The effect of microaggregation by individual ranking on the estimation of moments

Schmid, Matthias; Schneeweiss, Hans

Accepted Manuscript

The effect of microaggregation by individual ranking on the estimation of moments

Matthias Schmid, Hans Schneeweiss

PII: S0304-4076(09)00142-0
DOI: 10.1016/j.jeconom.2009.06.001
Reference: ECONOM 3208

To appear in: *Journal of Econometrics*

Received date: 6 May 2008
Revised date: 5 May 2009
Accepted date: 6 June 2009

Please cite this article as: Schmid, M., Schneeweiss, H., The effect of microaggregation by individual ranking on the estimation of moments. *Journal of Econometrics* (2009), doi:10.1016/j.jeconom.2009.06.001

The Effect of Microaggregation by Individual Ranking on the Estimation of Moments

Matthias Schmid [a,1] and Hans Schneeweiss [b]

[a] *Department of Medical Informatics, Biometry and Epidemiology, Friedrich-Alexander-University Erlangen-Nuremberg Waldstraße 6, D–91054 Erlangen, Germany*

[b] *Department of Statistics, University of Munich Ludwigstraße 33, D–80539 Munich, Germany*

Abstract

Microaggregation by individual ranking (IR) is an important technique for masking confidential econometric data. While being a successful method for controlling the disclosure risk of observations, IR also affects the results of statistical analyses. We conduct a theoretical analysis on the estimation of arbitrary moments from a data set that has been anonymized by means of the IR method. We show that classical moment estimators remain both consistent and asymptotically normal under weak assumptions. This theory provides the justification for applying standard statistical estimation techniques to the anonymized data without having to correct for a possible bias caused by anonymization.

Key words: consistent estimation; disclosure control; individual ranking; microaggregation; general moments.

JEL classification: C10; C12; C13.

1 Introduction

Confidential econometric data that have been collected by a statistical office are usually anonymized before publication. Anonymization is accomplished by making use of statistical disclosure control techniques. These techniques result in a reduction of the information content of the data and thus in a low re-identification risk of the observations in the published data set. As a consequence, data users (e.g., econometricians employed by universities or research institutes) have to rely on the quality of the results obtained from the masked data. A drawback of disclosure control techniques is that the reduction of the information content often leads to an efficiency loss and/or to biased statistical analysis (Willenborg and de Waal (2001), Doyle et al. (2001), Domingo-Ferrer and Torra (2004), Ronning et al. (2005), Aggarwal and Yu (2008)). Due to confidentiality requirements, a certain amount of efficiency loss cannot be avoided. However, if the efficiency loss is not too large, econometricians will still benefit from the published data. In order to control the efficiency loss arising from the anonymization of data sets, the effect of statistical disclosure control techniques on statistical analysis has to be carefully examined.

In this paper the focus is on the effect of microaggregation by individual ranking (IR) on the estimation of general moments and, by implication, on the least squares (LS) estimation of a linear model in transformed variables. IR, which has been popularized by the Statistical Office of the European Communities (Eurostat), is an important statistical disclosure control technique

[1] Corresponding author. *Email:* matthias.schmid@imbe.imed.uni-erlangen.de, *Tel.:* +49 9131 85 22706, *Fax:* +49 9131 85 25740

for continuous microdata (see Defays and Anwar (1998)). The idea of IR is to anonymize each continuous variable in a data set one after another by forming small groups (usually of size 3 or 5) of "similar" data values and by replacing the original data values with the respective group means. It is thus hoped that the values of (potentially sensitive) outliers at the tails of a distribution are masked while the multivariate distribution of the data is approximately preserved. A drawback of IR is that the technique seems to result in a relatively high identification risk of individual observations (Domingo-Ferrer and Torra (2001), Domingo-Ferrer et al. (2002), Winkler (2002)). For this reason, IR has often been used in combination with other disclosure control techniques, especially with techniques designed for masking discrete data. In the United States, for example, IR has been applied to anonymize various versions of the Individual Income Tax Return Public Use File (Strudler et al. (1986), see also `http://www.nber.org/~taxsim/gdb/`). Similarly, according to a survey conducted by the United Nations Economic Commission for Europe (UNECE Secretariat (2001)), IR has been used by several European statistical offices to mask officially collected data before publication. Another review stating that IR is used by European statistical offices was conducted by Felsö et al. (2001). In a recent publication on disclosure control, Eurostat considered IR to be a "perfectly acceptable strategy if a slight modification of the data is deemed sufficient" for avoiding disclosure of respondents (Hundepool et al. (2009), p. 115). Similarly, the members of a major German research project recommended to anonymize business data by using a combination of IR and disclosure techniques for discrete variables (Ronning et al. (2005), Rosemann et al. (2006)). This strategy has been adopted by the German statistical offices (Rosemann et al. (2006)).

In two previous papers (Schmid (2006), Schmid and Schneeweiss (2008)) we have analyzed the effect of IR on the estimation of linear models. In Schmid (2006) it was shown analytically that a linear model can be consistently estimated from the microaggregated data by standard LS estimation techniques. In addition, if the continuous variables in a data set are assumed to follow a mixed normal distribution each, the efficiency loss due to IR is asymptotically zero. In Schmid and Schneeweiss (2008) we have extended this theory by considering linear models in transformed variables where nonlinear variable transformations are applied to the data *after* microaggregation. We have shown that even in this case the LS estimators of a linear model remain consistent under mild regularity assumptions.

It should be pointed out that the consistency results derived for transformed data (Schmid and Schneeweiss (2008)) do not automatically follow from the results for untransformed data (Schmid (2006)). This is because nonlinear transformations of microaggregated data introduce an additional (finite sample) bias in the LS estimators. For instance, the empirical mean of three logarithmized data values is usually different from the logarithmized mean of the three values.

The purpose of this paper is to provide a generalization of the theory presented in Schmid (2006) and Schmid and Schneeweiss (2008) to the estimation of *arbitrary* moments based on transformed and untransformed microaggregated data. The variables involved need not be continuous variables as in Schmid and Schneeweiss (2008), so the consistency proof is adapted to this more general case. In addition, arbitrary multivariate moments are considered and not only product moments as in Schmid and Schneeweiss (2008). We will not only prove the consistency of the empirical moments computed from microaggre-

gated data but will also specify conditions and regularity assumptions under which the moments are asymptotically normal. Arbitrary moments include first, second, and product moments of the transformed and untransformed data as special cases. Thus, the consistency results for linear models presented in Schmid (2006) and Schmid and Schneeweiss (2008) are confirmed. Moreover, since the consistent estimation of arbitrary moments from the microaggregated data is guaranteed, any method-of-moments estimator is in turn consistent if computed from the microaggregated data. It should be noted that these results (obtained for the IR method) are fundamentally different from previous results obtained for other microaggregation techniques, such as multivariate microaggregation with a sorting variable (Mateo-Sanz and Domingo-Ferrer (1998), Domingo-Ferrer and Torra (2001)). In the latter case, moment estimators have been shown to be asymptotically biased, see Schmid et al. (2007). Other approaches to multivariate microaggregation (Laszlo and Mukherjee (2005), Domingo-Ferrer et al. (2006), Martinez-Balleste et al. (2007), Domingo-Ferrer et al. (2008)) have not yet been analyzed analytically with respect to their impact on the analytic potential of microdata sets.

The paper is organized as follows: In Section 2 we give an example of the IR method and illustrate the problems arising from nonlinear transformations of the microaggregated data. In Section 3 the consistency of the empirical moments computed from transformed micraggregated data is shown. Section 4 deals with the asymptotic normality of these estimators. Section 5 contains a simulation study on the theoretical results derived in Sections 3 and 4, as well as the econometric analysis of an officially collected example data set. A summary of the results presented in this paper is given in Section 6. Proofs of theorems are given in the Appendix.

2 Microaggregation by individual ranking

Microaggregation by individual ranking works as follows: First, the data set is sorted by the first continuous variable, and a fixed group size K is chosen for this variable. Next, groups of K consecutive observations are formed. The values of the variable in each group are replaced by their corresponding group means, while the values of the other variables in the data set are left unchanged. Then the same procedure is repeated for the second continuous variable, and so on. If the number of observations n is not a multiple of K, it is common practice to alter the procedure such that the groups around the medians contain $K + \mathrm{mod}\,(n/K)$ adjacent data values (see Domingo-Ferrer et al. (2002)). It is generally considered necessary to form groups of at least $K = 3$ observations, as data attackers can easily identify an observation in a group of less than 3 observations if they have sufficient background knowledge on only *one* of the observations. In practice, it is common to form groups of sizes 3 or 5. Note that the group size may differ for different variables and can also be equal to 1, meaning that this variable has not been microaggregated.

