Semiparametric Estimation of the Distribution of Episodically Consumed Foods Measured with Error

Félix Camirand Lemyre
Département de mathématiques, Université de Sherbrooke, Sherbrooke, Qc, Canada,
Centre de Recherche du Centre Hospitalier Universitaire de Sherbrooke,
felix.camirand.lemyre@usherbrooke.ca

Raymond J. Carroll
Department of Statistics, Texas A&M University, 3143 TAMU, College Station, TX
77843-3143, and School of Mathematical and Physical Sciences, University of Technology
Sydney, Broadway NSW 2007, Australia, carroll@stat.tamu.edu

Aurore Delaigle
School of Mathematics and Statistics and Australian Research Council Centre of
Excellence for Mathematical and Statistical Frontiers, University of Melbourne, Parkville,
VIC, 3010, Australia, aurored@unimelb.edu.au

Abstract: Dietary data collected from 24-hour dietary recalls are observed with significant measurement errors. In the nonparametric curve estimation literature, much of the effort has been devoted to designing methods that are consistent under contamination by noise, and which have been traditionally applied for analysing those data. However, some foods such as alcohol or fruits are consumed only episodically, and may not be consumed during the day when the 24-hour recall is administered. These so-called excess zeros make existing nonparametric estimators break down, and new techniques need to be developed for such data. We develop two new consistent semiparametric estimators of the distribution of such episodically consumed food data, making parametric assumptions only on some less important parts of the model. We establish its theoretical properties and illustrate the good performance of our fully data-driven method in simulated and real data.

Some Key Words: Asymptotic theory; Deconvolution; Excess zeros; Measurement error; Nonparametric deconvolution.

Short title: Estimating the Distribution of Error-Prone Episodically Consumed Foods

1 Introduction

For many years, national nutritional surveys in the U.S. (NHANES: National Health and Nutrition Examination Survey), Australia (e.g., the 2007 Australian National Children's Nutrition and Physical Activity Survey), Canada (the Canadian Community Health Survey) and elsewhere have used the 24-hour dietary recall (24HR). The 24HR aims at collecting precise information on nutrient and food intake over the past 24-hour period as the primary self-report instrument for dietary assessment (Dwyer et al., 2003). The main purpose of collecting dietary data in national surveys has been to estimate the distribution of usual (that is, average long-term) intake of various nutrients and food groups in populations and subpopulations, and to monitor such intakes over time. Another important use of the data has been to relate individual usual intakes to health outcome measures, such as obesity or blood pressure. Because a 24HR captures food/nutrient intakes only for a single day, it is not a good or even reliable estimate of long-term average intake. Thus, using just a 24HR to estimate the distribution of a usual intake without accounting for the inherent measurement error of a 24HR is badly biased for important quantities such as percentiles.

There is a vast literature on measurement errors for estimating usual intake distributions when the food/nutrient is consumed daily. In that standard setting, if X is the unobserved and unobservable usual intake, and W the result of a single 24HR, the typical model for the data $W_1, \ldots, W_n$ is the classical error model, where $W_i = X_i + U_i$ with X_i and U_i independent, and where U_i represents the measurement error. Most of the literature uses this simple model or variants of it, and assumes a fully known distribution for the day-to-day variability, or measurement error, of the 24HR. Often the W_i's do not satisfy the classical additive error model, and it is a transformed version of the data that satisfies it. For example, it is common to assume that the model is satisfied after log transformation. There, it is typical to assume that $W_i' = X_i' + U_i'$, where $W_i' = \log(W_i)$, $X_i' = \log(X_i)$ and $U_i' = \log(U_i)$. More generally, the model is often assumed to hold after the W_i's have been transformed by a monotone

function $\mathfrak{h}$.

In nutrition studies, there is considerable interest in the usual intake of episodically consumed foods, e.g., fish, whole grains, whole fruits, fruit juices, dark green and orange vegetables and legumes (DOL), milk, etc., see for example Guenther et al. (2008a,b) and Guenther et al. (2014). For example, in the 2001–2004 NHANES, the percentages of reported nonconsumption of total fruit, whole fruit, whole grains, DOL and milk on any single day are 17%, 40%, 42%, 50% and 12%, respectively. Such variables are equal to zero on days of nonconsumption and are strictly positive on consumption days.

