Density estimation of a biomedical variable subject to measurement error using an auxiliary set of replicate observations

Julien Stirnemann, Fabienne Comte, Adeline Samson

DENSITY ESTIMATION OF A BIOMEDICAL VARIABLE SUBJECT TO MEASUREMENT ERROR USING AN AUXILIARY SET OF REPLICATE OBSERVATIONS

J. J. STIRNEMANN[1,2], F. COMTE[1] AND A. SAMSON[1]

Abstract. Correcting for measurement error the density of a routinely collected biomedical variable is an important issue when describing reference values for both healthy and pathological states. The present work addresses the problem of estimating the density of a biomedical variable observed with measurement error without any *a priori* knowledge on the error density. Assuming the availability of a sample of replicate observations, either internal or external, which is generally easily obtained in clinical settings, an estimator is proposed based on non-parametric deconvolution theory with an adaptive procedure for cut-off selection, the replicates being used for an estimation of the error density. This approach is illustrated in two applicative examples: i) the systolic blood pressure distribution density using the Framingham Study dataset and ii) the distribution of the timing of onset of pregnancy within the female cycle, using ultrasound measurements in the first trimester of pregnancy.

Keywords. Density estimation; non-parametric methods; measurement error; gestational age; replicate measurements.

1. Introduction

The variability of a biomedical variable in a target population depends on several factors. However, its measurement inevitably exposes to measurement error, which adds to the observed variability. This is particularly important in clinical settings or large scale epidemiological studies, when dealing with routinely collected variables, like blood pressure or kaliema, the measurement of which relies on simple devices and/or depends on the circumstances and the operator. Depending on the magnitude of the measurement error, a direct consequence is that the probability distribution function of the observable variable deviates from the probability distribution function of the unobservable true underlying variable. Therefore, measurement error will alter any statistical inference depending on the amount of error the true measurement is contaminated with. In epidemiology and regression for example, measurement error in the exposure variable will bias the result towards the null hypothesis. Numerous examples of methods for dealing with such an error have been suggested in epidemiology (see Budtz-Jørgensen et al. [2003] for an example in environmental epidemiology and Freedman et al. [2008] for an example in a dietary intake cohort study). Similar considerations have been discussed in clinical trials, mostly when the design involves surrogate end-points and biomarkers Sarkar and Qu [2007], Li and Qu

[1] Applied Mathematics, MAP5, UMR CNRS 8145, Université Paris Descartes
[2] Department of Obstetrics and Maternal-Fetal Medicine, GHU Necker - Enfants Malades, Université Paris Descartes.

[2010]. Other common examples include genetics and population association studies, involving both clinical Lobach et al. [2008], Barendse [2011] and technical Bergemann and Zhao [2010] aspects of variables with measurement error.

In the present work, we are interested in the basic problem of estimating the density of a variable when it is measured with an unknown error. There are numerous examples that justify the need for an accurate estimation of the density of a biomedical variable such as the establishment of reference ranges. More particularly, taking potential measurement noise into account should improve the threshold values when planning a screening program based on blood biomarkers. In descriptive epidemiology, removing the measurement error might help in uncovering a bimodal distribution of the variable of interest, thus suggesting a mixture of two different subpopulations. However, since the error is not directly observed, the issue of estimating the density of the error-free unobserved variable is not straightforward. Depending on the setting, several solutions exist for estimating density of the true variable: i) when the density of the error is assumed to be known, classical deconvolution algorithm may be used to achieve a non-parametric estimation using kernel Fan [1991], Liu and Taylor [1989], Stefanski and Carroll [1990], Hesse [1999], Delaigle and Gijbels [2004] or wavelet methods Fan and Koo [2002], Pensky and Vidakovic [1999]. With biomedical variables, this may occur when gross inspection of the data strongly suggests a Gaussian error. However, in many circumstances Gaussian approximation of the error may not hold, as exemplified later. Furthermore, when the parameters of the error distribution are unknown, these methods require a sample estimation of both the mean and variance of the error as an additional estimation step for plugging into a density estimation algorithm. Many of these efficient algorithms are presently available in the R statistical `decon` package R Development Core Team [2010], Wang and Wang [2011]; ii) the density of the error may be unknown but a sample of pure errors without signal is available. In this circumstance, deconvolution algorithms have been implemented and have shown good properties regarding convergence Diggle and Hall [1993], Neumann [1997], Comte and Lacour [2011], Wang and Ye [2012]. However, this situation is less likely to occur in biomedical research especially epidemiological and clinical data, since a sample of signal-free errors requires a specific experimental design, in which signal- free samples have some meaning and are available. However, this situation occurs when dealing with measurement devices such as biological assays or imaging intensity measures. Indeed, assessing the intrinsic characteristics of any new measuring technique is a mandatory step although not always fully informative for the real-life setting.

The present work focuses on a situation where no hypothesis regarding the measurement error distribution can be made and where there is no access to a sample of signal-free measurement errors. However, as we will see in the following Section, errors are considered random, additive and homoscedastic. An estimator of the density of the measurement error-free variable is proposed based upon a second sample comprising replicate observations obtained in the same way as the primary sample of single observations. This estimator has shown good asymptotic properties both theoretically and in simulations Comte et al. [2011]. To illustrate its potential use in biomedical data, we apply this method in two real-data case studies:

i. Systolic blood pressure in the Framingham Study on coronary heart disease. Replicate data is used to estimate the distribution of systolic blood pressure, measured

twice on the same visit. The estimate is compared to both a naive estimator and to an estimation assuming a Gaussian approximation of the error density.

ii. Timing of onset of pregnancy within the female cycle in a cohort of pregnant women. We are interested here in estimating the density of the time interval between last menstrual period and onset of pregnancy. Since the true date of onset of pregnancy is unknown, we may only observe a noisy observation of this time interval. A first trimester ultrasound measurement of crown-rump length of the embryo is used as a noisy observation for dating pregnancy. Replicate observations are obtained from a sample of twin pregnancies using ultrasound measurements on each twin.

