ACTA ET COMMENTATIONES UNIVERSITATIS TARTUENSIS DE MATHEMATICA
Volume 12, 2008

Linear models with measurement errors arising from mixture distributions

Gerd Ronning

Abstract. The paper considers the linear model with multiplicative measurement errors. In particular, errors arising from mixture distributions will be analyzed. Such a model has to be used if the micro data have been protected by multiplicative noise. If all (continuous) variables are anonymized jointly by this approach, measurement errors will be correlated which is of special concern if the dependent variable is measured with error, too. The paper presents results for the biased naive least-squares estimator both in case of cross-section data and panel data. Moreover, derivation of consistent estimators is shortly discussed.

1. Introduction

Empirical research in economics has for a long time suffered from the unavailability of individual micro data and has forced econometricians to use (aggregate) time series data in order to estimate, for example, a consumption function. On the contrary other disciplines like psychology, sociology and, last not least, biometry have analyzed micro data already for decades. The software for microeconometric models has created growing demand for micro data in economic research, in particular data describing firm behaviour. However, such data are not easily available when collected by the Statistical Office because of confidentiality.

On the other hand these data would be very useful for testing microeconomic models. Therefore, the German Statistical Office initiated research on the question whether it is possible to produce files for scientific use from these data. The files have to be anonymized in a way that re-identification is almost impossible and, at the same time, distributional properties of the data do not change too much. In particular, data from enterprises have been

Received October 11, 2007.

2000 *Mathematics Subject Classification.* 62J05, 62P20.

Key words and phrases. Biased estimation, data protection, error factor model, panel data.

considered in this project. Results have been published quite recently (see Ronning et al. (2005)). Most known anonymization procedures have been rated both with regard to data protection and to informational content left after perturbation. In particular, micro-aggregation or addition of stochastic noise has been found convenient for continuous variables whereas "Post Randomization" (PRAM) can be recommended with some reservations for discrete variables.[1]

In case of anonymization by stochastic noise we have the situation of "errors in variables". In particular, errors in the regressor variables of a linear model will lead to biased estimation of parameters. Usually, additive measurement errors are considered whereas in our project we favor perturbation by multiplicative noise. More formally, the anonymized random variable X^a is obtained from

$$X^a \quad = \quad X \cdot U \ , \tag{1.1}$$

where X is the "original" variable and U an error with $E[U] = 1$. We consider this approach as superior to *addition* of noise since larger values are much better protected. For example, if two firms with sales of 1 million and 100 million Euro will be anonymized multiplicatively, both will receive an error of, say, ± 10 % whereas in the additive case an error of, say, ± 10,000 Euro ist added to both sales. Of course, the larger sale will be almost unaffected by this error.

In this paper we consider only stochastic noise generated from a mixture distribution which has been first suggested by Roque (2000). In particular, a **bimodal** mixture distribution will move anonymized values away from the original values with high probability.[2] If this method is applied to a set of variables then a multivariate (bimodal) mixture distribution will be involved. It will be shown below that such a distribution will imply correlation of measurement errors. This is of special concern if linear (or nonlinear) models are estimated from data anonymized in this way. Note that usually measurement errors are assumed to be independent across variables. As we will see, in particular the measurement error of the dependent variable no longer can be considered as harmless to estimation.

First we consider estimation from anonymized cross-section data, secondly the case of panel data is examined. For both cases we present results with regard to estimation of linear models. Moreover, we consider the possibility of constructing corrected unbiased estimators. All results in this paper are presented with proofs in Ronning (2007a).

[1] Additionally, most recently multiple imputation has been suggested by Donald Rubin for data protection.

[2] The idea has been suggested also by Massell, Zayatz und Funk (2006) although the authors do not explicitly refer to a mixture distribution.

The paper is organized as follows: In Section 2 we present an error factor model which has the attractive property of preserving proportionality among variables and which can be shown to be equivalent to anonymization by a multivariate mixture distribution. Section 3 reports results of estimating linear models from anonymized cross-section data and Section 4 presents modifications when panel data are used. The Appendix contains some results regarding mixture distributions for easier reference.

