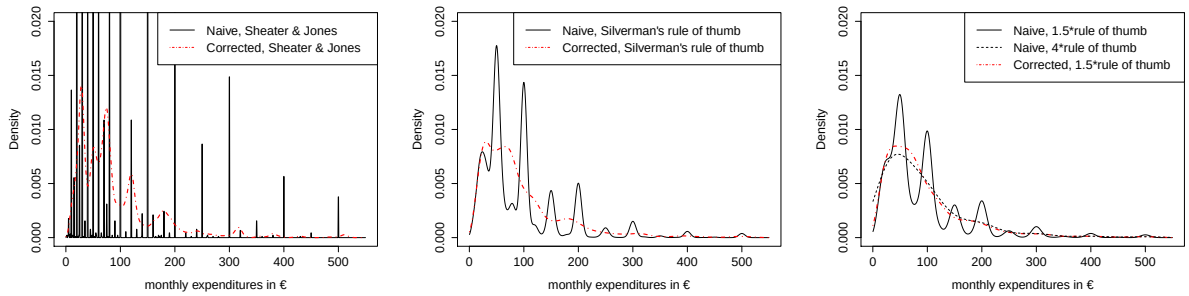

Figure 4: Kernel density estimation of food and drink expenditures outside home in € for *Naive* and *Corrected* method for different bandwidth choices

are utilized for higher monthly expenditures. Specifically, for expenditures of 25€ the model suggests rounding probabilities $\boldsymbol{p} = (0.8\%, 0.4\%, 19.4\%, 48.4\%, 11.9\%, 18.9\%, 0.1\%)$, while $\boldsymbol{p}$ equals $(0.0\%, 0.0\%, 0.2\%, 5.6\%, 5.9\%, 73.8\%, 14.5\%)$ for monthly expenditures of 150€.

Parameter	Mean	SD	95% uncertainty interval
τ_1	1.322	0.183	[0.951, 1.662]
τ_2	1.479	0.129	[1.223, 1.738]
τ_3	2.896	0.124	[2.656, 3.138]
τ_4	4.213	0.137	[3.952, 4.484]
τ_5	4.592	0.135	[4.333, 4.854]
τ_6	6.840	0.172	[6.508, 7.189]
β	-1.154	0.0316	[-1.215, -1.093]

Table 8: Food and drink expenditures outside home: Mean, standard deviation and 95% coverage intervals for rounding parameters

The algorithm converged to the same parameter values under multiple runs and different starting values for both examples (and the simulation scenarios). Trace plots for rounding parameters of the SOEP data are shown in Figures 5 and 6. Convergence is achieved after a burn-in period of about 50 iterations. The density estimates and the rounding parameters a and β were relatively robust to different choices of rounding values (for example $\boldsymbol{r} = (1, 5, 10)$ or $\boldsymbol{r} = (1, 2, 5, 10, 20)$ in the body weight example). However, in general, for rounding values which are not or very weakly supported by the data, the estimates (especially the threshold values as well as β) can get pretty unstable. The user should always consult the trace plots and eliminate the concerned rounding values if necessary.

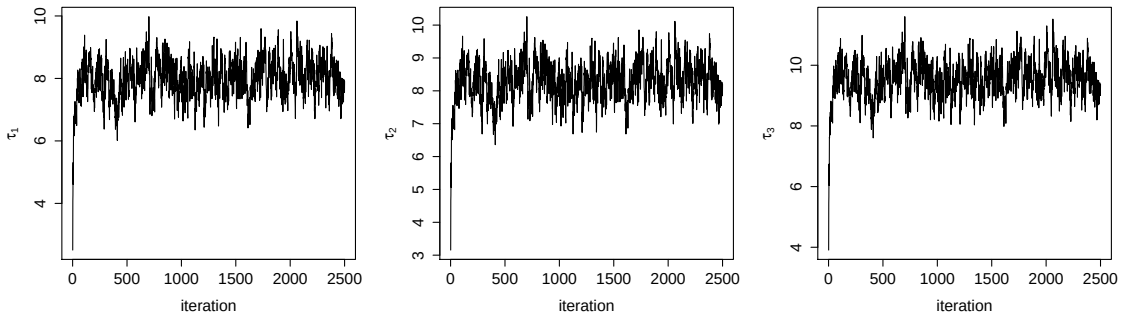

Figure 5: Trace plots for $\boldsymbol{\tau}$ (SOEP female body weight)