In Section 2 we introduce the general frame and notations of deconvolution models followed by our estimator with replicate measurements. Case-studies are presented in Sections 3 and 4. We finish with some discussion in Section 5.

2. Deconvolution with replicate measurements

Only the outline of the model and methods is presented here. Technical details regarding theoretical developments and convergence properties may be found in Comte, Samson and Stirnemann (2011) Comte et al. [2011].

2.1. **Model and notations.**

We will denote Y_j the noisy observation and X_j the unobserved variable for subject j. The goal is to estimate f the density of X. The classical formulation of measurement-error models yields:

$$Y_j = X_j + \varepsilon_j, \qquad j = 1, \ldots, n \tag{1}$$

where ε_j is an error term identically and independently distributed with an unknown density f_ε. Furthermore, we consider the specific case of a homoscedastic error: the sequences $(\varepsilon_j)_{1\leq j\leq n}$ and $(X_j)_{1\leq j\leq n}$ are assumed independent. The density of Y_j denoted f_Y is the convolution of the densities of X_j and ε. Denoting $\star$ the convolution operator, we have:

$$f_Y(x) = (f \star f_\varepsilon)(x) = \int f(x-u) f_\varepsilon(u) du \tag{2}$$

Equivalently, taking the characteristic function of each density denoted by an asterisk, we have:

$$f_Y^*(u) = f^*(u) \times f_\varepsilon^*(u) \tag{3}$$

Since Y is observed, a natural estimator of f_Y^* will be the empirical characteristic function. However, in most biomedical circumstances, the density of the error will be unknown prior to analysis and a sample of pure errors will not be available and sometimes altogether non-realistic. On the contrary, it is often possible to obtain a sample of repeated measurements, which we call replicates throughout this article not to confuse the reader with longitudinal observations. We define replicates as two or more error-contaminated observations of same nature (with identically distributed errors) of a single unobserved true quantity. This is often the case in reproducibility studies for example, when two or more observers will measure the same biological parameter. We suggest that an estimator of the density of an error-contaminated variable can be obtained using replicate observations in this circumstance.

2.2. **Estimators and deconvolution with replicate measurements.** Let us assume we have a sample of size M of $2L$ replicate noisy observations of X_k:

$$Y_{k,2\ell-1} = X_k + \varepsilon_{k,2\ell-1}, \quad Y_{k,2\ell} = X_k + \varepsilon_{k,2\ell}, \quad k = 1, \dots, M, \ell = 1, \dots, L \tag{4}$$

with X_k, $\varepsilon_{k,2\ell-1}$ and $\varepsilon_{k,2\ell}$, for $k = 1, \dots, M$, $\ell = 1, \dots, L$, independent and identically distributed. We assume the sequences $(X_k)_{1\leq k\leq M}$, $(\varepsilon_{k,2\ell-1})_{1\leq k\leq M}$ and $(\varepsilon_{k,2\ell})_{1\leq k\leq M}$ are independent. Therefore, we consider that two independent samples are available: the first, of size M, containing replicate observations and the second, of size n containing non-replicate observations. Density estimation by deconvolution with replicate observations has been previously studied Delaigle et al. [2008], Li and Vuong [1998], Meister and Neumann [2009], showing that an estimator may be achieved using the sample of replicates to estimate the noise density and to improve the deconvolution step. Although related to the estimator given by Delaigle, Hall and Meister Delaigle et al. [2008], we suggest a new estimator that has shown good finite sample and asymptotic properties Comte et al. [2011]. Aside from the assumption of independence, a reasonable assumption is that ε is symmetric and that its characteristic function never vanishes. This assumption implies that the characteristic function is real-valued and strictly positive. Under this hypothesis, observing that $Y_{k,2\ell-1} - Y_{k,2\ell}$ reduces to the difference of two independent, identically distributed random variables $\varepsilon_{k,2\ell-1} - \varepsilon_{k,2\ell}$, we have the following relationship between the corresponding characteristic functions

$$f^*_{Y_{k,2\ell-1}-Y_{k,2\ell}}(u) = (f^*_\varepsilon(u))^2$$

Under the assumption of a symmetric error, f^*_ε is real-valued. Hence an estimator of the square of f^*_ε is obtained by taking the real part of the estimator of $f^*_{Y_{k,2\ell-1}-Y_{k,2\ell}}$:

$$\widehat{(f^*_\varepsilon)^2}(u) = \frac{1}{ML} \sum_{k=1}^{M} \sum_{\ell=1}^{L} \cos\left(u(Y_{k,2\ell-1} - Y_{k,2\ell})\right) \tag{5}$$

We also define an estimator for f^*_Y which is the empirical characteristic function of the independent noisy observations $(Y_j)_{1\leq j\leq n}$ and $(Y_{k,1})_{1\leq k\leq M}$:

$$\hat{f}^*_Y(u) = \frac{1}{n+M} \left(\sum_{j=1}^{n} e^{iuY_j} + \sum_{k=1}^{M} e^{iuY_{k,1}} \right) \tag{6}$$

Because of numerical tractability, (5) cannot be plugged as such in (3) together with (6). In order to prevent the effect of dividing by small numbers in (3), we define a truncated estimator of f^*_ε, as suggested by Neumann (1997) Neumann [1997]:

$$\frac{1}{\tilde{f}^*_\varepsilon(u)} = \frac{\mathbf{1}\{\widehat{(f^*_\varepsilon)^2}(u) \geq (LM)^{-1/2}\}}{\sqrt{\widehat{(f^*_\varepsilon)^2}(u)}} \tag{7}$$

Finally, plugging (7) and (6) in (3) and using Fourrier inversion with a πm cut-off, we have the final estimator of f using a πm cut-off for integrability purpose :

$$\hat{f}_m(x) = \frac{1}{2\pi} \int_{-\pi m}^{\pi m} e^{-ixu} \frac{\hat{f}^*_Y(u)}{\tilde{f}^*_\varepsilon(u)} du \tag{8}$$

This estimator can also be viewed as a deconvolution kernel estimator with bandwidth $1/(\pi m)$.