If there are discrete variables in the data set, they are left unchanged during the IR procedure. We assume that all discrete variables have either been left unchanged or have been masked by so-called *non-perturbative* disclosure control techniques, so that the multivariate distribution of these variables is unbiased. Non-perturbative disclosure control techniques, which include subsampling of observations and re-coding of data values, are widely used in practice. For an overview we refer to Willenborg and de Waal (2001) and Ronning et al. (2005).

As an example of IR we consider a data set consisting of two vectors x and y,

both containing continuous data. In addition, we consider a "dummy" vector z containing the values of a discrete binary variable. Assume that the original data set is given by

x	2	4	7	0	9	5	1	8	3
y	4	2	0	9	1	5	6	11	10
z	1	0	1	0	1	1	1	1	1

.

The first step of IR results in the sorted data set

x	0	1	2	3	4	5	7	8	9
y	9	6	4	10	2	5	0	11	1
z	0	1	1	1	0	1	1	1	1

,

where the rows of the original data set have been ordered according to the values of x. In the second step of IR, with K chosen to be 3, the values of x are microaggregated:

$\tilde{x}$	1	1	1	4	4	4	8	8	8
y	9	6	4	10	2	5	0	11	1
z	0	1	1	1	0	1	1	1	1

.

The third step of IR results in the sorted data set

$\tilde{x}$	8	8	4	1	4	1	1	4	8
y	0	1	2	4	5	6	9	10	11
z	1	1	0	1	1	1	0	1	1

,

where the rows have been ordered according to the values of y. Finally, in the fourth step of IR, again with K chosen to be 3, the values of y are microaggregated:

$\tilde{x}$	8	8	4	1	4	1	1	4	8
$\tilde{y}$	1	1	1	5	5	5	10	10	10
z	1	1	0	1	1	1	0	1	1

.

Now suppose that $\tilde{y}$ is additionally transformed by means of a quadratic transformation. Then

$$\tilde{y}^2 = (1, 1, 1, 25, 25, 25, 100, 100, 100) \, .$$

Obviously, taking the squares of the microaggregated values of y results in a different data set than when the squared values of y are microaggregated. In the latter case, one would have obtained

$$\widetilde{y^2} = (1.67, 1.67, 1.67, 25.67, 25.67, 25.67, 100.67, 100.67, 100.67) \, .$$

Now consider the estimation of a theoretical moment, i.e., the expectation of an arbitrary one-dimensional function of the random variables (X, Y, Z) from the microaggregated data. Since the original data have been altered by IR, the consistent estimation of the theoretical moment by its corresponding ordinary

empirical moment is not guaranteed any more. In the next sections we will address this problem.

3 Consistent estimation of moments

Let $\boldsymbol{X} = (X^{(1)}, \ldots, X^{(p)})$ be a p-dimensional random vector and let $\boldsymbol{x}_i = (x_i^{(1)}, \ldots, x_i^{(p)})$, $i = 1, \ldots, n$, be an i.i.d. sample taken from the distribution of $\boldsymbol{X}$. The corresponding individually microaggregated data are denoted by $\tilde{\boldsymbol{x}}_i = (\tilde{x}_i^{(1)}, \ldots, \tilde{x}_i^{(p)})$, $i = 1, \ldots, n$. The group size for the aggregation of $X^{(k)}$ is denoted by K_k. For simplicity we always assume n to be a multiple of K_k (this assumption does not affect the asymptotic results derived in the following). We want to prove that the usual consistent estimator of the moments of the distribution of $\boldsymbol{X}$ remains consistent if we replace the original data by their microaggregated data values.

Let us consider general moments: Suppose that the expectation $\mathbb{E}[h(\boldsymbol{X})]$ for some measurable real-valued function h exists. We know that, given an i.i.d. sample $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\mathbb{E}[h(\boldsymbol{X})]$ can be consistently estimated by the empirical mean $\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i)$. The following sufficient conditions guarantee that $\mathbb{E}[h(\boldsymbol{X})]$ can be consistently estimated by the empirical mean $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i)$ constructed from the individually microaggregated data $\tilde{\boldsymbol{x}}_i$,, $i = 1, \ldots, n$:

$\mathcal{H}1$: h is defined on an open p-dimensional rectangle $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_p$, where $\mathcal{D}_k$ is a finite or infinite open interval on the real line, $k = 1, \ldots, p$. The support of $\boldsymbol{X}$ is contained in $\mathcal{D}$.

$\mathcal{H}2$: h is Lipschitz continuous on every finite closed p-dimensional subrectangle $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_p \subset \mathcal{D}$, where each $\mathcal{B}_k$, $k = 1, \ldots, p$, is a finite closed interval on the real line (i.e., there exists a positive bound H' depending

only on $\mathcal{B}$ such that, for any two points $\boldsymbol{x}_1 \in \mathcal{B}$ and $\boldsymbol{x}_2 \in \mathcal{B}$, $|h(\boldsymbol{x}_1) - h(\boldsymbol{x}_2)| \leq H' d(\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $d(\boldsymbol{x}_1, \boldsymbol{x}_2)$ is the Euclidean distance between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$).

$\mathcal{H}3$: There exist non-negative real-valued functions $h_k(x^{(k)})$ defined on $\mathcal{D}_k$, $k = 1, \ldots p$, which are Lipschitz continuous on every finite closed interval $\mathcal{B}_k \subset \mathcal{D}_k$, such that $|h(\boldsymbol{x})| \leq \sum_{k=1}^{p} h_k(x^{(k)})$ for all $\boldsymbol{x} = (x^{(1)}, .., x^{(p)}) \in \mathcal{D}$.

$\mathcal{H}4$: $\mathbb{E}[h_k^2(X^{(k)})] < \infty,\ k = 1, \ldots p.$

$\mathcal{H}5$: h_k is monotone on each side of $\overline{\mathcal{C}}_k$, where $\mathcal{C}_k \subset \mathcal{D}_k$ is a fixed finite closed interval and $\overline{\mathcal{C}}_k$ its complement in $\mathcal{D}_k$.

Remark 1: The Lipschitz condition in $\mathcal{H}2$ and $\mathcal{H}3$ can be replaced with the more familiar condition that h and h_k are continuously differentiable on their domains. Of course, the Lipschitz condition is weaker and therefore to be preferred. E.g., the function $h(x) = |x|$ is Lipschitz but not differentiable in its domain.

Theorem 1 *Let $\boldsymbol{X}$ be a p-dimensional random vector and $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ an i.i.d. sample from the distribution of $\boldsymbol{X}$. Let $(\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n)$ be the corresponding microaggregated sample with fixed aggregation group sizes K_k, $k = 1, \ldots, p$, assuming (w.l.o.g.) n to be a multiple of K_k. Let $h(\boldsymbol{x})$ satisfy the conditions $\mathcal{H}1$ to $\mathcal{H}5$. Then a. s.*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i) = \mathbb{E}[h(\boldsymbol{X})] \ .$$

In the univariate case ($p = 1$), condition $\mathcal{H}4$ can be replaced with the weaker condition $\mathcal{H}4'$: $\mathbb{E}[h_k(X^{(k)})] < \infty,\ k = 1, \ldots p.$

PROOF. See Appendix.

Example 1: For the one-dimensional case ($p = 1$), Theorem 1 applies to ordi-

nary moments of X and moments of transformed variables $h(X)$, such as $|X|$, $\sin(X)$, $\log(X)$, or $X^\lambda, \lambda \in \mathbb{R}^+$, where in the last two cases $\mathcal{D} = (0, \infty)$. These moments can be consistently estimated from the microaggregated sample via the empirical moments $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i)$. An example for $p = 2$ is the product moment $\mathbb{E}[h_1(X^{(1)})h_2(X^{(2)})]$ of two transformed variables with h_k Lipschitz on each $\mathcal{B}_k$, $h_k(x^{(k)})$ monotone on each side of some $\bar{\mathcal{C}}_k$, and $\mathbb{E}[h_k^4(X^{(k)})] < \infty$, $k = 1, 2$. Indeed, as $|h_1(X^{(1)})h_2(X^{(2)})| \leq h_1^2(X^{(1)}) + h_2^2(X^{(2)})$, condition $\mathcal{H}3$ is satisfied.

