Measurement Error in A Bivariate Model – Application in Nutrition Epidemiology

Hui Lin

March 21, 2014

Abstract

We consider the problem of estimating the joint distribution of two correlated random variables where one is observed with error. An interest in human nutrition is to estimate the joint distribution of usual energy intake and usual micronutrient intake. While precise biomarkers for energy consumption are available, there are no reliable biomarkers of consumption for nutrients including vitamins and minerals (vitamin K is an exception). Yet, nutritionists are interested in estimating the distribution of usual intake of micronutrients per unit of caloric intake. This is denoted the nutrient density of the diet and involves estimation of the distribution of the ratio of two non-normal random variables, one of which is observed with measurement error. We develop an approach that combines a deconvolution kernel method (DKM) and the method of copulas to estimate the joint distribution of two non-normal variables where one is contaminated. DKM is first used to adjust the univariate measurement error. A Gaussian copula is then used to model the correlation structure between the two variables after error adjustment. We carried out a small simulation study to investigate whether the two-step method we propose is promising. At least in the context of our simulation, we found that the approach produces good results when the correlation between the two random variables is reasonably high. Our findings are tentative, however and more research is needed before we can recommend the methodology for use more broadly.

1 Introduction

We consider estimation of a bivariate non-normal distribution given pairs of observations where one of the variables is contaminated with measurement error. This problem, which falls in the general category of a deconvolution

problem, arises frequently in applications since, in practice, we often find that variables of interest are subject to measurement error.

Our work is motivated by the need to estimate a bivariate usual intake distribution. There has been a lot of work in this area, but research has mostly focused on the univariate case (see below for a review of some of the literature). For practical reasons – cost and respondent burden among them – intake data are collected from individuals in a sample of the population for only a few days per individual. Even though intake information for each individual in the sample is limited, epidemiologists and nutritionists are typically interested in the long-run average intake, denoted usual intake, and in particular, in the distribution of usual intakes in the population. Given an estimate of the distribution of usual intakes, it is then possible to estimate, for example, the proportion of the population whose intakes fall below a threshold such as the estimated average requirement (EAR). Excessive intakes, as in the case of cholesterol and sodium, are also of interest, and the proportion of the population with high intakes of a nutrient can also be assessed from the usual intake distribution.

One simple approach to estimate the distribution of usual intakes is to use the distribution of observed individual mean intakes as the estimator. However, even if we were to assume that the observed intake is unbiased for usual intake, an individual's mean daily intake for a particular dietary component has a variance that contains some within-individual variability. Thus, the variance of the observed means is inflated by the day-to-day variability in daily intake. Because of this, using the distribution of the mean of a few days as an estimate of the usual intake distribution can lead to erroneous inference regarding dietary status.

In the univariate case, adjustment for measurement error can be formulated as the problem of estimating the distribution of a random variable that is observed with error. In 1986, the National Research Council (NRC, 1986) proposed a simple measurement error model to describe the relation between the observed daily intake for person i on day j, Y_{ij} and the unobservable usual intake for that person, y_i. In their formulation,

$$Y_{ij} = y_i + u_{ij},$$

where $y_i \sim N(\mu, \sigma_y^2)$ and $u_{ij} \sim N(0, \sigma_u^2)$. The measurement error u_{ij} is assumed to be independent of the unobservable usual intakes y_i and also of each other, within a person. Under the model, $y_i = E(Y_{ij}|i) = E(\bar{Y}_i|i)$, where $\bar{Y}_i$ is the observed mean intake of the ith person calculated over r_i daily intake observations. Further,

$Var(\bar{Y}_i) = \sigma_y^2 + \sigma_u^2/r_i$.

The NRC suggested estimating y_i using a best linear unbiased predictor (BLUP) and then estimating the usual intake distribution as the distribution of those BLUPS. Since observed daily intakes Y_{ij} are typically non-normal, the NRC proposed that the model be fit after log-transforming the daily intakes. Nusser et al. [18] revisited this problem and recognized that estimating $f(y)$ is a deconvolution problem. They proposed an approximation to the deconvolution estimate of $f(y)$ that assumed that a univariate transformation of Y_{ij} into the normal scale implies that both y_i and u_{ij} are also normally distributed. In the normal scale, Nusser et al. (1996) fitted the simple measurement error model, estimated the unobservable, normal-scale usual intakes $\ddot{y}_i$ and then, using a suitable back-transformation, obtained the estimated distribution of the y in the original scale.

While the model described above is simple, the areas in which the model can be applied are multiple and include astronomy, biology, chemistry, economy and public health [3], [17]. Estimation of the density of a univariate non-normal random variable with measurement error has been extensively studied. Mendelsohn and Rice (1982) [**?**] presented an example of estimation of a density given observations contaminated with normal error. Stefanski(1990) [23] considered estimation of a continuous bounded probability density when observations from the density are contaminated by additive measurement errors having a known distribution. These studies have focused on the univariate case. An exception is a recent paper by Zhang et al. [21] in which the authors propose a method for estimating a highly multivariate distribution when only short-term measurements are available. Overall, however, there is little work published for the case where the density of interest is multivariate.

We consider the problem of estimating the joint distribution of two correlated random variables where one of the variables is observed with error. An example in nutrition is estimation of the joint distribution of usual energy intake and usual micronutrient intake. While precise biomarkers for energy consumption are available (e.g., doubly-labeled water, Trabulsi and Schoeller, 2001), there are no reliable biomarkers of consumption for nutrients including vitamins and minerals (vitamin K is an exception). Yet, nutritionists are interested in estimating the distribution of usual intake of micronutrients per unit of caloric intake. This is referred to as the nutrient density of the diet and involves estimation of the distribution of the ratio of two non-normal random variables, one of which is observed with measurement error.

The main objective of this paper is to explore whether the method of copulas can be used to estimate the densities of two non-normal random variables when one is contaminated by normal measurement error. In our

set-up, we do not observe the marginal distributions of the two variables, but have access to independent replicate observations, at least of the contaminated variable. While the unobservable bivariate distribution is of interest, we focus on estimation of the density of functions of the two random variables, and in particular, of the ratio of the two random variables. In summary, we develop an approach that combines a deconvolution kernel method (DKM) and the method of copulas to estimate the joint distribution of two non-normal variables where one is contaminated by normal measurement error. DKM is first used to adjust the univariate measurement error. A Gaussian copula is then used to model the correlation structure between the two variables after error adjustment.

This paper is organized as follows. In the next section we describe the model and introduce some notation. We also discuss the methods we propose in this same section. A simulation study is presented in Section 3. We investigate the performance of the algorithm we propose in this section, with emphasis on the accuracy with which we can estimate the density of the ratio of the two random variables. Section 5 includes a discussion of our findings, and gives some directions for future work.