2.3. **Outline of technical calculation of $\hat{f}_m(x)$.** Technical details regarding calculation of $\hat{f}_m(x)$ can be found in Comte and Lacour Comte and Lacour [2011], Comte, Rozenholc and Taupin Comte et al. [2006] and Comte, Samson and Stirnemann Comte et al. [2011]. Details regarding the selection of an appropriate cut-off πm in (8) can be found in Comte, Samson and Stirnemann Comte et al. [2011]. In brief, the estimator is projected on an orthonormal basis built from sinus cardinal functions defined by $\varphi(x) = \sin(\pi x)/\pi x$. Such an orthonormal basis allows a smooth estimate while offering desirable computational properties. Concurrently, the choice of the cut-off πm is based upon an estimation of an optimal value for m defined by a bias-variance compromise:

$$m^{opt} = \underset{m}{\operatorname{argmin}}\Big(\|f - f_m\|^2 + \operatorname{var}(m)\Big)$$

where $\|f - f_m\|^2$ is the bias between the true unknown density function f and the function that is estimated by (8) while $\operatorname{var}(m)$ is the unknown variance of the estimator, function of m. Each of these two quantities may be estimated and plugged in to estimate an appropriate cut-off for the estimator.

The final algorithm allows a fully automatized estimation of f with cut-off selection and has been implemented in R. Compared to the estimator presented in Delaigle et al. [2008], the presented method provides an adaptative selection procedure for the optimal bandwidth πm and uses a different orthonormal basis for projection. As presented in Comte et al. [2011], comparative simulations show that the presented estimator outperforms the earlier in terms of pointwise risk.

3. Case study: distribution of systolic blood pressure in the Framingham data

To exemplify the method in a simple case study, we consider data from the Framingham Study on coronary heart disease discussed in Carroll *et al.* (2006) Carroll et al. [2006] and in Wang and Wang (2011) Wang and Wang [2011] where systolic blood pressure (SBP) were measured. An accurate denoised estimation of SBP may be of interest in several ways: population reference ranges could be substantially refined; the results of epidemiologic studies or trials using systolic blood pressure as a covariate or as an outcome could be improved; considering the measuring devices are unbiased, the denoised estimate is independant of the measuring technique and may be regarded as a gold-standard since any device-specific noise has been removed.

The data consist of measurements of SBP in 1,615 males on several visits on an 8-year follow-up. During each visit, each patient had its blood pressure measured twice. Since each measurement is contaminated with an unknown error due to either human, biological or technical variations, the density of blood pressure is unknown. Considering only the first visit, our goal is to estimate the density of SBP, using the replicate measurements denoted $SBP1$ and $SBP2$ on each individual. In these data, a plot of $SBP1 - SBP2$ shows that the hypothesis of a Gaussian error would roughly hold in this case as shown in Figure 1 and standard packaged deconvolution algorithms could be used. However, for the sake of the experience, let's not assume a parametric known distribution for the error. To fit our setting, we split the sample into two separate datasets: one containing the first 500 replicate observations of $SBP1$ and $SBP2$ and the second containing only the 1115 last observations of $SBP2$. The hypothesis of homoscedastic error is investigated graphically by plotting

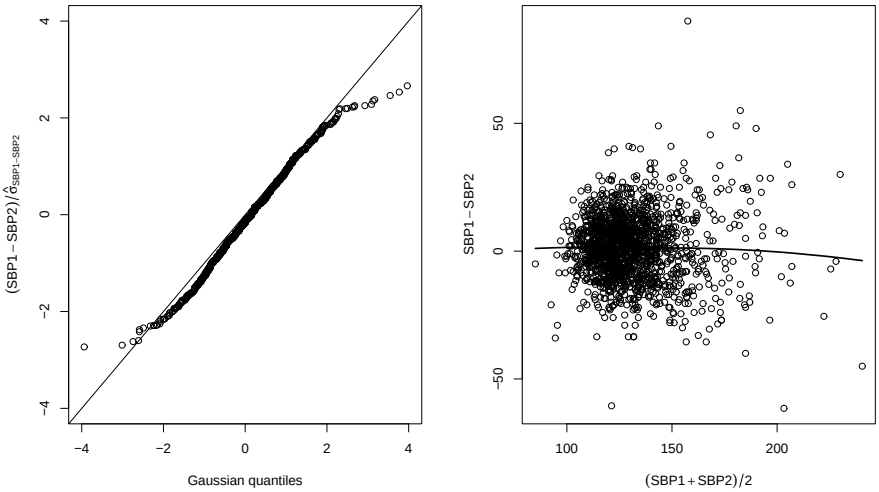

FIGURE 1. Diagnostic plots for the Framingham data. Left-side panel: normal quantile-quantile plot of the standardized difference $SBP1 - SBP2$. Right-side panel: A scatterplot of the difference versus the average of the replicate observations for graphical inspection of potential heteroscedasticity. A nonparametric local scatterplot smoother (LOESS) is added (solid line).

$SBP1 - SBP2$ versus $(SBP1 + SBP2)/2$ in Figure 1 showing no significant relationship. Deconvolving $SBP2$ using the replicate observations yields the estimate presented in Figure 2 (solid line). This estimate is compared to the deconvolution estimator obtained when considering the noise as Gaussian (dotted line) using the `decon` package in R Wang and Wang [2011]. Despite ungraceful ripples in the tails of our estimate, our estimate (solid line) shows a more peaked mode compared to the estimate with Gaussian error. This finding demonstrates that the hypothesis of a Gaussian error may in fact hamper the estimation since it seems to underestimate the amount of noise present in the data. Furthermore, the distributional mode is slightly right-shifted using our estimator. Compared to both deconvolution estimators, the naive kernel estimator (dashed-line) of the raw $SBP2$ data underestimates the peak since the unknown error adds variance to the true distribution.