2. A factor model of measurement errors

We start by considering one single error variable U_j which is generated from the following model:

$$U_j \quad = \quad 1 \, + \, \delta \, D_j \, + \, \varepsilon_j \ ,$$

where δ is a parameter and D_j a random variable satisfying

$$D_j \; = \; \begin{cases} +1 & \text{with probability } \gamma \\ -1 & \text{with probability } 1-\gamma \end{cases}$$

and ε_j is a continuous random variable with

$$E[\varepsilon_j] \; = \; 0 \ , \quad V[\varepsilon_j] \; = \; \sigma_\varepsilon^2 \ .$$

Since U_j in (1.1) is applied multiplicatively, the parameter δ determines a relative increase or decrease. For example, $\delta = 0.12$ means a change of ± 12 % for variable x. The error term ε adds some additional noise. In the following we set $\gamma = 0.5$ which implies $E[U_j] = 1$.

Let us now assume that the data set contains r different (continuous) variables which have to be anonymized. Following an idea first mentioned by Jörg Höhne the different U_j are generated by

$$U_j \quad = \quad 1 \, + \, \delta \, D \, + \, \varepsilon_j \ , \ j = 1, \ldots, r \ , \tag{2.1}$$

that is, the same D is used for all variables in order to preserve proportionality of the variables at least approximately. For the ratio $Z = X/Y$ of the two variables X and Y the following should hold:

$$E[Z] = E\left[\frac{X}{Y}\right] \quad \approx \quad E[Z^a] \; = \; E\left[\frac{X \cdot (1 \, + \, \delta D \, + \, \varepsilon_X)}{Y \cdot (1 \, + \, \delta D \, + \, \varepsilon_Y)}\right] \ .$$

Ronning (2007b) shows that this is true indeed. However, already E[Z] differs notably from the much more relevant ratio $E[X]/E[Y]$ if correlation between X and Y is low or – even worse – negative.

Note that the variable D can be seen as a “common factor” which implies (positive) correlation between the two error variables U_j and U_k. This result also follows from the fact that the above error factor model (2.1) is equivalent to the statement that the vector $\mathbf{U} = (U_1, \ldots, U_r)$ has a multivariate mixture distribution (for a definition see Appendix). From (3A) in the Appendix it is evident that the covariance matrix can only be diagonal if the mean vector equals the null vector! Therefore, in general any two of its elements will be correlated.

We now prove the statement that the factor model (1.1) is equivalent to a multivariate mixture distribution. First note that the probability density of D in (2.1) is given by

$$h(d) \quad = \quad \gamma^{\frac{1+d}{2}} \, (1 \, - \, \gamma)^{\frac{1-d}{2}} \tag{2.2}$$

for $d \, \in \, \{+1 \, , \, -1\}$. For convenience we also assume normality for the vector $\boldsymbol{\varepsilon}$ which we write as

$$\boldsymbol{\varepsilon} \; \sim \; N(\mathbf{0} \, , \, \sigma_{\varepsilon}^2 \mathbf{I}) \; . \tag{2.3}$$

Notation (2.1) is used also for the density of ε. The conditional density of $\mathbf{U}$ given $D = d$ is given by[3]

$$\mathbf{U} | (D = d) \; \sim \; N((1 \, + \, \delta \, d) \, \boldsymbol{\iota} \; , \; \sigma_{\varepsilon}^2 \mathbf{I})$$

and for the joint density of $\mathbf{U}$ and D we obtain

$$g(\mathbf{u}, d) \quad = \quad h(d) \, \cdot N((1 \, + \, \delta \, d) \, \boldsymbol{\iota} \; , \; \sigma_{\varepsilon}^2 \mathbf{I}) \; .$$

The marginal density of $\mathbf{U}$ then is given by

$$\begin{aligned} f(\mathbf{u}) \quad &= \quad \textstyle\sum_{d \, \epsilon \, \{+1 \, , -1\}} \alpha^{\frac{1+d}{2}} \, (1 \, - \, \alpha)^{\frac{1-d}{2}} \, N((1 \, + \, \delta \, d) \, \boldsymbol{\iota} \; , \; \sigma_{\varepsilon}^2 \mathbf{I}) \\ &= \quad \alpha \, N((1 \, + \, \delta) \boldsymbol{\iota} \, , \, \sigma_{\varepsilon}^2 \mathbf{I}) \; + \; (1 - \alpha) \, N((1 \, - \, \delta) \boldsymbol{\iota} \, , \, \sigma_{\varepsilon}^2 \mathbf{I}) \; . \end{aligned}$$

This density has the form (2.1) of a multivariate mixture distribution with $k = 2$.