Corollary 1 *For two variables $X^{(1)}$ and $X^{(2)}$ ($p = 2$), the conclusion of Theorem 1 also holds true if the sum $\sum_{k=1}^{2} h_k(x^{(k)})$ in condition $\mathcal{H}3$ is replaced with the product: $h_1(x^{(1)})h_2(x^{(2)})$ of two corresponding functions.*

PROOF. See Appendix.

Example 2: Again the product moment is an example, but now we get the previous result with the weaker condition that both $\mathbb{E}[h_k^2(X^{(k)})]$, $k = 1, 2$, exist, while $\mathbb{E}[h_k^4(X^{(k)})]$ need not exist, see also Schmid and Schneeweiss (2008). Theorem 1 and Corollary 1 imply that, under mild regularity conditions, a linear regression model in transformed variables can be consistently estimated with microaggregated data.

4 Asymptotics

Theorem 1 states that we can consistently estimate any moment $\mathbb{E}[h(\boldsymbol{X})]$ from the transformed microaggregated data in the same way as we would estimate $\mathbb{E}[h(\boldsymbol{X})]$ from the non-microaggregated data. We now give conditions under which the estimator constructed from the microaggregated data

is asymptotically as efficient as the corresponding estimator constructed from the non-aggregated data. It is possible to prove even more: The two estimators are asymptotically equivalent under certain conditions, in the sense that $\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i) - \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i))$ tends to zero in probability with $n \to \infty$. There are two sets of conditions that we need, one concerning the transformation function h, the other one concerning the distribution of $\boldsymbol{X}$. The conditions on h are stronger than the corresponding conditions for Theorem 1.

$\mathcal{H}$ (Conditions on the transformation function h):

$\mathcal{H}1^*$: Same as $\mathcal{H}1$. As to notation, let $\mathcal{D}_k = (d_{lk}, d_{uk})$, where d_{lk} and/or d_{uk} may be finite or infinite.

$\mathcal{H}2^*$: The function $h(\boldsymbol{x})$ has continuous partial derivatives on $\mathcal{D}$ denoted by

$$h'_k(\boldsymbol{x}) := \frac{\partial}{\partial x^{(k)}} h(x^{(1)}, \ldots, x^{(p)}),\ k = 1, \ldots, p.$$

$\mathcal{H}3^*$: For each $k = 1, \ldots, p$, let $\mathcal{C}_k = [c_{lk}, c_{uk}] \subset \mathcal{D}_k$ be a finite closed interval lying in $\mathcal{D}_k$, where $c_{uk} > 0$ ($c_{lk} < 0$) if $d_{uk} = \infty$ ($d_{lk} = -\infty$). Suppose there are functions $h_k(x^{(k)})$, $k = 1, \ldots, p$, of the following form:

$$h_k(x^{(k)}) = \begin{cases} h_{lk}(x^{(k)}) & \text{for} \quad d_{lk} < x^{(k)} < c_{lk} \\ C & \text{for} \quad c_{lk} \le x^{(k)} \le c_{uk} \\ h_{uk}(x^{(k)}) & \text{for} \quad c_{uk} < x^{(k)} < d_{uk} \end{cases},$$

where C is a positive constant and

$$h_{lk}(x^{(k)}) = \begin{cases} a_{lk}(-x^{(k)})^{m_{lk}} & \text{if} \quad d_{lk} = -\infty \\ a_{lk}(x^{(k)} - d_{lk})^{-m_{lk}} & \text{if} \quad d_{lk} > -\infty \ , \end{cases}$$

$$h_{uk}(x^{(k)}) = \begin{cases} a_{uk}(x^{(k)})^{m_{uk}} & \text{if} \quad d_{uk} = \infty \\ a_{uk}(d_{uk} - x^{(k)})^{-m_{uk}} & \text{if} \quad d_{uk} < \infty \end{cases}$$

with some positive constants a_{lk}, a_{uk}, m_{lk}, and m_{uk}, such that

$$|h'_r(\boldsymbol{x})| \leq \sum_{k=1}^{p} h_k(x^{(k)}), \quad r = 1, \ldots, p.$$

Remark 2: The bounding functions $h_k(x^{(k)})$ of Condition $\mathcal{H}3^*$ increase like power functions when $x^{(k)}$ approaches the boundaries of $\mathcal{D}_k$. They are monotone near these boundaries, while they are constant in the middle region $\mathcal{C}_k$.

$\mathcal{F}$ (Condition on the distribution of X):

Let F_k be the distribution function of $X^{(k)}$. Then, with m_{lk} and m_{uk} from Assumption $\mathcal{H}$,

$$\begin{aligned} &\lim_{n\to\infty} \left[1 - F_k\left(-n^{\frac{1}{4(m_{lk}+1)}}\right)\right]^n = 1 \qquad \text{if } d_{lk} = -\infty \ , \\ &\lim_{n\to\infty} \left[1 - F_k\left(d_{lk} + n^{-\frac{1}{4m_{lk}}}\right)\right]^n = 1 \quad \text{if } d_{lk} > -\infty \ , \\ &\lim_{n\to\infty} \left[F_k\left(n^{\frac{1}{4(m_{uk}+1)}}\right)\right]^n = 1 \qquad \text{if } d_{uk} = \infty \ , \\ &\lim_{n\to\infty} \left[F_k\left(d_{uk} - n^{-\frac{1}{4m_{uk}}}\right)\right]^n = 1 \qquad \text{if } d_{uk} < \infty \ . \end{aligned}$$

Remark 3: While the conditions on h_k bound the growth of $h'(\boldsymbol{x})$ when $x^{(k)}$ approaches the boundaries of $\mathcal{D}_k$, the conditions on the distribution of X describe how fast the distribution function $F_k(x^{(k)})$ tends to 0 or 1 when $x^{(k)}$ approaches the boundaries of $\mathcal{D}_k$. The stronger $h_k(x^{(k)})$ grows at the

boundaries of $\mathcal{D}_k$ (i.e., the larger the numbers m_{lk} and m_{uk} are), the faster $F_k(x^{(k)})$ has to go to its limits 0 or 1.

Theorem 2 *Suppose an i.i.d. sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of a random vector $\boldsymbol{X}$ has been microaggregated. If the transformation function h and the distribution functions of $X^{(k)}$ satisfy conditions $\mathcal{H}$ and $\mathcal{F}$, respectively, then*

$$\text{plim}_{n\to\infty} \frac{1}{\sqrt{n}} \left[\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i) - \sum_{i=1}^{n} h(\boldsymbol{x}_i) \right] = 0 \ . \tag{1}$$

PROOF. See Appendix.

Assuming that the estimator $\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i)$ of $\mathbb{E}[h(\boldsymbol{X})]$ is asymptotically normal with asymptotic variance σ_h^2/n, the asymptotic equivalence of $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i)$ and $\frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i)$ implies that $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i)$ is also asymptotically normal with the same asymptotic variance:

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i) - \mathbb{E}[h(\boldsymbol{X})] \right) \overset{n\to\infty}{\longrightarrow} \ \mathrm{N}(0, \sigma_h^2) \ . \tag{2}$$

Thus the estimator with microaggregated data is (asymptotically) just as efficient as the estimator with the original data.

Example 3: Let $p = 1$ and (suppressing the index $k = 1$) let $X \sim \mathrm{N}(\mu, \sigma^2)$ and $h(x) = x^r$, $r \in \mathbb{Z}^+$. The estimator $\frac{1}{n}\sum_{i=1}^{n} h(x_i)$ then estimates the r-th moment of X. With microaggregated data the estimator is $\frac{1}{n}\sum_{i=1}^{n} h(\tilde{x}_i)$. We show that the conditions of Theorems 1 and 2 are satisfied. First note that $\mathcal{D} = (-\infty, \infty)$. Obviously, $h(x)$ is continuously differentiable and $|h(x)|$ is monotone for $x > 0$ as well as for $x < 0$. Also $\mathbb{E}(X^r)$ exists. By Theorem 1, $\frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i^r$ is a consistent estimator of $\mathbb{E}(X^r)$.