4. Case study: estimation of density of onset of pregnancy

Except in the specific case of in vitro fertilization, the precise date of pregnancy is unknown in women conceiving naturally. Therefore, it must be estimated using proxy measures that are subject to measurement error (see Dunson *et al.* (1999) Dunson et al. [1999] and Dunson *et al.* (2001) Dunson et al. [2001] for a comprehensive study regarding measurement error in markers of ovulation). Ultrasound biometry of the embryo in early pregnancy has been proven the most accurate method for dating pregnancy in clinical practice Gardosi et al. [1997b,a], Mongelli and Gardosi [1997], Tunón et al. [1996]. Numerous regression formulas have related early ultrasound biometry and gestational age both in

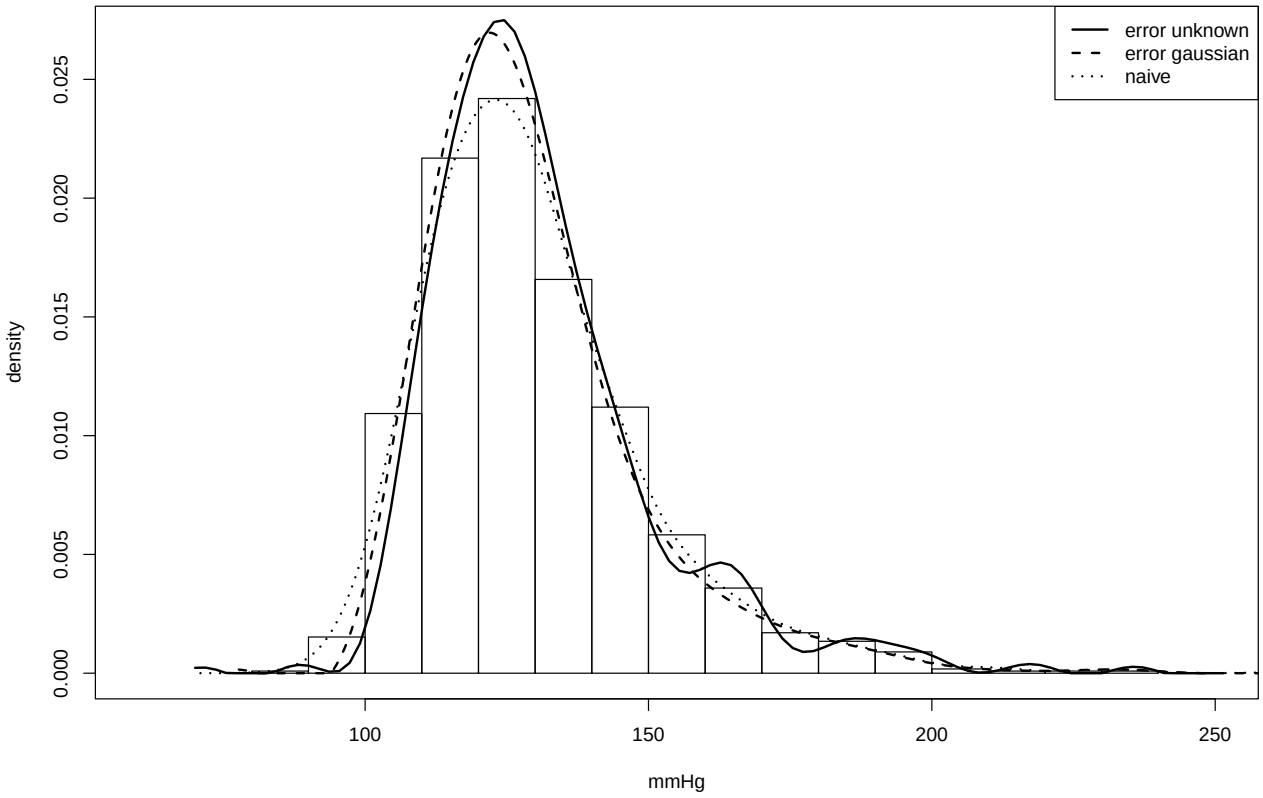

FIGURE 2. Estimation by deconvolution of the density of $SBP2$ in the Framingham Study. The solid line represents the proposed estimator without assumption on measurement error, whereas the dashed line represents the deconvolution estimator with assumption of Gaussian measurement errors. The histogram of the raw observations of $SBP2$ and the naive kernel estimator (dotted line) are given for comparison.

spontaneously conceived pregnancies and pregnancies conceived with assisted reproduction techniques Sladkevicius et al. [2005]. The most widely used formula is that of Robinson Robinson [1973] using crown-rump length in the first trimester of pregnancy. Therefore, for women conceiving spontaneously, early pregnancy ultrasound biometry provides a relevant noisy observation Y_j of X_j.

4.1. **The data.** The first dataset is a sample of singleton pregnancies with one noisy observation Y_j of X_j; the second is a sample of replicate observations obtained from twin pregnancies yielding noisy observations Y_{k1} and Y_{k2} of X_k.

4.1.1. *Single observation sample.* The data comprises basic information recorded at time of first trimester ultrasound and is provided by a single early pregnancy screening center

in Paris, France. Since screening is part of a national policy, women are unselected and are likely to be a representative sample of the general population of pregnant women. First trimester screening visits are planned at around 12 weeks following last menstrual period (about 10 weeks of gestation) and between 11 and 14 weeks as determined by ultrasound. Recorded data includes maternal age, characteristics of the cycles, last menstrual period and ultrasound biometric measurements in the fetus.

Cases were selected if conception was natural (i.e. without assisted reproductive techniques) and if the woman could recall last menstrual period. Because it could influence the recollection of last menstrual period, women were also excluded if ultrasound dating of the pregnancy had been performed prior to the visit.