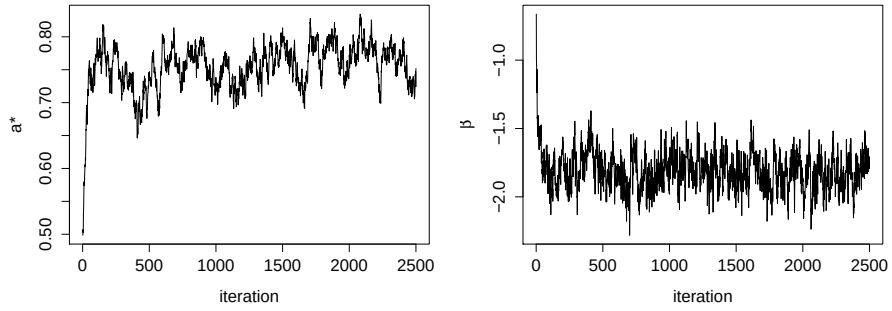

Figure 6: Trace plots for a and β (SOEP female body weight)

6. DISCUSSION

In this paper, a novel approach for kernel density estimation for heaped data was introduced. A Stochastic Expectation Maximization algorithm was presented, that generates smoother and more realistic non-parametric density estimates and gives additional insights into the rounding process. More specifically, the rounding probabilities as well as a rounding bias is estimated within the proposed algorithm. This can be very helpful for researches in assessing and validating self-reported data. In the presented example of self-reported body weight the approach was able to discover a biased response behaviour without validation data solely on the basis of reported values. The algorithm is easy to implement and is provided by the authors in a R-package. The algorithm exhibited very good statistical properties in the simulations.

However, it was necessary to make some restrictive assumptions on the heaping process. Both applications indicated that these assumptions are not completely fullfilled in real-world data. As Crawford et al. (2014) remarks, the assumption that, for example, a reported value of $W_i = 100$ with rounding value $R_i = 10$ means that the true unobserved value X_i lies inside the interval (95,105) is rather strong. A possible solution would be to decompose the reporting process into an recall error (i.e. the person does not know its body weight exactly) and a rounding error. This could be modeled by a measurement error model of classical error mixed with rounding but it is not clear how to estimate the recall error without validation data (one could set the recall error equal to the rounding error, but that would impose another assumption). Concerning the improved but still spiky density estimates under the Sheather-Jones bandwidth selector, the authors recommend to use the Silverman's rule of thumb instead and tune the bandwidth manually if necessary. However, a possible solution would be to introduce a random effect into the ordered probit model for the rounding probabilities. As the preference for some heaped values may not be captured by the model a grouping structure which assigns every X_i to the nearest possible rounded value is introduced (represented by design matrix $\boldsymbol{U}$ with rows u_i):

$$g_i = log(X_i)\beta + u_i'\boldsymbol{\gamma} + \epsilon_i \ , \ \epsilon_i \sim N(0,1), \boldsymbol{\gamma} \sim N(0, \boldsymbol{\tau})$$

The implementation is straightforward (a Metropolis-Hastings step is necessary) and first tests show very promising results, i.e. the estimated density is sufficiently smooth regardless of the bandwidth choice method. However, the authors are currently faced with stability and computing speed issues, but he is optimistic to solve these problems in the near future. Afterwards, this extension will be implemented into the R-package. A further extension could introduce a non-constant rounding bias as well. Respondents with owerweight, for

example, are possibly more inclined to round off than normal or underweight surveyed persons. Additionally, the estimation of parametric distributions is straightforward to integrate into this approach and with some minor modifications of the algorithm density estimation for classified data should be possible as well. In sum, the algorithm presented in this paper delivers a powerful and easy to use tool for users concerned with heaped data.