2 Bivariate random measurements with error in one margin

Suppose that we obtain two measurements on the i^{th} sample person on the j^{th} measurement occasion. Let X_{1ij} and X_{2ij} denote the observed values for the i^{th} subject on the j^{th} occasion, where $i = 1, ..., n;\ j = 1, ..., r_i$. For simplicity, we assume $r_i = r$ for all i. Suppose that X_{1ij} is an almost noise-free measurement of the usual value x_{1i} but that X_{2ij} measures x_{2i} with non-negligible error. A simple model in this case is

$$\begin{bmatrix} X_{1ij} \\ X_{2ij} \end{bmatrix} = \begin{bmatrix} x_{1i} \\ x_{2i} + \epsilon_{2ij} \end{bmatrix},$$

$$\begin{bmatrix} x_{1i} \\ x_{2i} \end{bmatrix} \sim \left[\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right] \quad \text{and} \quad \epsilon_{2ij} \sim N(0, \sigma_\epsilon^2).$$

In this model, (x_{1i}, x_{2i}) are assumed to be independent of ϵ_{2ij}. For a given person i, we also assume that the measurement errors $\epsilon_{2ij}, \epsilon_{2ij'}$ are independent. We wish to estimate $f(x_1, x_2)$ when we have at least one observation for x_1 and more than one independent replicate of X_{2i}. We make no distributional assumptions about X_2 but will

assume that the measurement error ϵ is normally distributed.

Let f_W, f_X and f_ϵ denote the densities of X_{2ij}, x_{2i} and ϵ_{2ij}. We propose a method for estimating $f(x_1, x_2)$ that consists of the following steps:

1. We first use the independent replicates $X_{2i1}, ..., X_{2ir}$ to obtain a moment estimator $\hat{\sigma}_\epsilon$ for the measurement error variance σ_ϵ. Then we have that $\hat{f}_\epsilon = \phi(0, \hat{\sigma_\epsilon})$ where $\phi(\mu, \sigma)$ is the normal density.

2. We then adjust for the measurement error in X_2 using a kernel deconvolution method to estimate the density of x_{2i}, denoted $\hat{f}_X$.

3. We use a copula approach to estimate the conditional density $\hat{f}_{X_{1ij}|X_{2ij}}$.

4. Finally, we draw pairs (x_{1i}, x_{2i}) from their estimated joint density as follows:

   (a) Simulate $\epsilon^*_{2ij} \sim \hat{f}_\epsilon$ and $x^*_{2i} \sim \hat{f}_X$ and compute $X^*_{2ij} = x^*_{2i} + \epsilon^*_{2ij}$, i.e. simulate observations contaminated with error.

   (b) Draw X^*_{1ij} from $\hat{f}_{X_{1ij}|X_{2ij}}$ with $X_{2ij} = X^*_{2ij}$

   (c) Calculate $x^*_{1i} = \frac{1}{r}\sum_{j=1}^{r} X^*_{1ij}$

   (d) Repeat a large number of times M to get pairs $(x^*_{1m}, x^*_{2m})_{m=1,...,M}$.

In the next sections, we describe these steps in more detail.

2.1 Deconvolution estimator of $f_{X_2}(x_2)$

Let φ_W, φ_X and φ_ϵ denote the characteristic functions of X_{2ij}, x_{2i} and ϵ_{2ij}. Let f_W, f_X and f_ϵ be probability density functions of X_{2ij}, x_{2i} and ϵ_{2ij}, respectively. By the inversion formula,

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\varphi_X(t)dt = \frac{1}{2\pi}\int e^{-itx}\frac{\varphi_W(t)}{\hat{\varphi}_\epsilon(t)}dt, \tag{1}$$

where we have omitted the subscript 2 to simplify notation. A kernel estimator of $\varphi_W(t)$ is given by

$$\hat{\varphi}_W(t) = \int e^{itw}\hat{f}_W(w)dw \tag{2}$$

where $\hat{f}_W(w) = \frac{1}{nh}\sum_{j=1}^{n} K(\frac{w-W_j}{h})$ is the conventional kernel density estimator of f_W and $K(\cdot)$ is a symmetric probability kernel with finite variance. The resulting estimator of f_X based on $\hat{\varphi}_W(t)$ is the deconvolution kernel density estimator [23]

$$\hat{f}_X(x) = \frac{1}{nh}\sum_{i=1}^{n} L(\frac{x-W_i}{h}), \tag{3}$$

where

$$L(z) = \frac{1}{2\pi}\int e^{-itz}\frac{\varphi_K(t)}{\hat{\varphi}_\epsilon(\frac{t}{h})}dt$$

is called the deconvoluting kernel and is such that φ_K is compactly supported and is the characteristic function of the kernel $K(\cdot)$. The parameter h is the bandwidth parameter. The distribution estimator $\hat{F}_X$ of F_X is thus defined as the integral of $\hat{f}_X$ over $(-\infty, x]$:

$$\hat{F}_X(x) = \frac{1}{2} + \frac{1}{2\pi n}\sum_{j=1}^{n}\int \frac{sin(t(x-W_j))\varphi_K(ht)}{t\hat{\varphi}_\epsilon(t)}dt. \tag{4}$$

We chose a standard kernel function for normal errors, a second-order kernel whose characteristic function has a compact and symmetric support (Fan, 1992) given by

$$K(x) = \frac{48cos(x)}{\pi x^4}(1-\frac{15}{x^2}) - \frac{144sin(x)}{\pi x^5}(2-\frac{5}{x^2}). \tag{5}$$

The characteristic function of the second-order kernel is given by:

$$\varphi_K(t) = (1-t^2)^3 I_{[-1,1]}(t). \tag{6}$$

The resulting deconvolution kernel when we assume normal errors is therefore:

$$L_1(x) = \frac{1}{\pi}\int_0^1 cos(tx)(1-t^2)^3 e^{\frac{\sigma^2 t^2}{2h^2}}dt. \tag{7}$$

The unknown bandwidth parameter h is difficult to determine from the data. There have been at least three different major approaches proposed to estimate the bandwidth parameter. The cross-validation approach proposed by Habbema, Hermans and Van Der Broek [12] while simple to formulate, has been shown to produce highly variable

results [1]. An alternative is what is known as 'plug-in' methods, of which there is a wide variety discussed in the literature [15, 1]. The approach discussed in Delaigle and Gijbels [1] is based on an asymptotic approximation to the mean integrated squared error (MISE), which we describe below. A third approach to estimating the bandwidth, is also based on the MISE, but instead of relying on an asymptotic approximation of the MISE, it relies on a bootstrap approximation to the MISE [2]. Here, we select the bandwidth h by minimizing the asymptotic approximation to the mean integrated error, as in the 'plug-in' method. The (MISE) is defined by

$$MISE(h) = E \int (\hat{f}_X(x,h) - \hat{f}_X(x))^2 dx. \tag{8}$$

Stefanski and Carrol [23] showed that an estimate of the MISE is given by:

$$\hat{MISE}(h) = \frac{1}{2\pi nh} \int \frac{|\varphi_K(t)|^2}{|\varphi_\epsilon(\frac{t}{h})|^2} dt + \frac{h^4}{4} R(f_X^{''}) \int x^2 K(x) dx, \tag{9}$$

where $R(f_X^{''}) = \int [f_X^{''}(x)]^2 dx$. If we were to assume that x_{2i} is normal, $R(\hat{f}_X^{''}) = 0.375 \hat{\sigma}_X^{-5} \pi^{-\frac{1}{2}}$ where $\hat{\sigma}_X = \sqrt{\hat{\sigma}_W^2 - \hat{\sigma}_\epsilon^2}$, $\hat{\sigma}_W^2$ is the sample variance of X_{2ij} and $\hat{\sigma}_\epsilon^2 = (\sum_{i=1}^n \sum_{j=1}^r (X_{2ij} - \bar{x}_{2i.})^2)(n(r-1))^{-1}$. The plug-in selection of h is the value of the bandwidth that minimizes $\hat{MISE}(h)$.