Over a one-year period, 1,706 cases met these criteria. Menstrual cycles were considered regular in 1386 women whereas 320 women had irregular cycles. For each woman, the observed time interval between the last menstrual period and the date given by the ultrasound measurement of crown-rump length (CRL) was computed, representing the noisy observation Y_j discussed in Section 2.1.

4.1.2. *Replicate observation sample.* Replicate independent observations are obtained from 86 spontaneous twin pregnancies in the same clinical setting as singleton pregnancies. Since onset of pregnancy is the same for both twins, the difference in CRL between the twins may be considered a measurement error. In more technical terms, we have shown in (5), that the empirical characteristic function of the difference of replicate noisy observations is an estimator of the square of the characteristic function of the noise itself. Therefore, each twin pregnancy has two observed time-intervals between LMP and DP based upon ultrasound, thus defining replicates of noisy observations Y_{k1} and Y_{k2} defined in Section 2.2.

It would be arguable here to check if the assumption of normality regarding ε_j would hold, therefore allowing standard deconvolution algorithms with f_ε considered as known. Denote $\hat{\sigma}^2$ the empirical estimator of the variance of $Y_{k1} - Y_{k2}$. Since both twins are exchangeable, we chose to compare the square of the standardized difference $((Y_{k1} - Y_{k2})/\hat{\sigma})^2$ to a $\chi^2(df = 1)$ distribution. Graphically, Figure 3 shows that the hypothesis of a normal error is difficult to sustain, mostly because of departure in the quantile-quantile plot. Therefore, standard deconvolution methods will not apply since we cannot guess any further which potential density the error ε_j originates from. Thus, the only way around is to consider the density of the error unknown as discussed in Section 2.

4.2. **Estimation.** We first consider only women with regular cycles. Deconvolution estimation of f in this population is presented in Figure 4. On the X-axis, $t = 0$ corresponds to the last menstrual period. The mode of the distribution is reached at 13 days meaning that in pregnant women with regular cycles, pregnancy is most likely to occur at that date within the cycle. Noteworthy, this distribution is skewed with a very low probability of pregnancy before 7 days and after 28 days. Therefore, even in women with regular cycles, there is a wide variation in the onset of pregnancy within a female cycle. Compared to the naive estimator not taking into account the error using standard kernel estimation routines (dashed line), our estimate shows a narrower and higher peak because of noise removal.

4.3. **Influence of covariates.** Several covariates have been shown to affect menstrual cycle characteristics including age, ethnicity, smoking, alcohol and other toxic substances

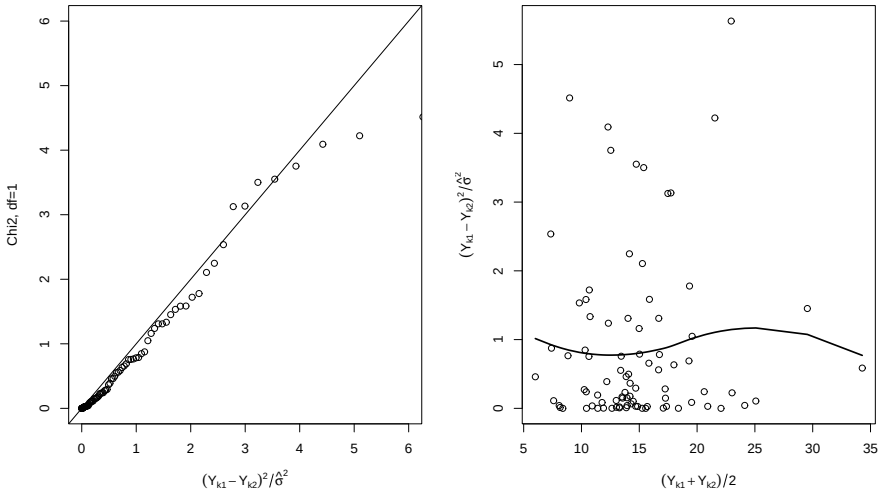

FIGURE 3. Diagnostic plots in the pregnancy data. Left-side panel: quantile-quantile plot comparing $(Y_{k1} - Y_{k2})^2/\hat{\sigma}^2$ and a $\chi^2(df = 1)$ distribution. Right-side panel: investigation of potential heteroscedasticity by plotting $(Y_{k1} - Y_{k2})^2/\hat{\sigma}^2$ versus the average $(Y_{k1} - Y_{k2})/2$. No significant trend was found using a nonparametric LOESS scatterplot smoother (solid line).

Liu et al. [2004]. The influences of age and cycle regularity are studied here as potential covariates for the timing of pregnancy.

4.3.1. *Influence of maternal age.* Maternal age was broken into four classes according to quartiles of increasing age in women with reported regular cycles. In each class of age, the density of onset of pregnancy was estimated independently. The results are presented in the top panel of Figure 5. As expected, a noticeable shortening of the time interval between LMP and date of pregnancy is noticed with increasing age, consistent with the shortening of the first phase of menstrual cycle with age. In each class of age, the modes of the densities were 14.1 days, 13.1 days, 13 days and 13 days for women aged [18.1-29.3],]29.3, 32.4],]32.4,36] and]36,48.1] respectively. The shape of the density also changes across age classes with positive skewness in younger women and negative skewness in older women consistent with the hypothesis of a shortening in the first phase of menstrual cycle with age. Noteworthy, the densities are more peaked in middle aged classes than in extreme classes showing that the probability of date of pregnancy is narrower around the mode in women between 29 and 36 compared to older and younger women.