As to the conditions of Theorem 2, first note that obviously $\mathcal{H}$ is satisfied with $m_u = m_l = r - 1$. $\mathcal{F}$ is also satisfied because (assuming w.l.o.g. $\mu = 0$ and

$\sigma^2 = 1$)

$$\lim_{n\to\infty} \left[\Phi(n^{\frac{1}{4r}})\right]^n = 1 \ ,$$

see Schmid et al. (2007). Similarly,

$$\left[1 - \Phi(-n^{\frac{1}{4r}})\right]^n = \left[\Phi(n^{\frac{1}{4r}})\right]^n \ \stackrel{n\to\infty}{\longrightarrow} \ 1 \ .$$

Thus, by Theorem 2, $\frac{1}{n}\sum_{i=1}^n \tilde{x}_i^r$ is an asymptotically normal estimator of $\mathbb{E}(X^r)$ with the same asymptotic variance as $\frac{1}{n}\sum_{i=1}^n x_i^r$.

Example 4: Let X be lognormally distributed, i.e., $\log(X) \sim \mathrm{N}(\mu, \sigma^2)$ and let $h(x) = (\log(X))^r$, $r \in \mathbb{Z}^+$. (Again $p = 1$ and the index $k = 1$ is suppressed). The estimator $\frac{1}{n}\sum_{i=1}^n h(x_i)$ then estimates the r-th moment of $\log(X)$. We show that the conditions of Theorems 1 and 2 are satisfied.

First note that the domain of $\log(X)$ is $\mathcal{D} = (0, \infty)$ and the support of X coincides with $\mathcal{D}$. Obviously, h is continuously differentiable. In addition, $|h(x)| = |\log(x)|^r$ is monotone for $0 < x < 1$ and for $x > 1$. Also, $\mathbb{E}[h(X)]$ exists. Thus Theorem 1 can be applied showing that $\frac{1}{n}\sum_{i=1}^n (\log(\tilde{x}_i))^r$ is a consistent estimator of $\mathbb{E}[(\log(X))^r]$.

To verify the conditions for Theorem 2, first note that $d_l = 0$ and $d_u = \infty$. As $|h'(x)| = r|\log(x)|^{r-1}x^{-1}$,

$$|h'(x)| \leq \max_{x\geq 1} |h'(x)| \ = \ r(r-1)^{r-1}e^{1-r} \ =: \ a \quad \text{for } x \geq 1 \tag{3}$$

and

$$|h'(x)| \ < \ rx^{-r} \quad \text{for } 0 < x < 1 \tag{4}$$

because $|\log(x)| < \frac{1}{x}$ for $0 < x < 1$. Thus, condition $\mathcal{H}$ is satisfied with $m_u = 0$ and $m_l = r$. Without loss of generality we may assume X to be

standard log-normally distributed, i.e., $\log(X) \sim \mathrm{N}(0,1)$. Then (with $m_u = 0$)

$$F\Big(n^{\frac{1}{4(m_u+1)}}\Big) = \Phi\Big(\tfrac{1}{4}\log(n)\Big) =: \Phi(b_n)\ .$$

Now, $\Phi(b) > 1 - \frac{1}{\sqrt{2\pi}}\frac{1}{b}e^{-\frac{1}{2}b^2}$ for any $b > 0$ and, in particular, for $b = b_n$, see Durrett (1991), Theorem (1.3). Let $a_n := \frac{n}{\sqrt{2\pi}b_n}e^{-\frac{1}{2}b_n^2}$. Then, since $a_n = \frac{4n}{\sqrt{2\pi\log(n)}}e^{-\frac{1}{32}\log^2(n)} \to 0$ for $n \to \infty$, we have $(1 - \frac{a_n}{n})^n \to e^0 = 1$, and hence $[\Phi(b_n)]^n \to 1$. Thus

$$\left[F\left(n^{\frac{1}{4(m_u+1)}}\right)\right]^n \to 1$$

for $n \to \infty$. Similarly (with $d_l = 0$ and $m_l = r$)

$$\begin{aligned}&\left[1 - F\left(d_l + n^{-\frac{1}{4m_l}}\right)\right]^n = \left[1 - F\left(n^{-\frac{1}{4r}}\right)\right]^n \\ =\ &\left[1 - \Phi\left(-\tfrac{1}{4r}\log(n)\right)\right]^n = \left[\Phi\left(\tfrac{1}{4r}\log(n)\right)\right]^n \to 1\ .\end{aligned}$$

This shows that condition $\mathcal{F}$ is satisfied as well. Thus Theorem 2 can be applied showing that $\frac{1}{n}\sum_{i=1}^n(\log(\tilde{x}_i))^r$ is an asymptotically normal estimator of $\mathbb{E}[(\log(X))^r]$.

5 Simulations and application example

We start with a simulation study on the quadratic regression

$$Y = 5 \cdot X^2 + \epsilon\ , \tag{5}$$

where X and ϵ are independent and standard normally distributed each. The slope parameter $\beta = 5$ can be expressed as a continuously differentiable function of the moments $\mathbb{E}[Y \cdot X^2]$ and $\mathbb{E}[(X^2)^2]$. Now suppose that Y and X have both been microaggregated with group size $K = 3$, and that the quadratic

transformation has to be applied to the data values of X after microaggregation. Then Theorem 1 guarantees the consistent estimation of $\mathbb{E}[Y \cdot X^2]$ and $\mathbb{E}[(X^2)^2]$ from the data (see Examples 1 and 2), and thus also the consistent least squares estimation of β. Moreover, due to Theorem 2, application of the delta method guarantees the asymptotic normality and efficiency of the least squares estimator of β computed from the microaggregated data. It is straightforward to extend these results to the case of a multiple polynomial regression. Table 1 shows the estimation results for $n = 300$ and 100 simulation runs. The similarities between the least squares estimator based on the non-aggregated data and the least squares estimator based on the transformed microaggregated data are obvious.

Our next example is the method-of-moments estimator of the shape and scale parameters of a Gamma distributed random variable X. Denote the shape parameter by α and the scale parameter by β. It is well known that the method-of-moments estimators computed from an i.i.d. sample $x_1, \ldots, x_n$ are $\hat{\alpha} = m_1^2/(m_2 - m_1^2)$ and $\hat{\beta} = (m_2 - m_1^2)/m_1$, where $m_1 := \sum_{i=1}^n x_i/n$ and $m_2 := \sum_{i=1}^n x_i^2/n$ are the first and second empirical moments of X. Since we have shown in Theorem 1 that the corresponding empirical moments computed from a microaggregated data set $\tilde{x}_1, \ldots, \tilde{x}_n$ converge a. s. to m_1 and m_2 as $n \to \infty$, estimation of α and β based on the microaggregated data yields asymptotically the same values as estimation based on the original data. Table 2, where the estimation results of a simulation study with 100 simulation runs are shown, confirms this result ($n = 300$, $K = 3$, $\alpha = 0.5$, $\beta = 2$).

Our third example concerns the maximum likelihood estimation of the scale

parameter c of a Levy distribution with density function

$$f(x) = \sqrt{\frac{c}{2\pi}} \frac{e^{-c/(2x)}}{x^{3/2}} \ . \tag{6}$$

The score function of a Levy distributed i.i.d. data sample $x_1, \ldots, x_n$ is given by

$$\frac{\partial l}{\partial c}(x_1, \ldots, x_n) = \frac{n}{2c} - \sum_{i=1}^{n} \frac{1}{2x_i} \ . \tag{7}$$

As the maximum likelihood estimator $\hat{c} := \left[\sum_{i=1}^{n}(1/x_i)/n\right]^{-1}$ is a consistent estimator of c, $\left[\sum_{i=1}^{n}(1/\tilde{x}_i)/n\right]^{-1}$ is also consistent (which is guaranteed by Theorem 1 and the monotonicity of $h(x) = 1/x$). Table 3, where the estimation results of a simulation study with 100 simulation runs are shown, confirms this result ($n = 300$, $K = 3$, $c = 2$).