2.2 A copula approach to conditional density estimation

Once we have estimated the marginal densities of x_1 and x_2, we can use the method of copulas to approximate their joint distribution. The history of the copula traces back to Frechet [5]. Formally, a copula is a bi-(or multi) variable distribution function whose marginal distribution functions are uniform on the interval [0,1]. Suppose that we have a g-dimensional random vector $(Z_1, Z_2, ..., Z_g)$ with continuous marginal cumulative distribution functions $F_i(z) = P[Z_i \leq z]$. If we apply the probability integral transform to each marginal, the vector

$$\begin{pmatrix} U_1 & U_2 & ... & U_g \end{pmatrix} = \begin{pmatrix} F(z_1) & F(z_2) & ... & F(z_g) \end{pmatrix}$$

has marginal distributions that are uniform. The copula of the vector Z is then defined as the joint cumulative distribution function of the vector U. More formally,

Definition 2.1. A g-dimensional copula $C : [0,1]^g \longrightarrow [0,1]$ is a cumulative distribution function with uniform

marginals.

Sklar [22] proved the following fundamental result:

Theorem 2.2. ***(Sklar1959)*** *Consider a g-dimensional cdf H with marginals $F_1, ..., F_g$. There exists a copula C, such that*

$$H(x_1, ..., x_g) = C(F_1(x_1), ..., F_g(x_g)) \tag{10}$$

for all $x_i \in \bar{R}$. If F_i is continuous for all $i = 1, ..., g$ then C is unique; otherwise C is uniquely determined only on $RanF_1 \times \cdots \times RanF_g$, where $RanF_i$ denotes the range of the cdf F_i .

This theorem gives a representation of a multivariate c.d.f as a function of each univariate c.d.f. In other words, the copula function captures the dependence structure among the components irrespective of the marginal distributions.

We estimate the conditional density $f_{x_{1i}|X_{2ij}}$ using a copula. By Theorem 2.2, we have that

$$H(X_{1ij}, X_{2ij}) \quad = \quad C(F_1(X_{1ij}), F_2(X_{2ij})), \tag{11}$$

where F_1 and F_2 are marginal cumulative density functions of $X_{1i}$and X_{2ij} and H is joint cumulative density function of X_{1i} and X_{2ij}. Then the joint probability density function is:

$$h(X_{1ij}, X_{2ij}) = \frac{\partial^2 H(X_{1ij}, X_{2ij})}{\partial X_{1ij} \partial X_{2ij}} = \frac{\partial^2 C(F_1(X_{1ij}), F_2(X_{2ij}))}{\partial X_{1ij} \partial X_{2ij}} = f_1(X_{1ij}) f_2(X_{2ij}) c(F_1(X_{1ij}), F_2(X_{2ij})), \tag{12}$$

and the conditional distribution of x_1 given x_2 is given by

$$f_{x_{1i}|X_{2ij}} = \frac{h(X_{1ij}, X_{2ij})}{f_2(X_{2ij})} = f_1(X_{1ij}) c(F_1(X_{1ij}), F_2(X_{2ij})). \tag{13}$$

We used a Gaussian copula to model the correlation structure between X_{1ij} and X_{2ij}, so that

$$C_\rho^{Ga}(F_1(X_{1ij}), F_2(X_{2ij})) = \int_{-\infty}^{\Phi^{-1}(F_1(X_{1ij}))} \int_{-\infty}^{\Phi^{-1}(F_2(X_{2ij}))} \frac{1}{2\pi\sqrt{1-\rho^2}} exp\{-\frac{s^2 - 2\rho st + t^2}{2(1-\rho^2)}\} ds dt. \tag{14}$$

The following corrected rank-based estimate was used to estimate the marginal cumulative distribution functions of X_{1ij} and X_{2ij} [24]:

$$\hat{F}(x^{(k)}) = \frac{r(k) - 0.326}{n + 0.348}, \tag{15}$$

where $r(k)$ is the rank of the k^{th} observation in a vector of observations $\mathbf{x}$. A pseudo(partial)-likelihood for ρ is (Genest et al., 1995):

$$\tilde{ln}(\rho) = \sum_{i=1,\ldots,n;j=1,\ldots,r} lnC_{\rho}^{Ga}(\hat{F}_1(X_{1ij}), \hat{F}_2(X_{2ij})). \tag{16}$$

To estimate ρ we find the value that maximizes the equation (16):

$$\hat{\rho} = \underset{\rho}{argsup}\{\tilde{ln}(\rho)\}. \tag{17}$$

3 Simulation study

We carried out a simulation study to assess the performance of the method we propose to estimate the bivariate density of x_{1i}, x_{2i}. We first generate x_2 from a non-normal distribution as described below. To ensure that the simulated observations are positive, we assume that the additive measurement error model holds after a log transformation of the observations. We generated identically distributed x_{2i} according to (18).

$$x_{2i} \sim 2 * \Gamma(5, 2) + \chi(12). \tag{18}$$

We considered three different structures for the correlation between x_{1i}, x_{2i}. Under the first correlation structure, x_{1i} and x_{2i} are highly correlated. Under the third structure, x_{1i} and x_{2i} are almost uncorrelated. In the third case, knowing the value of X_{2ij} does not provide much information about the value of x_{1i}. More precisely, the three conditional distributions from which we draw the value of x_{1i} are

1. $x_{1i}|x_{2i} \sim \chi^2(x_{2i})$
2. $x_{1i}|x_{2i} \sim \Gamma(5, 1) + \sqrt{x_{2i}}$

3. $x_{1i}|x_{2i} \sim e^{\Gamma(3,5)} + sin(x_{2i})$.

A graphic illustration of the joint distributions of samples from the three schemes is shown in Figure 1 for a single realization.

The simulated observations X_{2ij} are then contaminated by either normal or t errors. Recall that the deconvolution kernel estimator is based on a normal error assumption, so we wished to explore whether the approach we propose is to robust to departures from the normality assumption for the measurement errors. The errors in the study are generated as:

1. $\epsilon_{2ij} \sim N(0, 0.5)$

2. $\epsilon_{2ij} \sim t_3$.