4.3.2. *Influence of regularity of cycles.* Prior knowledge regarding individual menstrual characteristics may influence the dating of pregnancy since patients with irregular cycles are subject to more variation in timing of ovulation and therefore in timing of pregnancy as shown by Wilcox, Dunson and Baird (2000) Wilcox et al. [2000] in a study of fertility in non-pregnant fertile women. A comparison of timing of pregnancy in patients with reported

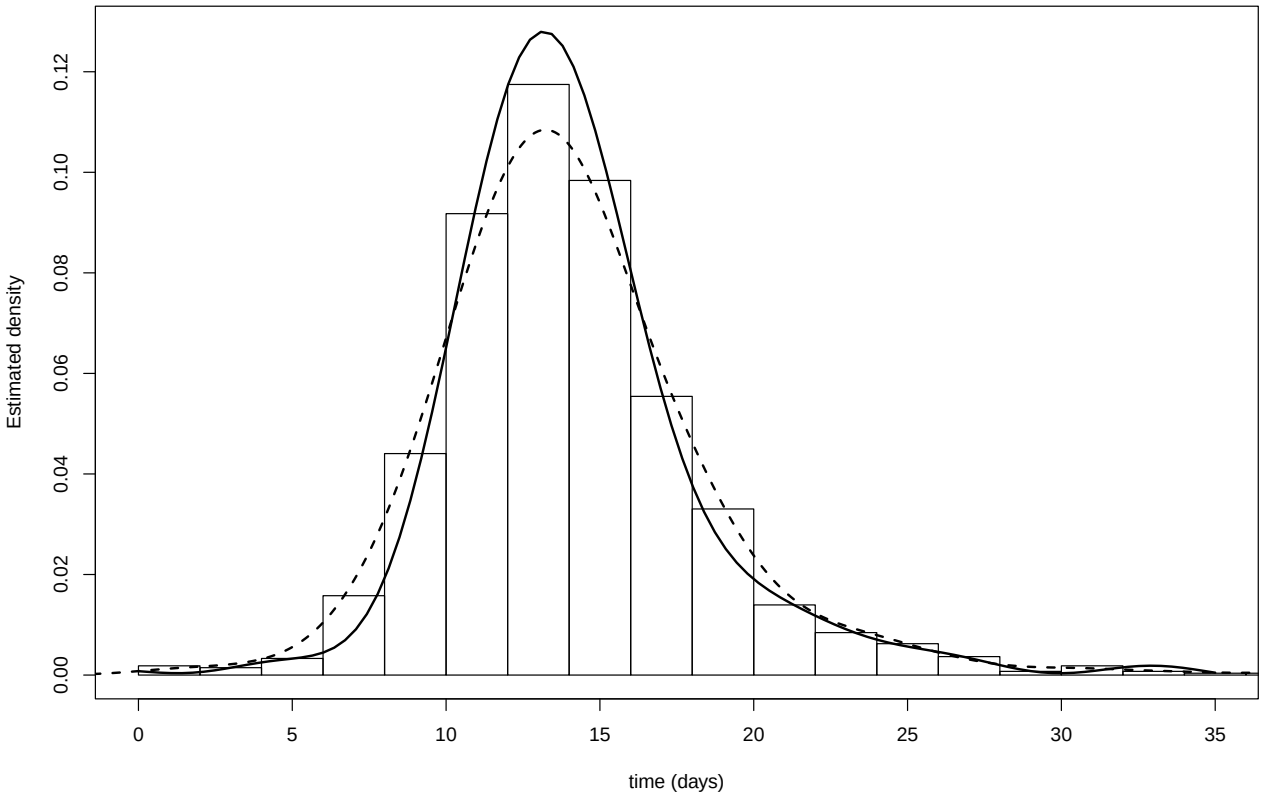

FIGURE 4. Estimated density of onset of pregnancy for women with regular cycles (solid line). The histogram and naive estimator (dashed line) of the raw data Y_j is given for comparison. The X-axis is the time interval between onset of pregnancy and last menstrual period (t=0).

regular and irregular cycles is presented in the bottom panel of Figure 5. Consistent with Wilcox, Dunson and Baird (2000) Wilcox et al. [2000], the density of onset of pregnancy is shifted towards longer time intervals with a wider variation than women with regular cycles.

5. Concluding remarks

We have presented a general method for estimating the density of a variable with error in measurement. This method is conceived for frequent situations occurring in biological and clinical research settings and should allow researchers to improve analyses and provide more meaningful results. We emphasize that only reasonable hypotheses are required. In particular, estimation does not require a prior hypothesis regarding a parametric density of the error. Such parametric distributions for errors usually include Laplace and Gaussian densities which may often be an undesirable approximation. Owing to replicate