Often several 24-hour recalls are available for each subject, so that the observations are a sample of replicated recalls W_{ij} for $i = 1, \ldots, n,\ j = 1, ..., J,\ J \geqslant 2$. A major goal of nutritional surveys is to estimate the distribution of the usual intake for an individual, that is, the distribution of $T_i = E(W_{ij}|X_i)$. It is important to recognize that this is the definition of the usual intake in the *original* data scale. The W_{ij}'s can still be regarded as a version of the usual intake contaminated by the day-to-day variability, but because of the excess zeros on nonconsumption days, we cannot use the classical additive error model to describe them.

There is a literature and associated multiple modeling approaches on the problem of excess zeros plus measurement error, and we use a model in line with that literature; see Tooze et al. (2002, 2006), Li et al. (2005), Kipnis et al. (2009), Keogh et al. (2011), Zhang et al. (2011a,b) and Carroll (2014). However, all these references are fully parametrics and assume that, after data transformation, every random variable is exactly normally distributed; see Section 2. In this work we are the first to propose an estimator of the distribution of the usual intake T_i that breaks out of the parametric distribution assumption. As we show, the problem is difficult technically, and requires completely new ways of analysis.

Although the motivation of our work comes from nutrition, the excess-zeros problem arises in many different areas such as ecology (Fletcher et al., 2005), environment (Tian, 2005) and drug testing (Zhou and Tu, 1999). There, the data are often modelled via two

components: one that accounts for the probability of observing nonzero values and one that reflects the distribution of nonzero values. The interest is then often in, but not limited to, estimating each component of the model. The methodology that we propose in this work allows us to derive such estimates as a by-product of our estimation procedure.

This paper is organised as follows. We introduce our model and data in Section 2. Our methodology for estimating the distribution of usual intake in introduced in Section 3, where we give two different estimators for the respective cases where the error density is known and that where it is unknown. We derive asymptotic properties of our estimators in Section 4. In Section 5, after showing how to choose in practice the smoothing parameters needed by our procedures, we illustrate their numerical performance on simulated data and apply the method to estimate the distribution of usual alcohol and fruit intake from an American study on eating. Proofs of the results are very lengthy and are given in the Supplementary Material.

2 Model and data

We are interested in the usual (i.e. long term daily average) intake T of a food. Instead of observing T, we observe data W_{ij}, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, where W_{ij} represents the reported food intake by the i^{th} individual on day j. On some days the i^{th} individual does not consume the food, so that $W_{ij} = 0$. On the days where the food is consumed, the food intake is measured by W_{ij}, which, after appropriate transformation, is a version of the usual intake X_i contaminated by a classical measurement error.

Tooze et al. (2006, 2010) and Kipnis et al. (2009) consider a parametric model for this problem. In their formulation, they suppose that the probability of eating a dietary component can be described via a known strictly increasing cumulative distribution function H, e.g. the logistic distribution function. An equivalent formulation in our context is the following. Specifically, for $i = 1, ..., n$ and $j = 1, ..., J$, they assume that for some latent

variables X_i related to usual intake,

$$\mathbb{P}(W_{ij} > 0 | X_i) = H(\beta_0 + \beta_1 X_i) \equiv H_{\boldsymbol{\beta}}(X_i)\,,$$

where β_0 and β_1 are unknown parameters and $\boldsymbol{\beta}^T = (\beta_0, \beta_1)$. For example, in most cases, the probability of eating a food on a given day is an increasing function of the usual intake of that food: the more we eat the food on average, the more likely we are to eat it on a given day.