The contaminated observations are then calculated as in (19).

$$X_{2ij} = x_{2i} * e^{\epsilon_{2ij}}. \tag{19}$$

Finally, we varied the number of individuals and the number of independent replicates available for each individual. We considered the case where we had $n = 200$ individuals, each with $r = 7$ independent replicate observations and the case where we had $n = 350$ individuals, each with $r = 4$ replicate observations. Overall, we considered 12 scenarios and the entire simulation study was repeated 15 times. Except where noted, all results presented below are averaged over the 15 simulation replicates.

We proceeded as described in Section 2.1. To illustrate the performance of the deconvolution kernel estimator, we show the estimated density curves corresponding to different sample sizes and two error distributions in Figure 2 and Figure 3. In each case, the average (over 5 simulation replicates) target curve is represented by a solid black line. Figure 2 and Figure 3 compare, for various sample sizes, the results obtained for estimating densities with respect to the two error distributions. The average deconvolution estimators appear to be more skewed to the right relative to the real values. In our study, violating the normal error assumption appears to significantly affect the performance of the kernel deconvolution estimator as suggested by the density estimators shown in Figure 3. However, this is not an issue that we explored in depth and findings are tentative. When the measurement error is normal, we would expect to have better deconvolution estimators when seven (rather than four) independent

replicates are available for each sample person, even though the number of individuals in the sample is half as large. This is because the accuracy with which we can estimate the variance of the measurement error depends more directly on the number of replicates within subject than on the number of subjects. Yet Figure 2 indicates that there is little advantage – at least in these particular simulation scenarios – in increasing the number of replicates per subject from four to seven.

Table 1 contains the mean, variance, and skewness coefficient of the distributions of the true x_{2i}, the x^*_{2i} drawn from the deconvolution estimator of $f(x_2)$ and the distribution of the contaminated sample. Note that, in all cases, the standard deviation of the contaminated sample is larger than that of the sample from the deconvolution estimator. This, in turn, tends to be larger than the standard deviation of the true values. This suggests that the deconvolution estimator of $f(x_2)$ has succeeded in at least partially removing the within-subject variability in the measurements. The mean of the contaminated values X_{2ij} tends to be larger than the means of x_{2i}, x^*_{2i}. This is unexpected at first glance, given that errors are drawn from distributions with zero mean. The reason for the difference in means is that contamination is multiplicative rather than additive (see expression 19).

Table 1: Moments of the distributions of the target values x_{2i}, deconvolution estimates x^*_{2i} and contaminated observations X_{2ij} for different sample sizes and error distributions.

	Mean	Standard Deviation	Skewness
$\epsilon_{2ij} \sim N(0, 0.5), n = 200, r = 7$			
x_{2i}	16.95	1.40	0.30
x^*_{2i}	16.19	2.03	0.30
X_{2ij}	19.12	3.06	0.33
$\epsilon_{2ij} \sim N(0, 0.5), n = 350, r = 4$			
x_{2i}	17.05	1.32	0.20
x^*_{2i}	16.30	2.04	0.45
X_{2ij}	19.27	3.17	0.40
$\epsilon_{2ij} \sim t(3), n = 200, r = 7$			
x_{2i}	16.85	1.35	0.39
x^*_{2i}	18.35	4.00	1.22
X_{2ij}	22.80	53.22	11.71
$\epsilon_{2ij} \sim t(3), n = 350, r = 4$			
x_{2i}	16.90	1.36	0.19
x^*_{2i}	16.14	3.44	0.21
X_{2ij}	21.94	20.41	8.80

Because the deconvolution estimator of $f(x_2)$ appears to deteriorate significantly when the errors are drawn from a heavy-tailed distribution such as the t_3 distribution, we did not consider these cases further in the simulation

study. In the remainder, we present results for the bivariate case, but only when the measurement errors in X_2 are normally distributed. As discussed earlier, the distribution of the ratio of two variables is of interest in some practical applications. For example, estimating the population distribution of the usual intake of a nutrient in energy consumption units requires determination of individual-level ratios, i.e the percent of all calories consumed that are attributable to dietary fat, or the usual dietary density of vitamin C consumption per 1000 calories in the diet. We therefore continued with the simulation study and computed the joint distribution of x_1, x_2 for the case where the measurement error is normal, but the strength of the correlation between the two random variables varies from strong, to moderate to weak and for the two sample size scenarios. We then used our estimated joint distribution to obtain the density of the ratio x_2/x_1 to explore how well the estimated ratio density compares to the true ratio density.

Figures 4, 5 and 6 below show the true ratio density (black curve) and the two estimated densities. The red dashed curves are obtained using a deconvolution estimate of $f(x_2)$ and a Gaussian copula estimate of the joint distribution of x_1, x_2. The blue dotted curves are naive estimates of the ratio density, computed as the empirical distribution of the observed mean ratios. In the three figures, the left panel corresponds to the case where 7 replicates are available for 200 subjects; the right panel corresponds to the case where 4 replicates are available for 350 subjects.

We note from the figures, that the estimator we propose approximates the true ratio density quite well when the correlation between the two variables is high. The performance of the method, however, deteriorates as the correlation decreases. Tables 2 and 3 display estimated percentiles of the distribution of the ratio under different simulation scenarios. The mean percentiles and estimated standard deviations were computed over the 5 replicated simulation samples. Overall, our approach performs better than the naive approach, at least when the two random variables are highly or moderately correlated. When the correlation between x_1, x_2 is high, the performance of our approach improves as we approach the upper tail of the ratio distribution; in this case, only the lower tail percentiles of the estimated ratio distribution are significantly different from the true ratio percentiles. Even when the correlation between x_1, x_2 is only moderate or even low and the estimated percentiles are significantly different from the true percentiles, the naive estimated distribution has percentiles that are even further away from the true values.

Table 2: Percentiles of the ratio $\frac{x_2}{x_1}$ under the three correlation structures. The measurement error distribution is N(0,0.5) and the size is 200 subjects with 7 replicates each. $\hat{r}_k$ is estimated ratio; r_k is the true ratio; r_k^o is the observed ratio with measurement error, and k indicates the corresponding correlation structure.

Quantile	$\hat{r}_1$	r_1	r_1^o	$\hat{r}_2$	r_2	r_2^o	$\hat{r}_3$	r_3	r_3^o
1%	0.3 (0.016)	0.47	0.25	0.55 (0.027)	0.75	0.44	2.35 (0.106)	2.33	1.56
5%	0.44 (0.012)	0.6	0.38	0.81 (0.014)	1.00	0.67	3.57 (0.054)	3.4	2.59
10%	0.53 (0.010)	0.67	0.47	0.96 (0.012)	1.17	0.84	4.28 (0.057)	4.31	3.37
25%	0.7 (0.007)	0.82	0.69	1.28 (0.010)	1.48	1.22	5.88 (0.078)	6.03	5.36
50%	0.96 (0.012)	1.03	1.04	1.78 (0.017)	1.88	1.84	8.32 (0.144)	8.91	9.23
75%	1.31 (0.019)	1.32	1.6	2.5 (0.032)	2.34	2.76	12.02 (0.257)	14.93	17.14
90%	1.72 (0.026)	1.73	2.36	3.28 (0.045)	2.8	3.99	16.2 (0.310)	27.4	32.1
95%	2.05 (0.030)	1.98	2.95	3.89 (0.072)	3.11	4.91	19.5 (0.481)	38.41	47.52
99%	2.88 (0.105)	2.73	4.59	5.59 (0.290)	3.54	7.18	28.17 (1.235)	71.97	101.77

NOTE: Values in parentheses are estimated standard errors for the Monte Carlo mean percentiles.