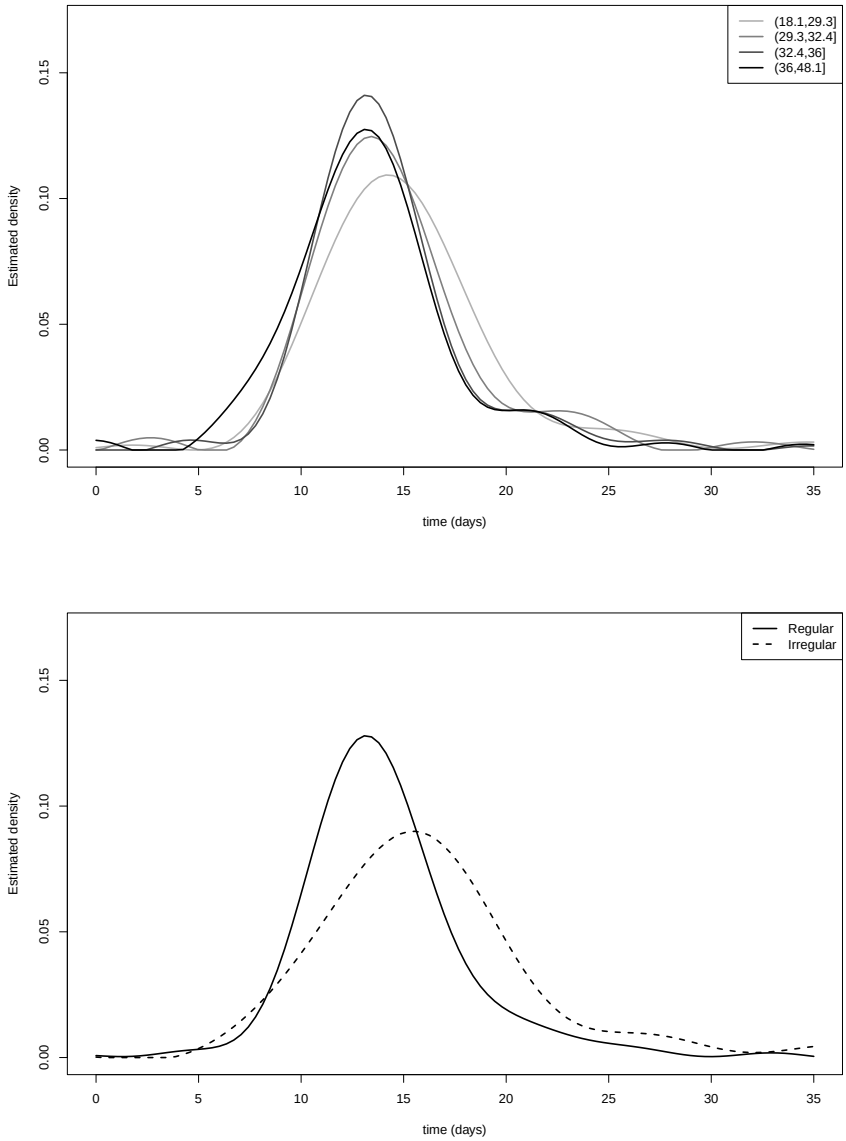

FIGURE 5. Top panel: estimated density of onset of pregnancy according to maternal age in patients with regular menstrual cycles. Maternal age is divided into quartiles, represented by increasing intensities of gray. Bottom panel: estimated density of onset of pregnancy in women with regular cycles (solid line) and women with irregular cycles (dashed line). The X-axis is the time interval between onset of pregnancy and last menstrual period (t=0).

observations, only the hypothesis of Gaussian errors may be easily investigated since the convolution of other distributions is unlikely to have a known parametric solution, except in very specific cases. The only hypothesis regarding distribution of the errors is symmetry, which seems a reasonable assumption in many situations involving biomedical data.

We wish to emphasize that replicate observations are not independent: only the errors ε_{k1} and ε_{k2} must be independent and identically distributed as well as independent of the underlying true variable X_k. Independence between errors in replicate observations is a classical and reasonable assumption. We acknowledge that independence between the error and the true variable, defining homoscedasticity, may be difficult to verify. However, following Bland and Altman (1999) Bland and Altman [1999], in most cases graphical inspection of a scatterplot of $(Y_{k1}-Y_{k2})$ versus $(Y_{k1}+Y_{k2})/2$ should be sufficiently conclusive. In case of evidence of heteroscedasticity, solutions have been recently suggested for replicate measurements under the assumption of a Gaussian error, considering the variances as random McIntyre and Stefanski [2011].

Two case-studies were used to exemplify the use of deconvolution in analyses of biomedical data: in the Framingham data, we showed that using an unknown-error algorithm does not alter the estimation compared to classical algorithm assuming a fully known density for the error. In the pregnancy data, we conducted a more complete data analysis presenting novel results regarding timing of onset of pregnancy. The results of this second case study deserve further discussion regarding interpretation and potential applications. Using twin pregnancies may be arguable as a choice of replicate measurements since they may differ in some way from the population of singleton pregnancies. However, as far as we know, there does not seem to be a significant difference between singleton and twins for ultrasound dating of pregnancy Dias et al. [2010]. Although it is related to previous results regarding the probability of ovulation during the menstrual cycle in healthy women often referred to as a "fertile window" Dunson et al. [1999, 2001], Wilcox et al. [2000], the clinical question is different: we are interested in the timing of pregnancy rather than the timing of ovulation. These two events differ since pregnancy does not necessarily occur just following ovulation and because the potency of an ovulation to lead to a pregnancy may depend upon its timing itself. Therefore, the density of onset of pregnancy and the density of ovulation are likely to be different. Furthermore, dating ovulation can be performed accurately only in a specific prospective setting with intensive monitoring of biological samples throughout successive cycles, whereas we are concerned by the general population of pregnant women without enrollment in a specific experimental study. In a clinical perspective, the estimation of the density of timing of pregnancy is interesting in several ways: i) The variability of onset of pregnancy based upon last menstrual period displayed by the estimated density is much wider than the +/- 5 days generally recognized as the interval of the error using ultrasound in the first trimester. Therefore, even in women with regular cycles and with a known date of last menstrual period, ultrasound will perform better than last menstrual period for dating pregnancy in clinical practice. ii) As discussed earlier, using any proxy for estimating onset of pregnancy in a single pregnant woman, such as crown-rump length ultrasound measurement, is subject to measurement error. However, in a Bayesian perspective and notwithstanding the larger variability in dating resulting from last menstrual period, the distribution of time to pregnancy may be used as a prior distribution for a likelihood function given by a proxy measurement. Thus, it may refine the estimate of a proxy measurement by incorporating last menstrual period as prior knowledge.

Acknowledgements.

We thank Professor J.C. Thalabard for his motivating input as well as for his contribution in correcting the manuscript.

References

W Barendse. The effect of measurement error of phenotypes on genome wide association studies. *BMC Genomics*, 12:232, 2011.

T L Bergemann and L P Zhao. Signal quality measurements for cDNA microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, 7(2):299–308, June 2010.

J M Bland and D G Altman. Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, 8(2):135–160, 1999.

E Budtz-Jørgensen, N Keiding, P Grandjean, P Weihe, and R F White. Consequences of exposure measurement error for confounder identification in environmental epidemiology. *Statistics in Medicine*, 22(19):3089–3100, October 2003.

R J Carroll, D Ruppert, L A Stefanski, and C M Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective, Second Edition*. Chapman and Hall/CRC, 2 edition, June 2006. ISBN 1584886331.

F. Comte and C. Lacour. Data-âĂŘdriven density estimation in the presence of additive noise with unknown distribution. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):601–627, 2011.