3. Estimation of linear models

We consider the linear model

$$\mathbf{y} \quad = \quad \beta_0 \, \boldsymbol{\iota} \; + \; \mathbf{X} \boldsymbol{\beta} \; + \; \boldsymbol{\eta} \tag{3.1}$$

where $\mathbf{X}$ is an $(n \times K)$ matrix, that is, we have n observations and K regressors. For $\boldsymbol{\eta}$ we assume:

$$E[\boldsymbol{\eta}] \; = \; \mathbf{0} \quad \text{and} \quad cov[\boldsymbol{\eta}] \; = \; \sigma_{\eta}^2 \, \mathbf{I}. \tag{3.2}$$

The least squares estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} \quad = \quad \left(\mathbf{X}'\mathbf{M}_{\iota}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{M}_{\iota}\mathbf{y} \; = \; \boldsymbol{\beta} \; + \; \left(\mathbf{X}'\mathbf{M}_{\iota}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{M}_{\iota}\boldsymbol{\eta} \tag{3.3}$$

with

$$\mathbf{M}_{\iota} \; = \; \mathbf{I}_n \; - \; (1/n) \, \boldsymbol{\iota}\boldsymbol{\iota}' \; .$$

In the following we assume that only anonymized variables $\mathbf{y}^a$ and $\mathbf{X}^a$ are available from

$$\mathbf{y}^a \; = \; \mathbf{y} \, \odot \, \mathbf{u}_y \quad \text{and} \quad \mathbf{X}^a \; = \; \mathbf{X} \, \odot \, \mathbf{U}_{\mathbf{x}} \; , \tag{3.4}$$

[3] The symbol $\boldsymbol{\iota}$ denotes a vector of ones.

where $\odot$ denotes the Hadamard product. More exactly, the vector $\mathbf{y}^a$ and the matrix $\mathbf{X}^a$ are given by

$$\begin{aligned}(\mathbf{y}^a)' &= (y\langle 1\rangle \cdot u_y\langle 1\rangle, y\langle 2\rangle \cdot u_y\langle 2\rangle, \ldots, y\langle n-1\rangle \cdot u_y\langle n-1\rangle, y\langle n\rangle \cdot u_y\langle n\rangle) \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (3.5) \\ (\mathbf{X}^a)' &= (\mathbf{x}^a\langle 1\rangle, \ldots, \mathbf{x}^a\langle K\rangle) = (\mathbf{x}\langle 1\rangle \odot \mathbf{u}_{\mathbf{x}}\langle 1\rangle, \ldots, \mathbf{x}\langle n\rangle \odot \mathbf{u}_{\mathbf{x}}\langle n\rangle),\end{aligned}$$

where $\mathbf{x}\langle i\rangle$ is K-dimensional and contains the elements of the i-th row of the matrix $\mathbf{X}$. Correspondingly, the vector $\mathbf{u}_{\mathbf{x}}\langle i\rangle$ contains the elements from the i-th row of $\mathbf{U}$. For each observation i, $i = 1, \ldots, n$, now the error factor model (2.1) holds:

$$\begin{aligned} u_y\langle i\rangle &= 1 + \delta\, D\langle i\rangle + \varepsilon_y\langle i\rangle\ , \\ \mathbf{u}_{\mathbf{x}}\langle i\rangle &= (1 + \delta\, D\langle i\rangle)\,\boldsymbol{\iota} + \boldsymbol{\varepsilon}_{\mathbf{x}}\langle i\rangle\ . \end{aligned} \tag{3.6}$$