Our final example is an analysis based on the data of the 2004 cost structure survey of enterprises of the mining and manufacturing industry in Germany (KSE). This survey is carried out regularly by the German Federal Statistical Office. As the data obtained from this survey contain comprehensive information on the German industry, they form an important basis for the national accounts of Germany. Also, they are a typical example of an officially collected data set that has to be anonymized before dissemination. The 2004 KSE data have been obtained from $n = 16\,099$ companies with 20 or more employees. Following the approach of Fritsch and Stephan (2003) and Ronning et al. (2005), we estimate a linear model of the form

$$\log(Y) = \gamma_0 + \sum_{j=1}^{5} \beta_j \log(X^{(j)}) + \epsilon \ , \tag{8}$$

where Y is an adjusted gross output of the companies and the regressors $X^{(1)}, \ldots, X^{(5)}$ are various cost factors. Model (8) corresponds to a logarithmized Cobb-Douglas production function whose production elasticities are

equal to the coefficients $\beta_1, \ldots, \beta_5$. As the least squares estimator of Model (8) from a microaggregated data set with variables $Y, X^{(1)}, \ldots, X^{(5)}$ is based on the first and second moments of $\log(Y), \log(X^{(1)}), \ldots, \log(X^{(5)})$, Theorem 1 applies (see Example 1), assuming that the regressor variables are (at least approximately) lognormally distributed.

Tables 4 and 5 show the estimation results obtained from the transformed original data and from the transformed microaggregated data (IR with group size $K = 3$). As expected, we see that IR has virtually no effect on the coefficient estimates of Model (8) and their estimated standard deviations.

6 Summary and conclusion

Microaggregation by individual ranking (IR) is a disclosure control technique which is generally considered to have a relatively small impact on the utility of an anonymized data set. In this paper we have shown analytically that IR has indeed favorable properties with respect to the estimation of statistical models: Any arbitrary moment which is defined as the expectation of a function h of a set of random variables can be consistently estimated from the microaggregated data by using the standard empirical moment estimators. Moreover, we did not assume the variables under consideration to be continuous. Thus, mixed moments between a microaggregated continuous and a non-microaggregated discrete variable can be estimated, as well as moments purely based on microaggregated continuous variables.

A further important result is the proof of asymptotic normality of the moment estimators based on the microaggregated data. This follows from the fact that the moment estimators are asymptotically equivalent to the corresponding

moment estimators computed from the non-aggregated data. Moment estimators with microaggregated and with the original data are thus equally efficient asymptotically. These results have been derived under suitable regularity conditions concerning the behavior of the transformation function h and of the distribution at the border of the domain of h. The simulation studies and data examples presented in Section 5 show that the asymptotic theory derived in this paper is already applicable when sample sizes are relatively small, i.e., when $n \geq 300$.

It should finally be pointed out that the favorable properties of the IR method go hand in hand with a relatively weak protection effect of IR, at least if no additional disclosure control procedures have been employed (there is generally a trade-off between analytic potential and protection effect of a disclosure control technique). However, as stated in Section 1, several statistical offices have used IR in combination with other disclosure control techniques to create safe public-use and scientific-use files. In particular, application of IR may be appropriate if the discrete variables in a data set (which serve as the main identifiers for attackers and data snoopers) are suitably anonymized by means of disclosure control techniques for discrete data. An overview of such methods is given in Willenborg and de Waal (2001). Also, the protection effect of IR can be enhanced if the group sizes K_k are taken sufficiently large. In this context, it is important to note that our asymptotic results do not depend on K_k. The consistency and asymptotic normality properties of the moment estimators based on the microaggregated data are not affected by the choice of the group size. Other microaggregation methods, such as distance-based microaggregation techniques (Domingo-Ferrer and Mateo-Sanz (2002), Laszlo and Mukherjee (2005), Domingo-Ferrer et al. (2006)) or microaggregation by

a sorting variable (Mateo-Sanz and Domingo-Ferrer (1998), Domingo-Ferrer and Torra (2001)), are generally considered to be more effective in protecting confidential data than the IR method. However, the analytic potential of data sets that have been masked by means of these methods seems to be limited (though partially retrievable, at least for the case of microaggregation with a sorting variable, see Schmid et al. (2007)).

While analyzing the protection effect of IR clearly is beyond the scope of this paper, we suggest that in those cases where the application of IR to an econometric data set sufficiently reduces disclosure risk, IR *should* be applied, since the method guarantees that many standard estimation techniques result in valid econometric findings.

Acknowledgements:

We thank Nadine Bartke and Mario Walter of the Bavarian State Office for Statistics and Data Processing for their support with the analysis of the 2004 KSE data. We also thank Daniel Rost and two anonymous referees for very helpful discussions and comments on the manuscript.

Appendix

Proof of Theorem 1:

Let $\mathcal{B}_k$ be finite closed intervals such that $\mathcal{C}_k \subset \mathcal{B}_k \subset \mathcal{D}_k$, $k = 1, \ldots, p$, and let $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_p$. Choose $\mathcal{B}_k$ such that

$$\mathbb{E}[h_k(X^{(k)})\mathrm{I}_{\bar{\mathcal{B}}_k}(X^{(k)})] < \epsilon \ , \tag{9}$$

$$\mathrm{P}(X^{(k)} \in \bar{\mathcal{B}}_k) < \epsilon, \quad k = 1, \ldots, p \ , \tag{10}$$

for some preassigned $\epsilon > 0$, $\bar{\mathcal{B}}_k$ being the complement of $\mathcal{B}_k$ in $\mathcal{D}_k$. This is possible because of Assumption $\mathcal{H}4$. Let $B_k := \{i : x_i^{(k)} \in \mathcal{B}_k\}$ and let $G_k(i)$ be the set of all indices j such that $x_i^{(k)}$ and $x_j^{(k)}$ belong to the same microaggregation group for $X^{(k)}$. We shall prove that

$$\lim_{n\to\infty} \left| \frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i) - \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i) \right| = 0 \quad \text{a. s.} \tag{11}$$

Now

$$\left| \frac{1}{n}\sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i) - \frac{1}{n}\sum_{i=1}^{n} h(\boldsymbol{x}_i) \right| \ \leq \ S_1 + S_2 + S_3 \ ,$$

where

$$\begin{aligned}
S_1 &:= \frac{1}{n} \sum_{\substack{i:G_k(i)\subset B_k \\ k=1,\ldots,p}} |h(\tilde{\boldsymbol{x}}_i) - h(\boldsymbol{x}_i)| \ , \\
S_2 &:= \frac{1}{n} \sum_{k=1}^{p} \sum_{i:G_k(i)\subset \bar{B}_k} |h(\tilde{\boldsymbol{x}}_i) - h(\boldsymbol{x}_i)| =: \sum_{k=1}^{p} S_{2k} \ , \\
S_3 &:= \frac{1}{n} \sum_{k=1}^{p} \sum_{\substack{i:G_k(i)\not\subset B_k \\ G_k(i)\not\subset \bar{B}_k}} |h(\tilde{\boldsymbol{x}}_i) - h(\boldsymbol{x}_i)| =: \sum_{k=1}^{p} S_{3k} \ .
\end{aligned}$$

We start with S_1. Because $\boldsymbol{x}_i \in \mathcal{B}$ as well as $\tilde{\boldsymbol{x}}_i \in \mathcal{B}$ for all i such that $G_k(i) \subset B_k$ for $k = 1, \ldots, p$, we have, by Assumption $\mathcal{H}2$,

$$\begin{aligned}
S_1 \ &\leq \ \frac{1}{n} H' \sum_{\substack{i:G_k(i)\subset B_k \\ k=1,\ldots,p}} \sum_{k=1}^{p} \left| x_i^{(k)} - \tilde{x}_i^{(k)} \right| \\
&\leq \ \frac{1}{n} H' \sum_{k=1}^{p} \sum_{i:G_k(i)\subset B_k} \left| x_i^{(k)} - \tilde{x}_i^{(k)} \right| \\
&\leq \frac{1}{n} H' \sum_{k=1}^{p} \sum_{i:G_k(i)\subset B_k} ||G_k(i)|| \\
&\leq \frac{1}{n} H' \sum_{k=1}^{p} K_k ||\mathcal{B}_k|| \ ,
\end{aligned}$$

where $||G_k(i)||$ is the range of the $x_j^{(k)}$ belonging to group $G_k(i)$ and $||\mathcal{B}_k||$ is the length of the interval $\mathcal{B}_k$. The last inequality follows because there are K_k elements in each $G_k(i)$. The last term converges to 0 as $n \to \infty$, so

$$S_1 < \epsilon \tag{12}$$

for sufficiently large n.