Table 3: Percentiles of the ratio $\frac{x_2}{x_1}$ under three correlation structures. The measurement error distribution is N(0,0.5) and the size is 350 subjects with 4 replicates each. $\hat{r}_k$ is estimated ratio; r_k is the true ratio; r_k^o is the observed ratio with measurement error.

Quantile	$\hat{r}_1$	r_1	r_1^o	$\hat{r}_2$	r_2	r_2^o	$\hat{r}_3$	r_3	r_3^o
1%	0.24 (0.016)	0.48	0.25	0.48 (0.034)	0.74	0.42	2.15 (0.135)	2.32	1.47
5%	0.42 (0.009)	0.60	0.39	0.83 (0.014)	1.00	0.67	3.61 (0.067)	3.48	2.57
10%	0.51 (0.009)	0.68	0.47	0.98 (0.015)	1.16	0.84	4.39 (0.064)	4.30	3.41
25%	0.68 (0.011)	0.82	0.69	1.3 (0.017)	1.47	1.23	5.99 (0.083)	6.09	5.46
50%	0.96 (0.017)	1.03	1.05	1.81 (0.027)	1.88	1.85	8.56 (0.118)	9.15	9.66
75%	1.34 (0.022)	1.33	1.61	2.48 (0.031)	2.36	2.80	12.28 (0.167)	15.59	17.59
90%	1.80 (0.027)	1.71	2.38	3.30 (0.029)	2.81	3.98	16.84 (0.180)	27.34	32.65
95%	2.16 (0.045)	2.00	3.02	3.87 (0.043)	3.12	4.94	20.32 (0.297)	38.79	47.90
99%	3.03 (0.074)	2.91	4.78	5.16 (0.144)	3.76	7.25	29.07 (0.699)	87.75	102.93

NOTE: Values in parentheses are estimated standard errors for the Monto Carlo mean percentiles.

4 Application

In this section, we illustrate our method using a data set from the Observing Protein and Energy Nutrition(OPEN) that was carried out by NCI. The variables of interest are energy intake measured using doubly labeled water (no measurement error) and protein intake collected using 24-hour recalls on two separate days. We want the distribution of usual protein intake per1000 kcal. We separate men and women before doing the estimation.

Table 4: Percentiles of the ratio $\frac{Protein}{Energy}$ (protein intake per 1000 kcal) with and without adjusting measurement error.

	1%	5%	10%	25%	50%	75%	90%	95%	99%
Adjusted ratio (Female)	10.85	14.66	18.44	23.82	32.96	42.00	54.16	65.38	78.70
Oberved ratio (Female)	10.85	16.92	19.30	25.39	33.04	41.74	50.79	58.11	74.42
Adjusted ratio (Male)	7.61	11.68	15.46	22.42	31.67	44.37	59.44	69.09	112.88
Observed ratio (Male)	8.48	13.71	17.39	22.87	30.92	38.46	47.26	54.96	72.78

5 Discussion

We have proposed an approach to estimate the joint distribution of two non-normal variables when one is contaminated with normal measurement error. The approach consists of two steps. First, we use a deconvolution method to estimate the marginal distribution of the unobservable variable that is observed with error. Next we use a Gaussian copula to estimate the joint distribution of x_1, x_2 using information about the marginals. Copulas are used to model the correlation structure among variables and requires few assumptions about the form of the multivariate distribution to be estimated. Therefore, this approach is applicable more broadly.

Estimation of the marginal distribution of the contaminated random variable is difficult if we wish to minimize assumptions about the form of the unobservable density. Here we have assumed that the errors are normally distributed, but it would be possible, given the independent replicates available for each person, to estimate the distribution of the measurement error empirically. The choice of deconvoluting kernel and of the bandwidth parameter is not straightforward and here we have made choices of convenience. It may be possible to improve the accuracy with which we estimate the marginal distribution of the contaminated random variable. On the other hand, the fact that even choices of convenience greatly improved over the naive estimator of the density suggests that the method we developed might be applicable in a wide range of problems.

The performance of the methods we implement is affected by the degree of association between the two random variables. When the correlation between them is high, the copula approach performs well and the distribution of the ratio of the two variables is closely approximated by the estimated density. When the two random variables are weakly correlated, however, the copula fails, because there is no association to model. In this case, while the estimated ratio density is still a better approximation to the true density that the observed empirical density, the performance of the estimator is poor, particular in the tails of the distribution.

Before settling on the deconvolution copula methodology, we investigated an approach that uses a piecewise

normal linear approximation to estimate the bivariate density. The method was proposed by Dimitris and Efthymia (2010) and an algorithm to implement the method was presented by Kugiumtzis and Bora-Senta (2010). We found that this approach required tuning a large number of model parameters and that it was difficult to account for the contamination in one margin.

References

[1] A. Delaigle and I. Gijbels (2002) Practical bandwidth selection in deconvolution kernel density estimation,Preprint submitted to Elsevier Science

[2] A. Delaigle and I. Gijbels (2004) Bootstrap bandwidth selection in kernel density estimation from a contaminated sample, Ann. Inst. Statist. Math. Vol. 56, No. 1, 19-47

[3] Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). Measurement Error in Nonlinear Models, 2nd ed. Boca Raton, FL: Chapman and Hall CRC Press. MR2243417

[4] Embrechts, P., McNeil, A., and Straumann, D. (2002). Correlation and dependence in risk management: Properties and pitfalls. In Dempster, M., editor, Risk Management: Value at Risk and Beyond., pages 176-223. Cambridge Univ. Press, Cambridge.

[5] M. Frechet. (1951) Sur les tableaux de correlation dont les marges sont donnes, Ann. Univ. Lyon, Science, 4, 13-84

[6] Frechet, M. R. (1958). Remarques au sujet de la note precedente. C. R. Acad. Sci. Paris Ser. I Math. 246, 2719-2720.

[7] G. Dall, Aglio. (1972) Frechet classes and compatibility of distribution functions, Symposia Math., 9, 131-150

[8] Genest, C., Ghoudi, K., and Rivest, L.-P. (1995). A semi-parametric estimation procedure of dependence parameters in multivariate families of distributions. Biometrika 82, 543-552

[9] Genest, C.,(1987). Frank's family of bivariate distributions, Biometrika 74 (3), 549-555.