F Comte, Y Rozenholc, and M L Taupin. Penalized contrast estimator for adaptive density deconvolution. *Canadian Journal of Statistics*, 34(3):431–452, 2006.

F Comte, A Samson, and J J Stirnemann. Deconvolution estimation of onset of pregnancy with replicate observations, 2011.

A Delaigle and I Gijbels. Bootstrap bandwidth selection in kernel density estimation from a contaminated sample. *Annals of the Institute of Statistical Mathematics*, 56(1):19–47, 2004.

A Delaigle, P Hall, and A Meister. On deconvolution with repeated measurements. *The Annals of Statistics*, 36(2):665–685, 2008.

T Dias, S Mahsud-Dornan, B Thilaganathan, A Papageorghiou, and A Bhide. First-trimester ultrasound dating of twin pregnancy: are singleton charts reliable? *BJOG: An International Journal of Obstetrics and Gynaecology*, 117(8):979–984, 2010.

P J Diggle and P Hall. A fourier approach to nonparametric deconvolution of a density estimate. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(2):523–531, 1993.

D B Dunson, D D Baird, A J Wilcox, and C R Weinberg. Day-specific probabilities of clinical pregnancy based on two studies with imperfect measures of ovulation. *Human Reproduction (Oxford, England)*, 14(7):1835–1839, July 1999.

D B Dunson, C R Weinberg, D D Baird, J S Kesner, and A J Wilcox. Assessing human fertility using several markers of ovulation. *Statistics in Medicine*, 20(6):965–978, 2001.

J Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272, September 1991.

J Fan and J Y Koo. Wavelet deconvolution. *Information Theory, IEEE Transactions on*, 48(3):734–747, 2002.

L S Freedman, D Midthune, R J Carroll, and V Kipnis. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, 27(25):5195–5216, November 2008.

J Gardosi, T Mul, A Francis, J Hall, and S Fishel. Comparison of second trimester biometry in singleton and twin pregnancies conceived with assisted reproductive techniques. *British Journal of Obstetrics and Gynaecology*, 104(6):737–740, 1997a.

J Gardosi, T Vanner, and A Francis. Gestational age and induction of labour for prolonged pregnancy. *British Journal of Obstetrics and Gynaecology*, 104(7):792–797, 1997b.

C H Hesse. Data-driven deconvolution. *Journal of Nonparametric Statistics*, 10(4):343–373, 1999.

T Li and Q Vuong. Nonparametric estimation of the measurement error model using multiple indicators. *Journal of Multivariate Analysis*, 65:139âĂŞ165, 1998.

W Li and Y Qu. Adjustment for the measurement error in evaluating biomarkers. *Statistics in Medicine*, 29(22):2338–2346, September 2010.

M C Liu and R L Taylor. A consistent nonparametric density estimator for the deconvolution problem. *Canadian Journal of Statistics*, 17(4):427–438, 1989.

Yan Liu, Ellen B Gold, Bill L Lasley, and Wesley O Johnson. Factors affecting menstrual cycle characteristics. *American Journal of Epidemiology*, 160(2):131–140, 2004.

I Lobach, R J Carroll, C Spinka, M H Gail, and N Chatterjee. Haplotype-based regression analysis and inference of case-control studies with unphased genotypes and measurement errors in environmental exposures. *Biometrics*, 64(3):673–684, September 2008.

J. McIntyre and L. A Stefanski. Density estimation with replicate heteroscedastic measurements. *Annals of the Institute of Statistical Mathematics*, 63 (1):81–99, 2011.

A. Meister and M. Neumann. Deconvolution from non-standard error densities under replicated measurements. *Statistica Sinica*, 20 (4):1609–1936, 2009.

M Mongelli and J Gardosi. Birth weight, prematurity and accuracy of gestational age. *International Journal of Gynaecology and Obstetrics: The Official Organ of the International Federation of Gynaecology and Obstetrics*, 56(3):251–256, 1997.

M H Neumann. On the effect of estimating the error density in nonparametric deconvolution. *Journal of Nonparametric Statistics*, 7(4):307, 1997.

M Pensky and B Vidakovic. Adaptive wavelet estimator for nonparametric density deconvolution. *Annals of Statistics*, pages 2033–2053, 1999.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

H P Robinson. Sonar measurement of fetal crown-rump length as means of assessing maturity in first trimester of pregnancy. *British Medical Journal*, 4(5883):28–31, 1973.

S Sarkar and Y Qu. Quantifying the treatment effect explained by markers in the presence of measurement error. *Statistics in Medicine*, 26(9):1955–1963, April 2007.

P Sladkevicius, S Saltvedt, H Almström, M Kublickas, C Grunewald, and L Valentin. Ultrasound dating at 12-14 weeks of gestation. a prospective cross-validation of established dating formulae in in-vitro fertilized pregnancies. *Ultrasound in Obstetrics and Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 26(5):504–511, 2005.

L A Stefanski and R J Carroll. Deconvoluting kernel density estimators. *Statistics*, 21(2): 169—184, 1990.

K Tunón, S H Eik-Nes, and P Grøttum. A comparison between ultrasound and a reliable last menstrual period as predictors of the day of delivery in 15,000 examinations. *Ultrasound in Obstetrics & Gynecology: The Official Journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 8(3):178–185, 1996.

Xiao-Feng Wang and Bin Wang. Deconvolution estimation in measurement error models: The R package decon. *Journal of Statistical Software*, 39(10), March 2011.

Xiao-Feng Wang and Deping Ye. The effects of error magnitude and bandwidth selection for deconvolution with unknown error distribution. *Journal of Nonparametric Statistics*, 24(1):153–167, 2012.

A J Wilcox, D Dunson, and D D Baird. The timing of the "fertile window" in the menstrual cycle: day specific estimates from a prospective study. *BMJ (Clinical Research Ed.)*, 321 (7271):1259–1262, 2000.