The sum S_3 is a borderline case: For each k there are at most two aggregation groups $G_k(i)$ for which $G_k(i) \not\subset B_k$ and $G_k(i) \not\subset \bar{B}_k$. (On rare occasions, there may be only one group such that some $x_i^{(k)}$ in this group lie to the left of $\mathcal{B}_k$ and some to the right of $\mathcal{B}_k$, but such a group can be treated in a similar way as the other ones). Let G_k denote the set of indices that belong to these two groups (G_k may be empty). There are at most $2K_k$ indices in G_k. Now

$$\begin{aligned} S_{3k} &= \frac{1}{n} \sum_{i \in G_k} |h(\tilde{\boldsymbol{x}}_i) - h(\boldsymbol{x}_i)| \\ &\leq \frac{1}{n} \sum_{i \in G_k} \sum_{l=1}^{p} \Big(h_l(\tilde{x}_i^{(l)}) + h_l(x_i^{(l)}) \Big) \end{aligned}$$

by Assumption $\mathcal{H}3$. As to the term $h_l(\tilde{x}_i^{(l)})$, we have

$$h_l(\tilde{x}_i^{(l)}) \;\leq\; H_l + \sum_{j=1}^{n} h_l(x_j^{(l)})\, \mathrm{I}_{\bar{\mathcal{B}}_l}(x_j^{(l)}) \;,$$

where $H_l = \max_{x^{(l)} \in \mathcal{B}_l} h_l(x^{(l)})$. This is because either $\tilde{x}_i^{(l)} \in \mathcal{B}_l$ and then $h(\tilde{x}_i^{(l)}) \leq H_l$ or $h_l(\tilde{x}_i^{(l)}) > H_l$ and then $\tilde{x}_i^{(l)} \in \bar{\mathcal{B}}_l$, and, because of the monotonicity property of h_l, there is at least one $x_j^{(l)} \in \bar{\mathcal{B}}_l$ such that $h_l(\tilde{x}_i^{(l)}) \leq h_l(x_j^{(l)})$. Consequently, $h_l(\tilde{x}_i^{(l)}) \leq \sum_{j=1}^{n} h_l(x_j^{(l)})\, \mathrm{I}_{\bar{\mathcal{B}}_l}(x_j^{(l)})$. As there are at most $2K_k$ indices in G_k, we have

$$\frac{1}{n} \sum_{i \in G_k} h_l(\tilde{x}_i^{(l)}) \;\leq\; \frac{2K_k}{n} H_l + \frac{2K_k}{n} \sum_{j=1}^{n} h_l(x_j^{(l)})\, \mathrm{I}_{\bar{\mathcal{B}}_l}(x_j^{(l)}) \;.$$

The first term on the right-hand side converges to zero and the last term con-

verges, with probability 1, to $2\,K_k\,\mathbb{E}[h_l(X^{(l)})\mathrm{I}_{\bar{\mathcal{B}}_l}(X^{(l)})]$, which is less than $2\,K_k\,\epsilon$ by Assumption (9). We have a similar bound for $\frac{1}{n}\sum_{i\in G_k} h_l(x_i^{(l)})$ (just replace K_k with 1) and so, with some constant C_1, $S_{3k} < C_1\,\epsilon$ a. s. for sufficiently large n. Hence, with some constant C_2,

$$S_3 < C_2\,\epsilon \tag{13}$$

a. s. for sufficiently large n.

Finally consider S_2. By Assumption $\mathcal{H}3$,

$$S_{2k} \;\leq\; \frac{1}{n} \sum_{i:G_k(i)\subset\bar{B}_k} \sum_{l=1}^{p} \left(h_l(\tilde{x}_i^{(l)}) + h_l(x_i^{(l)})\right). \tag{14}$$

Let us pause for a moment and let us first finish the proof for the univariate case ($p = 1$). The result will be used to prove the multivariate case. To simplify the notation, let us suppress the index $k = 1$. Since $G(i) \subset \bar{B}$ and $h(x_j)$ is monotone on each side of $\bar{\mathcal{B}}$,

$$h(\tilde{x}_i) \leq \sum_{j\in G(i)} h(x_j)\ .$$

Because each set $G(i)$ has K indices and because $G(i) \subset \bar{B}$ implies $x_i \in \bar{\mathcal{B}}$, it follows that

$$\frac{1}{n} \sum_{i:G(i)\subset\bar{B}} h(\tilde{x}_i) \;\leq\; \frac{K}{n} \sum_{i:G(i)\subset\bar{B}} h(x_i) \;\leq\; \frac{K}{n} \sum_{x_i\in\bar{\mathcal{B}}} h(x_i) \;=\; \frac{K}{n} \sum_{i=1}^{n} h(x_i)\,\mathrm{I}_{\bar{\mathcal{B}}}(x_i)\ . \tag{15}$$

The last term converges to $K\,\mathbb{E}[h(X)\mathrm{I}_{\bar{\mathcal{B}}}(X)]$ as $n \to \infty$. By similar arguments, the second sum in (14) is bounded by a term that converges to $\mathbb{E}[h(X)\mathrm{I}_{\bar{\mathcal{B}}}(X)]$. Hence S_2 is bounded by a term that converges to $(K+1)\,\mathbb{E}[h(X)\mathrm{I}_{\bar{\mathcal{B}}}(X)]$, which is less than $(K+1)\,\epsilon$ by Assumption (9). This shows that, for $p = 1$, $S_2 < (K+1)\,\epsilon$ a. s. for sufficiently large n. Together with (12) and (13) it follows that the whole sum S converges to 0 a. s. Thus, in the univariate case,

$\lim_{n\to\infty} \frac{1}{n} h(\tilde{x}_i) = \mathbb{E}[h(X)]$ a. s. (Note that we did not use Assumption $\mathcal{H}4$ in the proof for the univariate case, but rather the weaker Assumption $\mathcal{H}4'$).

Now let us continue with the multivariate case. The inequality in (14) implies

$$\begin{aligned} S_{2k} &\leq \frac{1}{n} \sum_{l=1}^{p} \sum_{i \in \bar{B}_k} \left(h_l(\tilde{x}_i^{(l)}) + h_l(x_i^{(l)}) \right) \\ &= \sum_{l=1}^{p} \left(\frac{1}{n} \sum_{i=1}^{n} h_l(\tilde{x}_i^{(l)}) \mathrm{I}_{\bar{\mathcal{B}}_k}(x_i^{(k)}) + \frac{1}{n} \sum_{i=1}^{n} h_l(x_i^{(l)}) \mathrm{I}_{\bar{\mathcal{B}}_k}(x_i^{(k)}) \right) . \end{aligned} \tag{16}$$

Considering the first sum in (16), we have by the Cauchy-Schwartz inequality that

$$\frac{1}{n} \sum_{i=1}^{n} h_l(\tilde{x}_i^{(l)}) \mathrm{I}_{\bar{\mathcal{B}}_k}(x_i^{(k)}) \;\leq\; \sqrt{\frac{1}{n} \sum_{i=1}^{n} h_l^2(\tilde{x}_i^{(l)})} \sqrt{\frac{1}{n} \sum_{i=1}^{n} \mathrm{I}_{\bar{\mathcal{B}}_k}(x_i^{(k)})} \, .$$

From what we have just proved for the univariate case (replacing h with h_l^2), the first square root converges to $\sqrt{\mathbb{E}[h_l^2(X^{(l)})]}$. The second square root converges to $\sqrt{\mathbb{E}[\mathrm{I}_{\bar{\mathcal{B}}_k}(X^{(l)})]}$, which is less than ϵ by Assumption (10). As the same can be said of the second sum in (16), it follows that $S_{2k} < C_3\, \epsilon$ and hence

$$S_2 < C_4\, \epsilon$$

a. s. for n sufficiently large.

Summing up, we have for sufficiently large n that

$$\left| \frac{1}{n} \sum_{i=1}^{n} h(\tilde{\boldsymbol{x}}_i) - \frac{1}{n} \sum_{i=1}^{n} h(\boldsymbol{x}_i) \right| < C\, \epsilon \quad \text{a. s.} \tag{17}$$

with some constant C. This proves (11).