If we use the "naive" least squares estimator

$$\hat{\boldsymbol{\beta}}^a = \left(\mathbf{X}^{a\prime}\mathbf{M}_\iota\mathbf{X}^a\right)^{-1} \mathbf{X}^{a\prime}\mathbf{M}_\iota\mathbf{y}^a$$

and assume that $\mathbf{Q}$ from

$$\text{plim}\, \frac{1}{n}\mathbf{X}'\mathbf{M}_\iota\mathbf{X} \equiv \mathbf{Q} \tag{3.7}$$

is nonsingular, then we obtain (see Ronning (2007a) for details)

$$\text{plim}\, \hat{\boldsymbol{\beta}}^a = \mathbf{S}^{-1}\left(\mathbf{Q}\,\boldsymbol{\beta} + \{\beta_0\,\boldsymbol{\mu}_x + (\mathbf{Q} + \boldsymbol{\mu}_x\boldsymbol{\mu}_x')\boldsymbol{\beta}\} \odot cov[\mathbf{u}_x, u_y]\right) \tag{3.8}$$

with

$$\mathbf{S} = cov[\mathbf{u}_x] \odot \left\{\mathbf{Q} + \boldsymbol{\mu}_x\,\boldsymbol{\mu}_x'\right\} + \mathbf{Q}$$

where $\boldsymbol{\mu}_x$ is the mean vector of the K regressors, $cov[\mathbf{u}_x]$ is the $(n \times n)$ covariance matrix of $\mathbf{u}_x$ and $cov[\mathbf{u}_x, u_y]$ denotes the n-dimensional vector of covariances of $\mathbf{u}_x$ with u_y. Consistency will only be obtained if the two just mentioned expressions regarding covariances are equal to zero. For the error factor model (2.1) these two expressions are given by

$$cov[\mathbf{u}_x] = \sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota}\boldsymbol{\iota}'\ , \quad cov[\mathbf{u}_x, u_y] = \delta^2\,\boldsymbol{\iota}\ .$$

Therefore, consistency will be obtained if both δ and σ_ε^2 are zero. If only the vector $cov[\mathbf{u}_x, u_y]$ is zero, then we have the result for the special case when only the regressors are anonymized. Note that in the multiplicative case considered here also the parameter β_0 from (3.1) influences the bias of the "naive" estimator.

4. Estimation of linear panel models

We now consider the linear panel model with individual random effects which we write as

$$y_{it} \quad = \quad \beta_0 \, + \, \sum_{k=1}^{K} \beta_k \, x_{itk} \, + \, \tau_i \, + \, \eta_{it} \, , \, i = 1, \ldots , n \, , \, t = 1, \ldots T \ , \tag{4.1}$$

or more compactly

$$\mathbf{y} \quad = \quad \beta_0 \, \boldsymbol{\iota} \, + \, \mathbf{X}\boldsymbol{\beta} \, + \, \boldsymbol{\tau} \, + \, \boldsymbol{\eta} \tag{4.2}$$

with

$$\mathbf{X} \quad = \quad \begin{pmatrix} \mathbf{X}\langle 1 \rangle \\ \mathbf{X}\langle 2 \rangle \\ \vdots \\ \mathbf{X}\langle n-1 \rangle \\ \mathbf{X}\langle n \rangle \end{pmatrix} \ , \quad \mathbf{y} \quad = \quad \begin{pmatrix} \mathbf{y}\langle 1 \rangle \\ \mathbf{y}\langle 2 \rangle \\ \vdots \\ \mathbf{y}\langle n-1 \rangle \\ \mathbf{y}\langle n \rangle \end{pmatrix} \ ,$$

$$\mathbf{X}\langle i \rangle \quad = \quad \begin{pmatrix} \mathbf{x}_1\langle i \rangle & \mathbf{x}_2\langle i \rangle & \ldots & \mathbf{x}_{K-1}\langle i \rangle & \mathbf{x}_K\langle i \rangle \end{pmatrix} \ , \quad i = 1, \ldots , n \ ,$$

$$\mathbf{x}'_k\langle i \rangle \quad = \quad (x_{i1k}, x_{i2k}, \ldots , x_{i,T-1,k}, x_{iTk}) \ , \quad k = 1, \ldots , K \ ,$$

$$\mathbf{y}'\langle i \rangle \quad = \quad (y_{i1}, y_{i2}, \ldots , y_{i,T-1}, y_{iT}) \ .$$