On days where a food has been consumed, after appropriate transformation, the measured food intake satisfies a classical error model. Specifically, we assume that there is a known strictly monotone function $\mathfrak{h} : (0, \infty) \to \mathbb{R}$, e.g. a log transformation, such that given the latent variable X_i, the W_{ij} are independent and

$$\mathbb{P}(\widetilde{W}_{ij} \leqslant v | W_{ij} > 0, X_i = x) = \mathbb{P}(X_i + U_{ij} \leqslant v | X_i = x)\,,$$

where $\widetilde{W}_{ij} = \mathfrak{h}(W_{ij})$ and the U_{ij} represent classical measurement errors which are independent across (i, j) and independent of the X_i. The assumption of classical measurement error after transformation is fairly standard in the dietary assessment literature, and has been used by Kipnis et al. (2009) and Zhang et al. (2011). Also, assuming independence between measurement errors is also typical in cases where the replications are far apart in time. The type of data that we analyze in Section 5.5 and was the inspiration for our work has replications which are taken at least 3 months distant, which is a standard sampling pattern for replicates, and hence are widely separated in time.

Depending on the cases that the density f_U of the U_{ij} is (a) known; or (b) unknown, f_U is usually assumed to be symmetric and continuous. Moreover the X_i are independent and identically distributed and their density f_X is unknown. Throughout, for any random variable Z, we use f_Z to denote its density.

Finally, following standard classical error model assumptions about replicated contaminated data, we assume that the measured food intake on two days where an individual has consumed the food are independent given X_i. Specifically, since the replicates satisfy the classical error model, we assume that, for $1 \leqslant i \leqslant n$ and $j \neq j'$,

$$\begin{aligned}\mathbb{P}(\widetilde{W}_{ij} \leqslant v, \widetilde{W}_{ij'} \leqslant w \mid W_{ij} > 0, W_{ij'} > 0, X_i = x) \\ = \mathbb{P}(X_i + U_{ij} \leqslant v | X_i = x) \mathbb{P}(X_i + U_{ij'} \leqslant w | X_i = x)\,.\end{aligned}$$

These assumptions imply that

$$\mathbb{P}(\widetilde{W}_{ij} \leqslant v | W_{ij} > 0, X_i = x) = \mathbb{P}(U_{ij} \leqslant v - x)\,, \tag{2.1}$$

$$\begin{aligned}\mathbb{P}(\widetilde{W}_{ij} \leqslant v, \widetilde{W}_{i'j} \leqslant w \mid W_{ij} > 0, W_{ij'} > 0, X_i = x) \\ = \mathbb{P}(U_{ij} \leqslant v - x) \mathbb{P}(U_{ij'} \leqslant w - x)\,.\end{aligned} \tag{2.2}$$

Our goal is to estimate the distribution function of the usual intake, defined in the literature as the random variable $T_i = \mathrm{E}(W_{ij} \mid X_i)$. We emphasize that X_i is not the i^{th} person's usual intake, T_i is, and that X_i is a latent variable related to usual intake.

3 Methodology

3.1 Basic calculations

To find an expression for $F_T(t) = \mathbb{P}(T_i \leqslant t)$, we first express the random variable T_i as

$$\begin{aligned}T_i = \mathrm{E}(W_{ij} \mid X_i) &= \mathrm{E}\left[W_{ij} \left\{\mathbb{I}(W_{ij} = 0) + \mathbb{I}(W_{ij} > 0)\right\} \mid X_i\right] \\ &= \mathrm{E}(W_{ij} \mid W_{ij} > 0, X_i)\, H_{\boldsymbol{\beta}}(X_i)\,, \\ &= H_{\boldsymbol{\beta}}(X_i)\, (\mathfrak{h}^{-1} * f_U)(X_i)\,,\end{aligned} \tag{3.1}$$

where to write the last equation we have used the fact that $\mathrm{E}\left(W_{ij} \mid W_{ij} > 0, X_i\right) = \int \mathfrak{h}^{-1}(v) f_U(v - X_i)\,\mathrm{d}v = (\mathfrak{h}^{-1} * f_U)(X_i)$, which follows from (2.1), and we have assumed that $\mathfrak{h}^{-1}$ is well defined on the whole real line. We deduce that

$$F_T(t) = \mathbb{P}\left\{H_{\boldsymbol{\beta}}(X_i)\,(\mathfrak{h}^{-1} * f_U)(X_i) \leqslant t\right\} = \int_{A_{\boldsymbol{\beta}}(t)} f_X(x)\,\mathrm{d}x\,,$$

where

$$A_{\boldsymbol{\beta}}(t) = \left\{x : H_{\boldsymbol{\beta}}(x)(\mathfrak{h}^{-1} * f_U)(x) \leqslant t\right\}. \tag{3.2}$$