Proof of Corollary 1:

We use the notation of the proof of Theorem 1. The data vector $\boldsymbol{x}_i$ is now equal to $\boldsymbol{x}_i = (x_i^{(1)}, x_i^{(2)})$. The partial sum S_1 is treated as before, so we only need to consider S_2 and S_3. Let $S_2 = S_{21} + S_{22}$ as in the proof of Theorem 1.

For S_{21} we have that

$$\begin{aligned} S_{21} &= \frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} |h(\tilde{\boldsymbol{x}}_i) - h(\boldsymbol{x}_i)| \\ &\leq \frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} |h(\tilde{\boldsymbol{x}}_i)| + \frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} |h(\boldsymbol{x}_i)| \, . \end{aligned}$$

Consider the first sum. By the modified assumption $\mathcal{H}_3$ and by the Cauchy-Schwartz inequality,

$$\begin{aligned} \frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} |h(\tilde{\boldsymbol{x}}_i)| &\leq \frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} h_1(\tilde{x}_i^{(1)}) h_2(\tilde{x}_i^{(2)}) \\ &\leq \sqrt{\frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} h_1^2(\tilde{x}_i^{(1)})} \sqrt{\frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} h_2^2(\tilde{x}_i^{(2)})} \\ &\leq \sqrt{\frac{K_1}{n} \sum_{i=1}^{n} h_1^2(x_i^{(1)}) \mathrm{I}_{\bar{\mathcal{B}}_1}(x_i^{(1)})} \sqrt{\frac{1}{n} \sum_{i=1}^{n} h_2^2(\tilde{x}_i^{(2)})} \, . \end{aligned}$$

The last inequality follows from similar arguments that led to (15), but with h_1^2 in place of h. The first radicand converges to $K_1 \, \mathbb{E}[h_1^2(X^{(1)}) \mathrm{I}_{\bar{\mathcal{B}}_1}(X^{(1)})]$, which is less than $K_1 \, \epsilon$ by assumption, and the second one to $\mathbb{E}[h_2^2(X^{(2)})]$ by Theorem 1. Thus $\frac{1}{n} \sum_{i:G_1(i)\subset \bar{B}_1} |h(\tilde{\boldsymbol{x}}_i)| < C_5 \, \epsilon$ a. s. for sufficiently large n. The other parts of S_2 can be treated in the same way, yielding $S_2 < C_6 \, \epsilon$ a. s. for sufficiently large n. In a similar way we can show that $S_3 < C_7 \, \epsilon$ and thus $S \leq C_8 \, \epsilon$ a. s. for sufficiently large n. This completes the proof.

Proof of Theorem 2:

We need some preliminary definitions. Let $\mathcal{B}_k = \mathcal{B}_k(n) = [b_{lk}(n), b_{uk}(n)]$ be closed finite intervals depending on n and $\mathcal{B} = \mathcal{B}_1 \times \cdots \times \mathcal{B}_p$ be the corresponding p-dimensional rectangle. (Note that $\mathcal{B}$ is not the same as the rectangle $\mathcal{B}$ used in the proof of Theorem 1). Let $b_{lk} = b_{lk}(n)$ and $b_{uk} = b_{uk}(n)$ be the following functions depending on whether the boundaries of $\mathcal{D}$ are finite or infinite:

$$b_{lk} = \begin{cases} -n^{\frac{1}{4(m_{lk}+1)}} & \text{if } d_{lk} = -\infty \\ d_{lk} + n^{-\frac{1}{4m_{lk}}} & \text{if } d_{lk} > -\infty \ , \end{cases} \tag{18}$$

$$b_{uk} = \begin{cases} n^{\frac{1}{4(m_{uk}+1)}} & \text{if } d_{uk} = \infty \\ d_{uk} - n^{-\frac{1}{4m_{uk}}} & \text{if } d_{uk} < \infty \ . \end{cases} \tag{19}$$

Note that $\mathcal{B}_k$ is contained in $\mathcal{D}_k$ and converges to $\mathcal{D}_k$ for $n \to \infty$. Let n be large enough so that $\mathcal{C}_k \subset \mathcal{B}_k(n)$ for all $k \in \{1, \ldots, p\}$. It follows from Assumption $\mathcal{H}3^*$ that $h_k(x^k) \leq C + h_{uk}(b_{uk}) + h_{lk}(b_{lk})$ for $x^k \in \mathcal{B}_k$, $r = 1 \ldots, p$, (see Remark 2), and thus

$$|h'_r(\boldsymbol{x})| \ \leq \ p\, C + \sum_{k=1}^{p} \{h_{uk}(b_{uk}) + h_{lk}(b_{lk})\} \quad \text{for } \boldsymbol{x} \in \mathcal{B} \ . \tag{20}$$

Let A_n be the event that $\frac{1}{\sqrt{n}} \left| \sum_{i=1}^n h(\tilde{\boldsymbol{x}}_i) - \sum_{i=1}^n h(\boldsymbol{x}_i) \right| > \epsilon$ and let B_n be the event that $\boldsymbol{x}_i \in \mathcal{B}$ for all $i = 1, \ldots, n$ (and hence also $\tilde{\boldsymbol{x}}_i \in \mathcal{B}$ for all i). We have to prove that $\lim_{n\to\infty} \mathrm{P}(A_n) = 0$ for all $\epsilon > 0$. Now

$$\mathrm{P}(A_n) \ \leq \ \mathrm{P}(A_n \cap B_n) + \mathrm{P}(\bar{B}_n) \ .$$

We want to prove that $\mathrm{P}(A_n \cap B_n) \to 0$ as well as $\mathrm{P}(\bar{B}_n) \to 0$. First consider the event $A_n \cap B_n$. Under this event (using a Taylor series approximation with $\boldsymbol{x}_i^* = t\tilde{\boldsymbol{x}}_i + (1-t)\boldsymbol{x}_i$, $0 < t < 1$, and using (20)),

$$\begin{aligned} \epsilon \ &< \ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} |h(\tilde{\boldsymbol{x}}_i) - h(\boldsymbol{x}_i)| \ \leq \ \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sum_{r=1}^{p} |h'_r(\boldsymbol{x}_i^*)| |\tilde{x}_i^{(r)} - x_i^{(r)}| \\ &\leq \frac{1}{\sqrt{n}} \left[pC + \sum_{k=1}^{p} \{h_{uk}(b_{uk}) + h_{lk}(b_{lk})\} \right] \sum_{i=1}^{n} \sum_{r=1}^{p} |\tilde{x}_i^{(r)} - x_i^{(r)}| \\ &\leq \frac{1}{\sqrt{n}} \left[pC + \sum_{k=1}^{p} \{h_{uk}(b_{uk}) + h_{lk}(b_{lk})\} \right] \sum_{r=1}^{p} K_r(b_{ur} - b_{lr}) \ . \end{aligned} \tag{21}$$

Now, by (18) and (19),

$$\frac{1}{\sqrt{n}}h_{lk}(b_{lk}) = \begin{cases} a_{lk}n^{-\frac{1}{4}-\frac{1}{4(m_{lk}+1)}} & \text{if } d_{lk} = -\infty \\ a_{lk}n^{-\frac{1}{4}} & \text{if } d_{lk} > -\infty \ , \end{cases} \tag{22}$$

$$\frac{1}{\sqrt{n}}h_{uk}(b_{uk}) = \begin{cases} a_{uk}n^{-\frac{1}{4}-\frac{1}{4(m_{uk}+1)}} & \text{if } d_{uk} = \infty \\ a_{uk}n^{-\frac{1}{4}} & \text{if } d_{uk} < \infty \ . \end{cases} \tag{23}$$

From (18), (19), (22), and (23) it is clear that the product terms that arise from evaluating (21) take the form

$$n^{-\frac{1}{4}-\frac{1}{4(m+1)}}n^{\frac{1}{4(m'+1)}} \leq n^{-\frac{1}{4(m+1)}} \ , \quad n^{\frac{1}{4(m+1)}}n^{-\frac{1}{4}} = n^{-\frac{m}{4(m+1)}} \ ,$$
$$n^{-\frac{1}{4m}}n^{-\frac{1}{4}-\frac{1}{4(m'+1)}} \ , \quad n^{-\frac{1}{4m}}n^{-\frac{1}{4}}$$

with positve real numbers m and m'. As all product terms go to zero with $n \to \infty$, (21) also tends to zero with $n \to \infty$. Hence $\mathrm{P}(A_n \cap B_n) \to 0$.