Moreover, the random vector of individual effects has the following form:

$$\boldsymbol{\tau} \quad = \quad \begin{pmatrix} \boldsymbol{\tau}\langle 1 \rangle \\ \\ \boldsymbol{\tau}\langle 2 \rangle \\ \\ \vdots \\ \boldsymbol{\tau}\langle n-1 \rangle \\ \\ \boldsymbol{\tau}\langle n \rangle \end{pmatrix} \quad \text{and} \quad \boldsymbol{\tau}\langle i \rangle \ = \ \tau_i \, \boldsymbol{\iota}_T \ .$$

We estimate the vector $\boldsymbol{\beta}$ from (4.2) by the so-called "within"-estimator

$$\hat{\boldsymbol{\beta}}_W \quad = \quad (\mathbf{X}' \, \mathbf{M}_W \, \mathbf{X})^{-1} \, \mathbf{X}' \, \mathbf{M}_W \, \mathbf{y} \, , \tag{4.3}$$

where the symmetric idempotent matrix $\mathbf{M}_W$ is given by

$$\mathbf{M}_W \quad = \quad \mathbf{I}_{nT} - \mathbf{W} \, (\mathbf{W}' \, \mathbf{W})^{-1} \, \mathbf{W}' \ = \ \mathbf{I}_n \otimes \left(\mathbf{I}_T \, - \, \frac{1}{T} \, \boldsymbol{\iota}_T \, \boldsymbol{\iota}'_T \right)$$

with

$$\mathbf{W} \quad = \quad \mathbf{I}_n \otimes \boldsymbol{\iota}_T \, .$$

If only anonymized variables are available, we use the "naive" estimator

$$\hat{\boldsymbol{\beta}}^a_W \quad = \quad (\mathbf{X}^{a\prime}\,\mathbf{M}_W\,\mathbf{X}^a)^{-1}\,\mathbf{X}^{a\prime}\,\mathbf{M}_W\,\mathbf{y}^a\,, \tag{4.4}$$

which has the same structure as (4.3).

Again we apply the error factor model which however now has to be specified for each period t. Since proportionality should also be preserved over periods (compare the discussion in Section 2), the following specification is used for anonymization:

$$\begin{array}{rcl} x^a_{itk} & = & x_{itk}\,(1\,+\,\delta D_i\,+\,\varepsilon_{itk}) \quad ,k=1,\dots,K, \\ y^a_{it} & = & y_{it}\,(1\,+\,\delta D_i\,+\,\varepsilon_{ity})\;. \end{array}$$

Note that the same random variable D_i is used for all x's as well as for y in all T periods!

In Ronning (2007a) it is shown that the probability limit (for $n \to \infty$) of the (naive) "within" estimator (4.4) is given by

$$\operatorname{plim}\hat{\boldsymbol{\beta}}^a_W \;=\; \left(cov[\mathbf{x}] \,+\, \tfrac{\sigma^2_\varepsilon}{(1+\delta^2)} \begin{pmatrix} \sigma_1^2+\mu_1^2 & & \\ & \ddots & \\ & & \sigma_K^2+\mu_K^2 \end{pmatrix} \right)^{-1} cov[\mathbf{x}]\,\boldsymbol{\beta}\,, \tag{4.5}$$

where $cov[\mathbf{x}]$ contains the second moments of the K regressors, that is,

$$cov[\mathbf{x}] \quad = \quad \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \dots & \sigma_K^2 \end{pmatrix} .$$

If only the regressors are anonymized, the probability limit is

$$\operatorname{plim}\hat{\boldsymbol{\beta}}^a_W = \tfrac{1}{1+\delta^2} \left(cov[\mathbf{x}] + \tfrac{\sigma^2_\varepsilon}{(1+\delta^2)} \begin{pmatrix} \sigma_1^2+\mu_1^2 & & \\ & \ddots & \\ & & \sigma_K^2+\mu_K^2 \end{pmatrix} \right)^{-1} cov[\mathbf{x}]\boldsymbol{\beta} \tag{4.6}$$

and therefore differs only by the factor $1/(1+\delta^2) < 1$ from the result in (4.5).