References

Bar, H. Y., and Lillard, D. R. (2012), "Accounting for Heaping in Retrospectively Reported Event Data–a Mixture-Model Approach," *Statistics in medicine*, 31(27), 3347–3365.

Battistin, E., Miniaci, R., and Weber, G. (2003), "What do we Learn from Recall Consumption Data?," *Journal of Human Resources*, 38(2), 354–385.

Camarda, C. G., Eilers, P. H., and Gampe, J. (2008), "Modelling general patterns of digit preference," *Statistical Modelling*, 8(4), 385–401.

Carroll, R., Ruppert, D., Stefanski, L., and Crainiceanu, C. (2010), *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability Taylor & Francis.

Celeux, G., Chauveau, D., and Diebolt, J. (1996), "Stochastic Versions of the EM algorithm: an Experimental Study in the Mixture Case," *Journal of Statistical Computation and Simulation*, 55(4), 287–314.

Crawford, F. W., Weiss, R. E., and Suchard, M. A. (2014), "Sex, Lies, and Self-Reported Counts: Bayesian Mixture Models for Longitudinal Heaped Count Data via Birth-Death Processes," *arXiv preprint arXiv:1405.4265*, .

Czajka, J. L., and Denmead, G. (2008), "Income Data for Policy Analysis: A Comparative Assessment of Eight Surveys," *Final report. Washington, DC: Mathematica Policy Research*, .

De Lusignan, S. d., Belsey, J., Hague, N., and Dzregah, B. (2004), "End-Digit Preference in Blood Pressure Recordings of Patients with Ischaemic Heart Disease in Primary Care," *Journal of Human Hypertension*, 18(4), 261–265.

Diebolt, J., Ip, E., and Olkin, I. (1994), A Stochastic EM Algorithm for Approximating the Maximum Likelihood Estimate,, Technical report, Technical Report 301, Department of Statistics, Stanford University, California.

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90(430), 577–588.

Gross, M. (2015), *Kernelheaping: Kernel Density Estimation for Heaped Data*. R package version 0.9.

Groß, M., Rendtel, U., Schmid, T., Schmon, S., and Tzavidis, N. (2015), Estimating the Density of Ethnic Minorities and Aged People in Berlin: Multivariate Kernel Density Estimation Applied to Sensitive Geo-Referenced Administrative Data Protected via Measurement Error,, Technical report, Discussion Paper, School of Business & Economics: Economics.

Hanisch, J. U. (2007), *Rounding of Income Data : An Empirical Analysis of the Quality of Income Data with Respect to Rounded Values and Income Brackets with Data from the European Community Household Panel* Peter Lang, Frankfurt.

Härdle, W., and Scott, D. W. (1992), "Smoothing by Weighted Averaging of Rounded Points," *Computational Statistics*, 7, 97–128.

Heitjan, D. F., and Rubin, D. B. (1990), "Inference from Coarse Data via Multiple Imputation with Application to Age Heaping," *Journal of the American Statistical Association*, 85(410), 304–314.

Heitjan, D. F., and Rubin, D. B. (1991), "Ignorability and Coarse Data," *Annals of Statistics*, 19(4), 2244–2253.

Izenman, A. J. (1991), "Recent Developments in Nonparametric Density Estimation," *Journal of the American Statistical Association*, 86(413), pp. 205–224.

Jones, M. C. (1993), "Simple Boundary Correction for Kernel Density Estimation," *Statistics and Computing*, 3(3), 135–146.

Marcus, J., Siegers, R., and Grabka, M. M. (2013), Preparation of Data from the New SOEP Consumption Module: Editing, Imputation, and Smoothing,, Technical Report 70, Data Documentation, DIW.