[10] Genest, C., Rivest, L.-P., 1993. Statistical inference procedures for bivariate Archimedean copulas. Journal of the American Statistical Association, 88 (423), 1034-1043.

[11] Dimitris Kugiumtzis, Efthymia Bora-Senta (2010) Normal correlation coefficient of non-normal variables using piece-wise linear approximation, Comput Stat 25:645-662

[12] Habbema, J.D., Hermans,J. and van den Broek, K. (1974). A stepwise discrimination analysis program using density estimation. Compstat 1974: Proceedings in computational statistics. (G. Bruckman, ed.) 101-110. Vienna: Physica Verlag.

[13] Johnson N, Kotz S (1970) Distributions in statistics, continuous univariate distributions. Houghton Mifflin Company, Boston

[14] Joe, H. (1997): Multivariate Models and Dependence Concepts. Chapman & Hall, London.

[15] M. C. Jones; J. S. Marron; S. J. Sheather (1996) A Brief Survey of Bandwidth Selection for Density Estimation, Journal of the American Statistical Association, Vol. 91, No. 433. , pp. 401-407.

[16] Mendelsohn,J. and Rice, J. (1982) Deconvolution of microfluorometric histograms with B splines. Journal of the American Statistical Association, 77 748-753.

[17] Merritt, D.(1997), Recovering velocity distributions via penalized likelihood. Astronomical J. 114 228-237.

[18] Nusser, S. M., A.L. Carriquiry, W.A.Fuller, and K.W.Dodd. 1996a. A semiparametric approach to estimating usual intake distributions. Jounal of the American Statistical Association.

[19] Oakes,D., (1982). A model for association in bivariate survival data, Journal of the Royal Statistical Society Series B 44, 414-422.

[20] Paul Embrechts, Filip Lindskog and Alexander McNeil,(2001), Modelling Dependence with Copulas and Applications to Risk Management, Working paper, ETH, Zurich, http://www.math.ethz.ch/Finance

[21] Saijuan zhang, Douglas Midthune, Patricia M. Guenther, Susan M. Krebs-Smith, Victor Kipnis, Kevin W. Dodd, Dennis W. Buckman, Janet A. Tooze, Laurence Freedman and Raymond J. Carroll (2011), A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment, The Annals of Applied Statistics, Vol.5,No.2B, 1456-1487

[22] Sklar, A. (1959). Fonctions de repartition a n dimensions e leurs marges. Publications de l'Institut de Statistique de l'Univiversite de Paris 8, 229-231.

[23] Stefanski, L.A. and Carroll.,R.J. (1990) Deconvoluting kernel density estimators. Statistics 2, 169-84

[24] Yu GH, Huang CC (2001) A distribution free plotting position. Stoch Environ Res Risk Assess 15:462-476.

Figure 1: Joint distribution for simulated x_{1i} and x_{2i}; top-left : $x_{1i}|x_{2i} \sim \chi^2(x_{2i})$; top-right: $x_{1i}|x_{2i} \sim \Gamma(5,1)+\sqrt{x_{2i}}$; bottom: $x_{1i}|x_{2i} \sim e^{\Gamma(3,5)} + sin(x_{2i})$

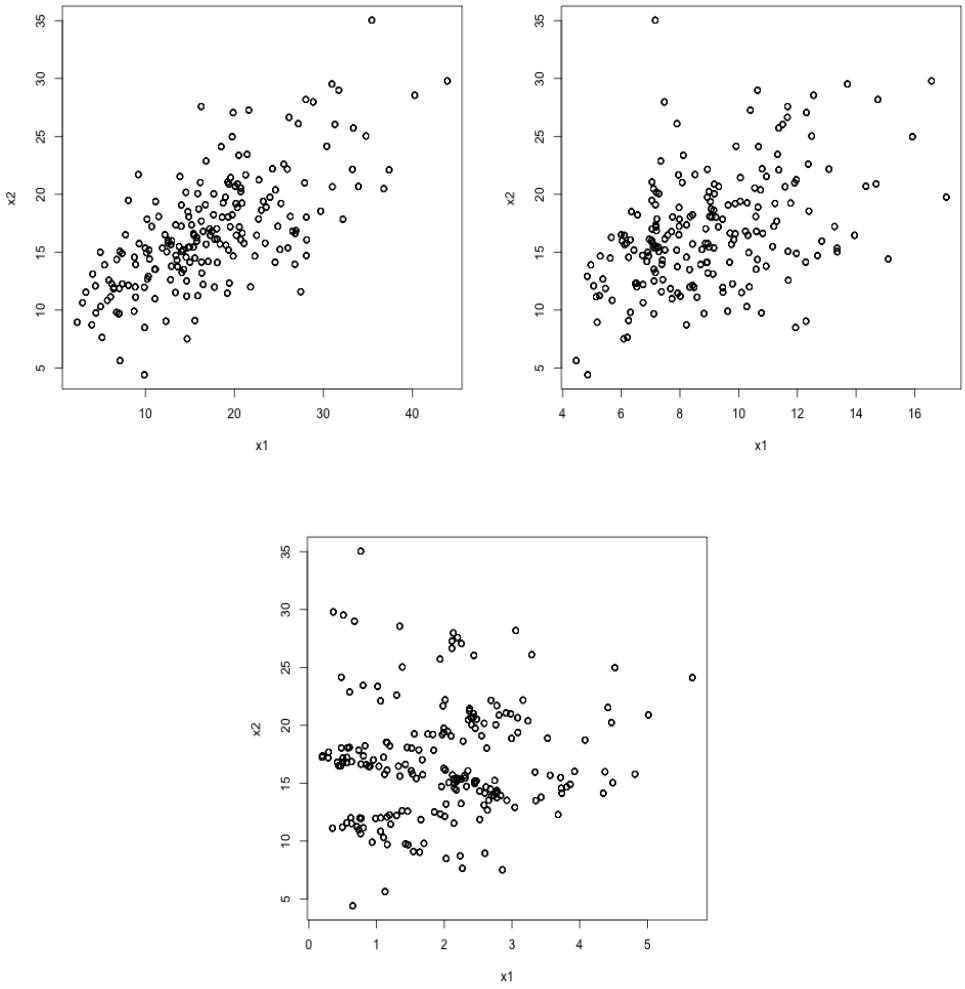

Figure 2: Errors are $\epsilon \sim N(0, 0.5)$. Black solid curve is the average (over 15 reps) of the true density of x_2; blue dotted curve is the average of the naive density estimator, ignoring measurement error; red dashed curve is the average of the deconvolution estimator. The left panel corresponds to the case where $n = 200$ and $r = 7$ and the right panel corresponds to the case where $n = 350$ and $r = 4$.