It is easy to see that $A_{\boldsymbol{\beta}}(t) = \bigcup_{i=1}^{p}[a_{2i-1}, a_{2i}]$, where $a_1 < a_2 < \ldots < a_{2p}$, a_1 and a_{2p} do not need to be finite and, unless $H_{\boldsymbol{\beta}}(x)(\mathfrak{h}^{-1} * f_U)(x)$ oscillates infinitely many times, p is finite. In this notation, $F_T(t)$ can be expressed as

$$\begin{aligned} F_T(t) = \sum_{i=1}^{p} \int_{a_{2i-1}}^{a_{2i}} f_X(x)\,\mathrm{d}x \; &= \sum_{i=1}^{2p} (-1)^i \int_{-\infty}^{a_i} f_X(x)\,\mathrm{d}x \\ &= \sum_{i=1}^{2p} (-1)^i \mathcal{I}_T(a_i)\,\mathbb{I}(|a_i| < \infty) + \mathbb{I}(a_{2p} = \infty)\,, \end{aligned} \tag{3.3}$$

where, for all $y \in \mathbb{R}$,

$$\mathcal{I}_T(y) = \int_{-\infty}^{y} f_X(x)\,\mathrm{d}x\,. \tag{3.4}$$

Thus, in order to estimate $F_T(t)$, it suffices to estimate the a_j's, which depend on the unknown β_0 and β_1, and the function $\mathcal{I}_T$, which depends on the unknown f_X. We show how to do this in Section 3.3.

Remark 3.1. *While p and the a_j's depend on t, to simplify presentation we do not make this dependence explicit in our notation.*

To save space, in what follows we define

$$\sum_{j<j'}^{J} = \sum_{j=1}^{J-1} \sum_{j'=j+1}^{J}. \tag{3.5}$$

3.2 When f_U is known

We start by showing how to estimate $\boldsymbol{\beta}$, $\mathcal{I}_T$ and F_T in the case that f_U is known.

The parameter $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ is defined implicitly through $H_{\boldsymbol{\beta}}(x) = \mathbb{P}(W_{ij} > 0 \mid X_i = x)$, which cannot be estimated directly since the X_i's are unobserved. Therefore, $\boldsymbol{\beta}$ cannot be estimated directly and we use an estimating equation approach. Since $\boldsymbol{\beta} \in \mathbb{R}^2$, we need to find two equations which depend on $H_{\boldsymbol{\beta}}$ and which can be estimated from the W_{ij}'s. The fact that $H_{\boldsymbol{\beta}}$ tends to zero in one of its tails restricts the choice of the equations. For example, although $g/H_{\boldsymbol{\beta}} = f_X$ integrates to 1 and we can estimate g from our data, the equation $1 = \int g(x)/H_{\boldsymbol{\beta}}(x)\, dx$ is not very stable numerically because it involves dividing by $H_{\boldsymbol{\beta}}$.

After investigation, in the case where only two replicates are available per individual, we found that the equations at (3.6) and (3.7) below were numerically stable and easy to estimate from our data. See Remark 3.2 below for an even more stable approach in the case with three or more replicates. To define (3.6) and (3.7), recall from Section 2 that, for $j \neq j'$, W_{ij} and $W_{ij'}$ are independent conditionally on X_i. Letting $g = f_X H_{\boldsymbol{\beta}}$, we deduce that