Note that both probability limits have the form

$$(\mathbf{A}\,+\,\mathbf{B})^{-1}\,\mathbf{A}\,\boldsymbol{\beta}$$

with both **A** and **B** being positive definite so that the bias has the well-known "shrinkage property" of the standard errors-in variables model with additive measurement errors whereas for cross-section data even the sign of the (asymptotic) bias cannot be assured (see (3.8)). Moreover consistency of the naive estimator (4.4) is obtained in both cases for **any** value of δ if the variance σ^2_ε is equal to zero.

5. Concluding remarks

Since for anonymized data the parameters of the anonymization method will be known, it is straightforward – at least in linear models – to derive bias-corrected (and therefore consistent) estimators from the above results. We will illustrate this for the estimation from panel data discussed in Section 4.

From (4.5) we have immediately the following consistent estimator for the case of all variables being jointly anonymized:

$$\hat{\boldsymbol{\beta}}_W^{a,corr} = cov[\mathbf{x}]^{-1}\left(cov[\mathbf{x}] + \frac{\sigma_\varepsilon^2}{1+\delta^2}\begin{pmatrix}\sigma_1^2+\mu_1^2 & & \\ & \ddots & \\ & & \sigma_K^2+\mu_K^2\end{pmatrix}\right)\hat{\boldsymbol{\beta}}_W^a \, . \tag{5.1}$$

Although δ and σ_ε^2 are known, this estimator cannot be used directly since $cov[\mathbf{x}]$ and $\boldsymbol{\mu}_x$ are unknown. However employing the method of moments we obtain from

$$\begin{aligned}\widehat{cov[\mathbf{X}^a]} &= \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\boldsymbol{\iota}\boldsymbol{\iota}'\right)\odot\left(\widehat{cov[\mathbf{x}]} + \widehat{\boldsymbol{\mu}_x}\widehat{\boldsymbol{\mu}_x}'\right) + \widehat{cov[\mathbf{x}]} \\ &= \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\boldsymbol{\iota}\boldsymbol{\iota}' + \boldsymbol{\iota}\boldsymbol{\iota}'\right)\odot\widehat{cov[\mathbf{x}]} + \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\boldsymbol{\iota}\boldsymbol{\iota}'\right)\odot\widehat{\boldsymbol{\mu}_x}\widehat{\boldsymbol{\mu}_x}'\end{aligned} \tag{5.2}$$

the following estimator for $cov[\mathbf{x}]$:

$$\widehat{cov[\mathbf{x}]} = \left\{\widehat{cov[\mathbf{X}^a]} - \left(\sigma_\varepsilon^2\mathbf{I} + \delta^2\,\boldsymbol{\iota}\boldsymbol{\iota}'\right)\odot\widehat{\boldsymbol{\mu}_x}\widehat{\boldsymbol{\mu}_x}'\right\} \div \left(\sigma_\varepsilon^2\mathbf{I} + (1+\delta^2)\,\boldsymbol{\iota}\boldsymbol{\iota}'\right) \, , \tag{5.3}$$

where $\div$ denote element-wise division (Hadamard division) and $\widehat{\ }$ indicate estimates. Additionally, it should be exploited that the mean of the "original" regressors and the anonymized regressors are equal and therefore

$$\widehat{\boldsymbol{\mu}_x} = \widehat{\boldsymbol{\mu}_x^a} \, .$$

Substituting these estimates for the unknown moments in (5.1) leads to an operational form of the consistent estimator of $\boldsymbol{\beta}$ in (4.2).

Note that such explicit solutions are only possible in case of linear models. For nonlinear models the SIMEX procedure first proposed by Cook and Stefanski (1994) (see also Carroll et al. (2006)) could be applied. However, in case of correlated errors some modifications are necessary since in the simulation step the correlation between errors of different variables should be taken into account. This has been examined in Ronning and Rosemann (2008).