Next consider $\bar{B}_n$. We have

$$\begin{aligned} \mathrm{P}(\bar{B}_n) &\leq \sum_{k=1}^{p} \mathrm{P}\Big(\max_{i=1,\ldots,n} x_i^{(k)} > b_{uk} \Big) + \sum_{k=1}^{p} \mathrm{P}\Big(\min_{i=1,\ldots,n} x_i^{(k)} < b_{lk} \Big) \\ &= \sum_{k=1}^{p} \Big(1 - [F(b_{uk})]^n + 1 - [1 - F(b_{lk})]^n \Big) \ . \end{aligned}$$

With b_{lk} and b_{uk} from (18) and (19) we have, by condition $\mathcal{F}$, $[F(b_{uk})]^n \to 1$ and $[1 - F(b_{lk})]^n \to 1$ for $n \to \infty$. Thus $\mathrm{P}(\bar{B}_n) \to 0$ and hence $\mathrm{P}(A_n) \to 0$.

References

Aggarwal, C. C., Yu, P. S., 2008. Privacy-Preserving Data Mining: Models and Algorithms. Springer, New York.

Defays, D., Anwar, M. N., 1998. Masking microdata using microaggregation. Journal of Official Statistics 14 (4), 449–461.

Domingo-Ferrer, J., Martinez-Balleste, A., Mateo-Sanz, J. M., Sebe, F., 2006. Efficient multivariate data-oriented microaggregation. International Journal on Very Large Data Bases 15 (4), 355–369.

Domingo-Ferrer, J., Mateo-Sanz, J. M., 2002. Practical data-oriented microaggregation for statistical disclosure control. IEEE Transactions on Knowledge and Data Engineering 14 (1), 189–201.

Domingo-Ferrer, J., Oganian, A., Torres, A., Mateo-Sanz, J. M., 2002. On the security of microaggregation with individual ranking: Analytical attacks. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10 (5), 477–491.

Domingo-Ferrer, J., Sebe, F., Solanas, A., 2008. A polynomial-time approximation to optimal multivariate microaggregation. Computers & Mathematics with Applications 55 (4), 714–732.

Domingo-Ferrer, J., Torra, V., 2001. A quantitative comparison of disclosure control methods for microdata. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (Eds.), Confidentiality, Disclosure, and Data Access. North-Holland, Amsterdam, pp. 111–133.

Domingo-Ferrer, J., Torra, V., 2004. Privacy in Statistical Databases. Springer, Berlin.

Doyle, P., Lane, J., Theeuwes, J., Zayatz, L., 2001. Confidentiality, Disclosure, and Data Access. North-Holland, Amsterdam.

Durrett, R., 1991. Probability: Theory and Examples. Wadsworth & Brooks/Cole, Pacific Grove.

Felsö, F., Theeuwes, J., Wagner, G. G., 2001. Disclosure limitation methods in use: Results of a survey. In: Doyle, P., Lane, J., Theeuwes, J., Zayatz, L. (Eds.), Confidentiality, Disclosure, and Data Access. North-Holland, Amsterdam, pp. 17–42.

Fritsch, M., Stephan, A., 2003. Die Heterogenität der technischen Effizienz innerhalb von Wirtschaftszweigen - Auswertungen auf Grundlage der Kostenstrukturstatistik des Statistischen Bundesamtes. In: Pohl, R., Fischer, J., Rockmann, U., Semlinger, K. (Eds.), Analysen zur regionalen Industrieentwicklung - Sonderauswertungen einzelbetrieblicher Daten der amtlichen Statistik. Statistisches Landesamt, Berlin, pp. 143–156, in German.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G., De Wolf, P.-P., 2009. Handbook on Statistical Disclosure Control. `http://neon.vb.cbs.nl/CASC/.%5CSDC_Handbook.pdf`.

Laszlo, M., Mukherjee, S., 2005. Minimum spanning tree partitioning algorithm for microaggregation. IEEE Transactions on Knowledge and Data Engineering 17 (7), 902–911.

Martinez-Balleste, A., Solanas, A., Domingo-Ferrer, J., Mateo-Sanz, J. M., 2007. A genetic approach to multivariate microaggregation for database privacy. In: 23rd Internationl Conference on Data Engineering. Workshop on Privacy Data Management. IEEE Computer Society, pp. 180–185.

Mateo-Sanz, J. M., Domingo-Ferrer, J., 1998. A comparative study of microag-

gregation methods. Questiio 22 (3), 511–526.

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M., Vorgrimler, D., 2005. Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. Statistik und Wissenschaft 4. Statistisches Bundesamt, Wiesbaden, in German.

Rosemann, M., Lenz, R., Vorgrimler, D., Sturm, R., 2006. Anonymising business micro data - results of a German project. Journal of Applied Social Sciences Studies 126 (4), 635–651.

Schmid, M., 2006. Estimation of a linear model under microaggregation by individual ranking. Journal of the German Statistical Society 90 (3), 419–438.

Schmid, M., Schneeweiss, H., 2008. Estimation of a linear model in transformed variables under microaggregation by individual ranking. AStA Advances in Statistical Analysis 92 (4), 359–374.

Schmid, M., Schneeweiss, H., Küchenhoff, H., 2007. Estimation of a linear regression under microaggregation with the response variable as a sorting variable. Statistica Neerlandica 61 (4), 407–431.

Strudler, M., Oh, H. L., Scheuren, F., 1986. Protection of taxpayer confidentiality with respect to the tax model. Proceedings of the Section on Survey Research Methods of the American Statistical Association, 375–381.

UNECE Secretariat, 2001. Statistical data confidentiality in the transition countries: 2000/2001 winter survey. In: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality. Skopje, Macedonia.

Willenborg, L., de Waal, T., 2001. Elements of Statistical Disclosure Control. Springer, New York.

Winkler, W. E., 2002. Single-ranking micro-aggregation and reidentification. Statistical Research Division Report RR 2002/08, U.S. Bureau of the Census, Washington.

List of Tables

Tables

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	4.904	4.978	5.009	5.008	5.032	5.157
Microaggregated data	4.902	4.987	5.011	5.017	5.049	5.155

Table 1
Simulation study on quadratic regression - summary statistics of the 100 least squares estimates of β in Model (5). The standard deviations of the least squares estimates, multiplied with $\sqrt{n} = \sqrt{300}$, were 0.792 (original data) and 0.855 (microaggregated data).

Estimates of α

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	0.372	0.468	0.510	0.514	0.561	0.729
Microaggregated data	0.394	0.479	0.517	0.520	0.564	0.729

Estimates of β

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	1.324	1.757	1.972	1.966	2.167	2.676
Microaggregated data	1.320	1.736	1.945	1.938	2.105	2.622

Table 2
Simulation study on the shape and scale parameter estimation of a gamma distribution - summary statistics of the 100 method-of-moments estimates ($\alpha = 0.5$, $\beta = 2$).

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Original data	1.710	1.900	2.002	2.012	2.124	2.375
Microaggregated data	1.712	1.905	2.012	2.018	2.134	2.377

Table 3
Simulation study on maximum likelihood estimation of the scale parameter c of a Levy distribution - summary statistics of the 100 maximum likelihood estimates ($c = 2$).

	β_1	β_2	β_3	β_4	β_5
non-aggregated data	0.426	0.313	0.065	0.131	0.046
IR (log trafo after MA)	0.426	0.313	0.065	0.131	0.046

Table 4
Least squares estimates of Model (8) obtained from the 2004 KSE data. The abbreviation "MA" stands for "microaggregation". The coefficient estimates have been rounded to two decimal places, so the results obtained from the non-aggregated data are not exactly the same as the results obtained from the microaggregated data.

	σ_{β_1}	σ_{β_2}	σ_{β_3}	σ_{β_4}	σ_{β_5}
non-aggregated data	0.0021	0.0036	0.0015	0.0023	0.0027
IR (log trafo after MA)	0.0021	0.0036	0.0015	0.0023	0.0027

Table 5
Estimated standard deviations of the least squares estimates of Model (8) obtained from the 2004 KSE data. The abbreviation "MA" stands for "microaggregation". All estimates have been rounded to two decimal places, so the results obtained from the non-aggregated data are not exactly the same as the results obtained from the microaggregated data.