Merrill, R. M., and Richardson, J. S. (2009), "Peer Reviewed: Validity of Self-Reported Height, Weight, and Body Mass Index: Findings from the National Health and Nutrition Examination Survey, 2001-2006," *Preventing Chronic Disease*, 6(4).

Otterbach, S., and Sousa-Poza, A. (2010), "How Accurate are German Work-Time Data? A Comparison of Time-Diary Reports and Stylized Estimates," *Social Indicators Research*, 97(3), 325–339.

Pudney, S. (2008), Heaping and leaping: Survey Response Behaviour and the Dynamics of Self-Reported Consumption Expenditure,, Technical report, ISER Working Paper Series.

R Core Team (2014), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rowland, M. L. (1990), "Self-Reported Weight and Height.," *The American Journal of Clinical Nutrition*, 52(6), 1125–1133.

Scarrott, C. J., and Hu, Y. (2014), "evmix: Extreme Value Mixture Modelling, Threshold Estimation and Boundary Corrected Kernel Density Estimation,". Available on CRAN.

Schneeweiß, H., and Komlos, J. (2009), "Probabilistic Rounding and Sheppard's Correction," *Statistical Methodology*, 6(6), 577–593.

Scott, D. W. (2009), *Multivariate Density Estimation: Theory, Practice, and Visualization*, Vol. 383 John Wiley & Sons.

Sheather, S. J. (2004), "Density Estimation," *Statistical Science*, 19(4), 588–597.

Sheather, S., and Jones, C. (1991), "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 53(3), 683–690.

Shields, M., Gorber, S. C., and Tremblay, M. S. (2008), "Estimates of Obesity Based on Self-Report Versus Direct Measures," *Health Rep*, 19(2), 61–76.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability Taylor & Francis.

Taylor, A. W., Grande, E. D., Gill, T. K., Chittleborough, C. R., Wilson, D. H., Adams, R. J., Grant, J. F., Phillips, P., Appleton, S., and Ruffin, R. E. (2006), "How Valid are Self-Reported Height and Weight? A Comparison Between CATI Self-Report and Clinic Measurements Using a Large Cohort Study," *Australian and New Zealand Journal of Public Health*, 30(3), 238–246.

Utts, J., and Heckard, R. (2014), *Mind on Statistics*, Boston, United States: Cengage Learning.

Wagner, G. G., Frick, J. R., and Schupp, J. (2007), *The German Socio-Economic Panel Study (SOEP)-Evolution, Scope and Enhancements*, Vol. 127 Schmollers Jahrbuch.

Wang, H., and Heitjan, D. F. (2008), "Modeling Heaping in Self-Reported Cigarette Counts," *Statistics in Medicine*, 27(19), 3789–3804.

Wasmer, M., Scholz, E., Blohm, M. et al. (2007), *Konzeption und Durchführung der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften"(ALLBUS) 2008* GESIS-ZUMA.

Zhang, X., King, M. L., and Hyndman, R. J. (2006), "A Bayesian Approach to Bandwidth Selection for Multivariate Kernel Density Estimation," *Computational Statistics & Data Analysis*, 50(11), 3009–3031.

Diskussionsbeiträge - Fachbereich Wirtschaftswissenschaft - Freie Universität Berlin
Discussion Paper - School of Business and Economics - Freie Universität Berlin

2015 erschienen:

2015/1 GÖRLITZ, Katja und Christina GRAVERT
The effects of increasing the standards of the high school curriculum on school dropout
Economics

2015/2 BÖNKE, Timm und Clive WERDT
Charitable giving and its persistent and transitory reactions to changes in tax incentives: evidence from the German Taxpayer Panel
Economics

2015/3 WERDT, Clive
What drives tax refund maximization from inter-temporal loss usage? Evidence from the German Taxpayer Panel
Economics

2015/4 FOSSEN, Frank M. und Johannes KÖNIG
Public health insurance and entry into self-employment
Economics