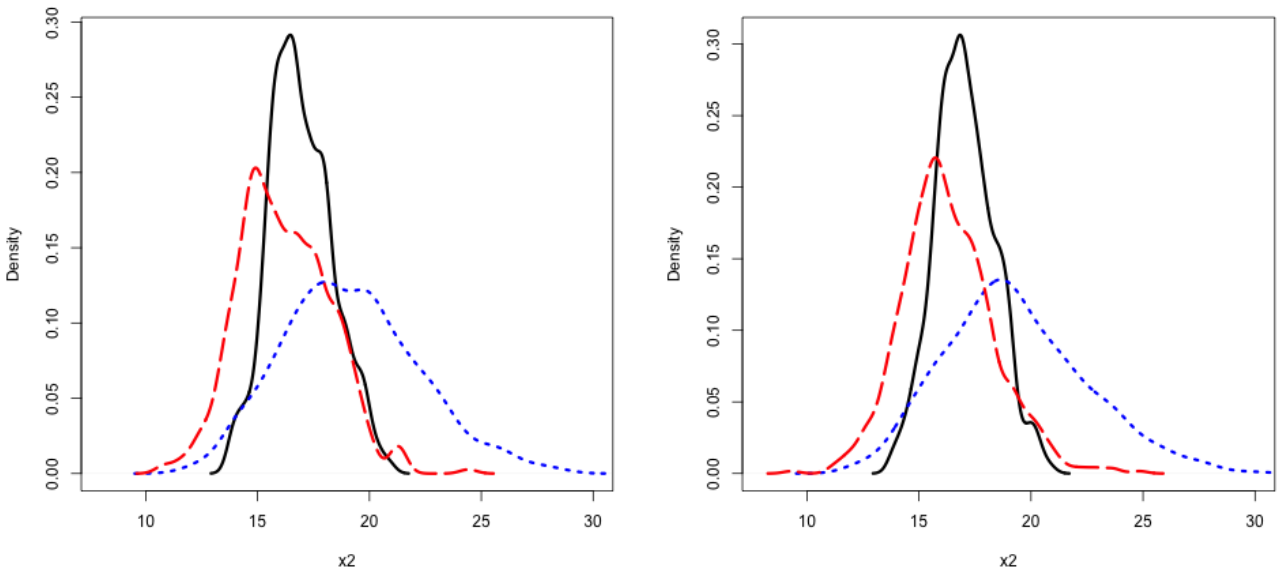

Figure 3: Errors are $\epsilon \sim t(3)$. Black solid curve is the average (over 15 reps) of the true density of x_2; blue dotted curve is the average of the naive density estimator, ignoring measurement error; red dashed curve is the average of the deconvolution estimator. The left panel corresponds to the case where $n = 200$ and $r = 7$ and the right panel corresponds to the case where $n = 350$ and $r = 4$.

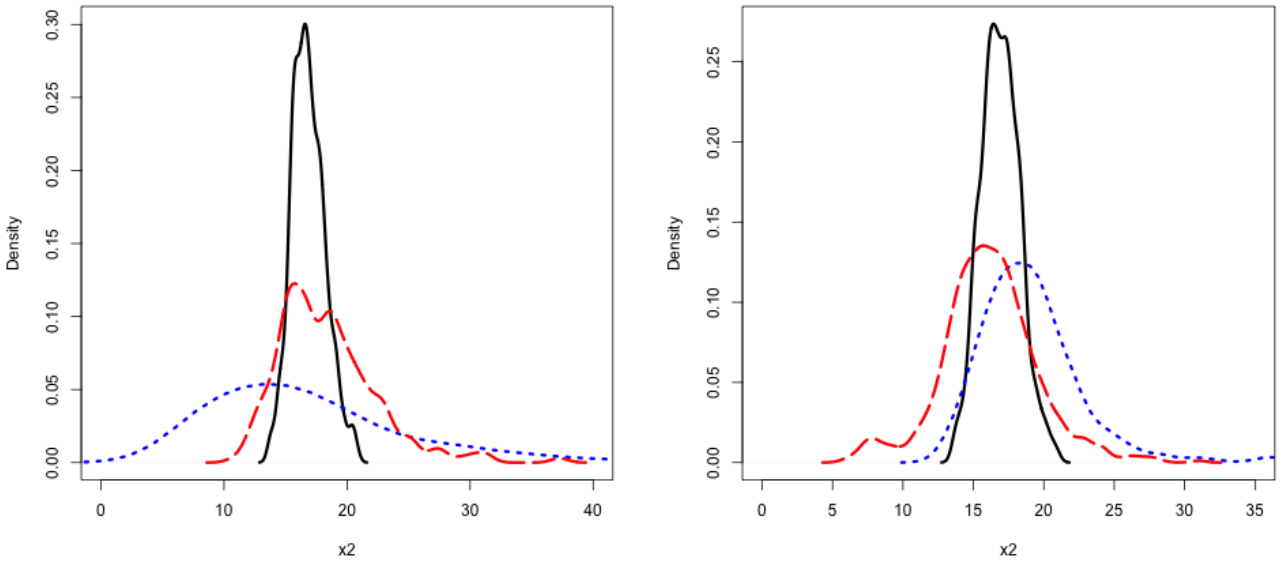

Figure 4: Density of the ratio x_2/x_1 with $x_{1i}|x_{2i} \sim \chi^2(x_{2i})$. Left panel corresponds to $n = 200$ subjects with $r = 7$ independent replicates each; right panel corresponds to $n = 350$ subjects with $r = 4$ replicates. The black solid curve is the true density; the blue dotted curve is the density of the observed ratio (ignoring measurement error); the red dashed curve is obtained using the deconvolution estimate of $f(x_2)$.