Appendix. Mixture distributions

For an arbitrary number k of random variables W_i with density function $f_i(w)$ the density of a mixture of these random variables is given by

$$g(u) \quad = \quad \sum_{i=1}^{k} \alpha_i \, f_i(u) \ , \quad 0 < \alpha_i < 1 \ , \ \sum_i \alpha_i \ = \ 1 \ ,$$

with expectation

$$E[U] \quad = \quad \sum_{i=1}^{k} \alpha_i \, \mu_i \ ,$$

and variance

$$V[U] \quad = \quad \sum_{i=1}^{k} \alpha_i \, (\sigma_i^2 \, + \, \mu_i^2) \, + \, \left(\sum_i \alpha_i \, \mu_i \right)^2 \ .$$

In the multivariate case the r-dimensional random vector $\mathbf{U}$ follows a (multivariate) mixture distribution if its joint density is given by

$$g(u_1, u_2, \dots u_r) \quad = \quad \sum_{i=1}^{k} \alpha_i \, f_i(u_1, u_2, \dots , u_r) \tag{1A}$$

with expectation

$$E[\mathbf{U}] \quad = \quad \sum_{i=1}^{k} \alpha_i \, \boldsymbol{\mu}_i \tag{2A}$$

and covariance matrix

$$cov[\mathbf{U}] \quad = \quad \sum_{i=1}^{k} \alpha_i \, \left(\boldsymbol{\Sigma}_i \, + \, \boldsymbol{\mu}_i \, \boldsymbol{\mu}_i{}' \right) \, - \, \left(\sum_{i=1}^{k} \alpha_i \, \boldsymbol{\mu}_i \right) \left(\sum_{i=1}^{k} \alpha_i \, \boldsymbol{\mu}_i \right)' \tag{3A}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ denote the expectation and the covariance matrix, respectively, of the i-th random vector $\mathbf{W}_i$, $i = 1, \dots , k$.

Acknowledgements. Results in this paper are related to the project "Wirtschaftsstatistische Paneldaten und faktische Anonymisierung" which is financially supported by the German Ministry of Education and Research. The critical remarks of an anonymous referee are gratefully acknowledged.

References

Carroll, R. J., Ruppert, D., Stefanski, L. A. and Crainiceanu, C. M. (2006), *Measurement Error in Nonlinear Models. A Modern Perspective*, London, Chapman and Hall.

Cook, J. R. and Stefanski, L. A.(1994). *Simulation-extrapolation estimation in parametric measurement error models*, J. Amer. Statist. Assoc. **89**, 1314–1328.

Massell, P., Zayatz, L. and Funk, J. (2006), *Protecting the confidentiality of survey tabular data by adding noise to the underlying micro data: Application to the commodity flow survey*; In: Privacy in Statistical Data Bases (eds. J. Domingo and L. Franconi), Springer, Berlin, pp. 304–317.

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. and Vorgrimler, D. (2005), *Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten*, Series "Statistik und Wissenschaft", Volume 4, German Statistical Office, Wiesbaden.

Ronning, G. (2007a), *Stochastische Überlagerung mit Hilfe der Mischungsverteilung*, IAW Discussion Paper No. 30 (October 2007). IAW, Tübingen. http://www.iaw.edu/pdf/dp2007-30.pdf

Ronning, G. (2007b), *Measuring Research Intensity From Anonymized Data: Does Multiplicative Noise With Factor Structure Save Results Regarding Quotients?*, Mimeo, University of Tübingen, July 2007.

Ronning, G. and Rosemann, M. (2008), *SIMEX estimation in case of correlated measurement errors*, AStA Adv. Stat. Anal. **92**, 391–404.

Roque, G. M. (2000), *Masking Microdata Files with Mixtures of Multivariate Normal Distributions*, Dissertation, June 2000, University of California, Riverside.

Department of Economics, University of Tübingen, Germany
E-mail address: gerd.ronning@uni-tuebingen.de