2015/5 WERDT, Clive
The elasticity of taxable income for Germany and its sensitivity to the appropriate model
Economics

2015/6 NIKODINOSKA, Dragana und Carsten SCHRÖDER
On the Emissions-Inequality Trade-off in Energy Taxation: Evidence on the German Car Fuel Tax
Economics

2015/7 GROß, Marcus; Ulrich RENDTEL; Timo SCHMID; Sebastian SCHMON und Nikos TZAVIDIS
Estimating the density of ethnic minorities and aged people in Berlin: Multivariate kernel density estimation applied to sensitive geo-referenced administrative data protected via measurement error
Economics

2015/8 SCHMID, Timo; Nikos TZAVIDIS; Ralf MÜNNICH und Ray CHAMBERS
Outlier robust small area estimation under spatial correlation
Economics

2015/9 GÖRLITZ, Katja und Marcus TAMM
Parenthood and risk preferences
Economics

2015/10 BÖNKE, Timm; Giacomo CORNEO und Christian WESTERMEIER
Erbschaft und Eigenleistung im Vermögen der Deutschen: eine Verteilungsanalyse
Economics

2015/11 GÖRLITZ, Katja und Marcus TAMM
The pecuniary and non-pecuinary returns to voucher-financed training
Economics

2015/12 CORNEO, Giacomo
Volkswirtschaftliche Bewertung öffentlicher Investitionen
Economics

2015/13 GÖRLITZ, Katja und Christina Gravert
The effects of a high school curriculum reform on university enrollment and the choice of college major
Economics

2015/14 BÖNKE, Timm und Carsten SCHRÖDER
European-wide inequality in times of the financial crisis
Economics

2015/15 BÖNKE, Timm; Beate JOACHIMSEN und Carsten SCHRÖDER
Fiscal federalism and tax enforcement
Economics

2015/16 DEMMER, Matthias
Improving Profitability Forecasts with Information on Earnings Quality
FACTS

2015/17 HAAN, Peter und Victoria PROWSE
Optimal Social Assistance and Unemployment Insurance in a Life-cycle Model of Family Labor Supply and Savings
Economics

2015/18 CORNEO, Giacomo, Carsten SCHRÖDER und Johannes KÖNIG
Distributional Effects of Subsidizing Retirement Savings Accounts: Evidence from Germany
Economics

2015/19 BORGONI, Riccardo; Paola DEL BIANCO; Nicola SALVATI; Timo SCHMID und Nikos TZAVIDIS
Modelling the distribution of health related quality of life of advanced melanoma patients in a longitudinal multi-centre clinical trial using M-quantile random effects regression
Economics

2015/20 HELLER, C.-Philipp; Johannes JOHNEN und Sebastian SCHMITZ
Congestion Pricing: A Mechanism Design Approach
Economics

2015/21 BARTELS, Charlotte und Nico PESTEL
The Impact of Short- and Long-term Participation Tax Rates on Labor Supply
Economics

2015/22 JESSEN, Robin; Davud ROSTAM-AFSCHAR und Viktor STEINER
Getting the Poor to Work: Three Welfare Increasing Reforms for a Busy Germany
Economics

2015/23 BLAUFUS, Kay; Matthias BRAUNE; Jochen HUNDSDOERFER und Martin JACOB
Does Legality Matter? : The Case of Tax Avoidance and Evasion
FACTS

2015/24 RENDTEL, Ulrich
Warum im Zensus die Ergebnisse der Stichprobenmethode keine Benachteiligung der großen Gemeinden darstellen: eine Detektivarbeit
Economics

2015/25 RENDTEL, Ulrich
Is there a fade-away effect of initial nonresponse bias in EU-SILC?
Economics

2015/26 BÖNKE, Timm; Matthias GIESECKE und Holger LÜTHEN
The Dynamics of Earnings in Germany: Evidence from Social Security Records
Economics