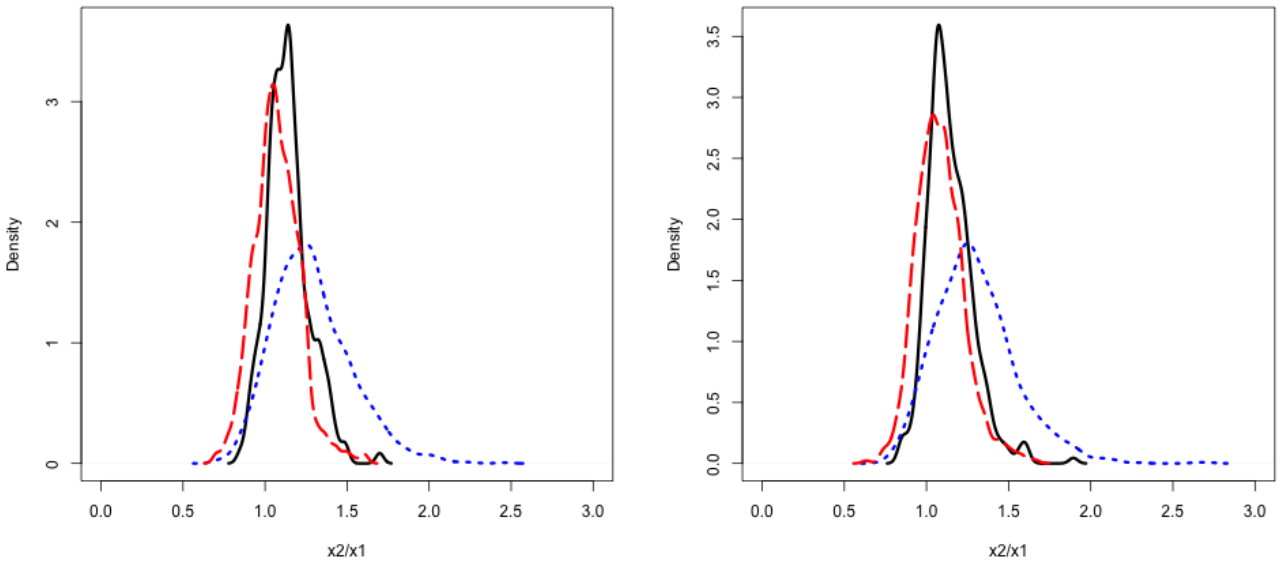

Figure 5: Density of the ratio x_2/x_1 with $x_{1i}|x_{2i} \sim \Gamma(5,1) + \sqrt{x_{2i}}$. Left panel corresponds to $n = 200$ subjects with $r = 7$ independent replicates each; right panel corresponds to $n = 350$ subjects with $r = 4$ replicates. The black solid curve is the true density; the blue dotted curve is the density of the observed ratio (ignoring measurement error); the red dashed curve is obtained using the deconvolution estimate of $f(x_2)$.

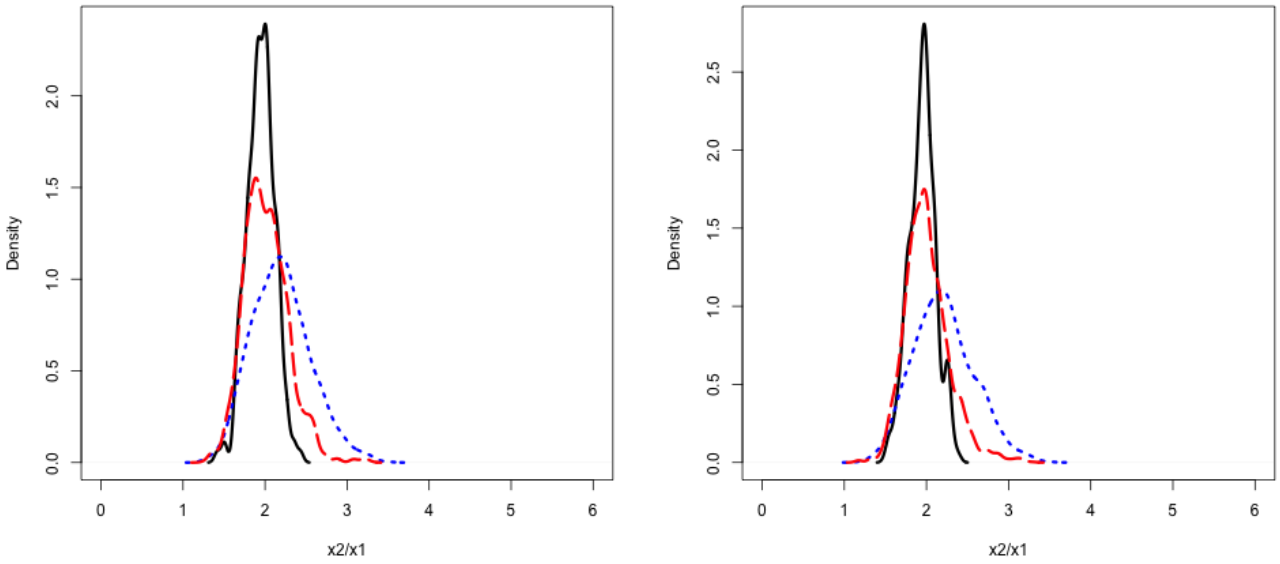

Figure 6: Density of the ratio x_2/x_1 with $x_{1i}|x_{2i} \sim e^{\Gamma(3,5)} + sin(x_{2i})$. . Left panel corresponds to $n = 200$ subjects with $r = 7$ independent replicates each; right panel corresponds to $n = 350$ subjects with $r = 4$ replicates. The black solid curve is the true density; the blue dotted curve is the density of the observed ratio (ignoring measurement error); the red dashed curve is obtained using the deconvolution estimate of $f(x_2)$.

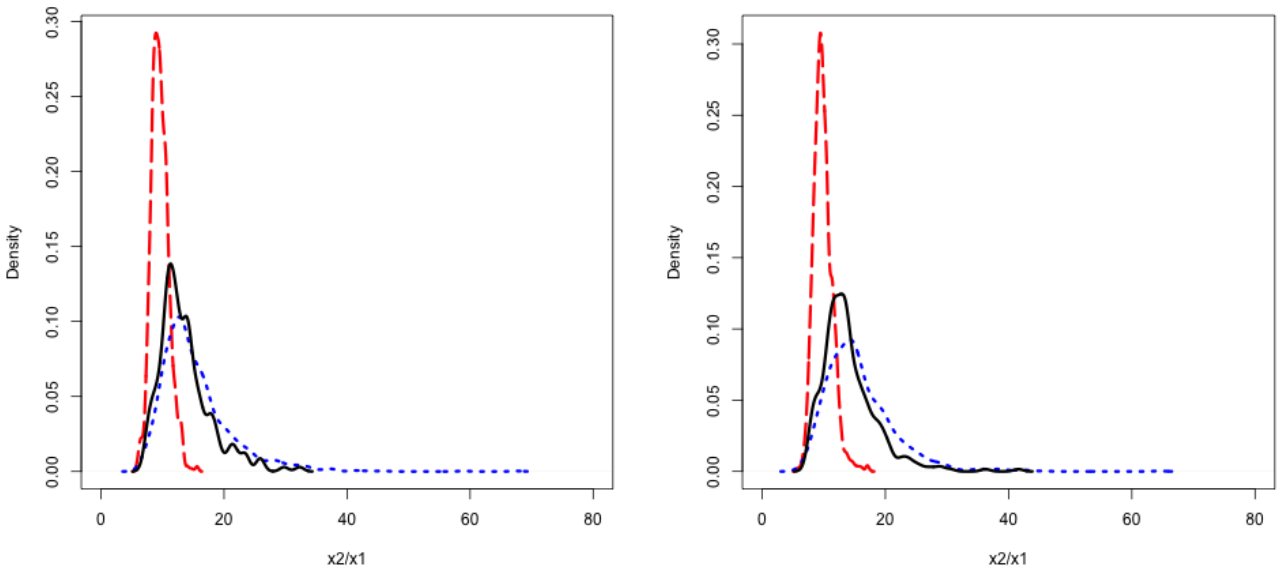

Figure 7: Density of the ratio $\frac{Protein}{Energy}$ (protein intake per 1000 kcal) with (black solid line) and without (red solid line) adjusting measurement error