$$\begin{aligned}
p_{W_+,W_+} &\equiv \mathbb{P}(W_{ij} > 0, W_{ij'} > 0) \\
&= \int \mathbb{P}(W_{ij} > 0, W_{ij'} > 0 \mid X_i = x) f_X(x)\, \mathrm{d}x = \int H_{\boldsymbol{\beta}}(x) g(x)\, \mathrm{d}x\,, \qquad (3.6) \\
m_+ &\equiv \mathrm{E}\{\widetilde{W}_{ij} \mathbb{I}(W_{ij} > 0) \mathbb{I}(W_{ij'} > 0)\} = \int \mathrm{E}\big\{\widetilde{W}_{ij}\, \mathbb{I}(W_{ij} > 0) \mid X_i = x\big\} g(x)\, \mathrm{d}x \\
&= \int \mathrm{E}\big(\widetilde{W}_{ij} \mid W_{ij} > 0, X_i = x\big) H_{\boldsymbol{\beta}}(x) g(x)\, \mathrm{d}x = \int x H_{\boldsymbol{\beta}}(x) g(x)\, \mathrm{d}x\,. \qquad (3.7)
\end{aligned}$$

Here we used the fact that $\mathrm{E}\big(\widetilde{W}_{ij} \mid W_{ij} > 0, X_i = x\big) = \int u f_U(u - x)\, \mathrm{d}u = x$, which follows

from (2.1) and the symmetry of f_U. Using (3.5), we can estimate p_{W_+,W_+} and m_+ by

$$\widehat{p}_{W_+,W_+} = \frac{2}{nJ(J-1)} \sum_{i=1}^{n} \sum_{j<j'}^{J} \mathbb{I}(W_{ij} > 0, W_{ij'} > 0)\,, \tag{3.8}$$

$$\widehat{m}_+ = \frac{1}{nJ(J-1)} \sum_{i=1}^{n} \sum_{j<j'}^{J} (\widetilde{W}_{ij} + \widetilde{W}_{ij'}) \mathbb{I}(W_{ij} > 0, W_{ij'} > 0)\,. \tag{3.9}$$

To deduce estimators of β_0 and β_1 from (3.6)–(3.9), it remains to estimate g on the right hand sides of (3.6) and (3.7). Now $g = f_X H_{\boldsymbol{\beta}}$, and f_X cannot be estimated easily since the X_i's are unobserved. However since we observe $\widetilde{W}_{ij}$ when $W_{ij} > 0$, we can estimate the density of $\widetilde{W}_{ij} | W_{ij} > 0$, which can be related to g by conditioning on X_i, as follows. Letting $p_{W_+} = \mathbb{P}(W_{jk} > 0)$, an unconditional probability, and using (2.1), we have

$$f_{\widetilde{W}_{jk}|W_{jk}>0}(v) = \int f_{\widetilde{W}_{ij}|W_{ij}>0, X_i=x}(v) H_{\boldsymbol{\beta}}(x) f_X(x)\,\mathrm{d}x / p_{W_+} = \int f_U(v-x) g(x)\,\mathrm{d}x / p_{W_+}\,.$$

If $\phi_{\widetilde{W}_+}$ and ϕ_U denote the Fourier transforms of $p_{W_+} f_{\widetilde{W}_{jk}|W_{jk}>0}$ and f_U, respectively, the Fourier inversion theorem implies that, if $\phi_U(t) > 0$ for all $t \in \mathbb{R}$,

$$g(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \frac{\phi_{\widetilde{W}_+}(t)}{\phi_U(t)}\,\mathrm{d}t\,. \tag{3.10}$$

Now when $p_{W_+} > 0$, $\phi_{\widetilde{W}_+}(t)$ can be estimated unbiasedly by

$$\widehat{\phi}_{\widetilde{W}_+}(t) = \frac{1}{nJ} \sum_{j=1}^{n} \sum_{k=1}^{J} e^{it\widetilde{W}_{jk}} \mathbb{I}\,(W_{jk} > 0), \tag{3.11}$$

see Lemma F.1 in the Supplementary Material. However, $\widehat{\phi}_{\widetilde{W}_+}$ is essentially an empirical characteristic function and is too unreliable in its tails to be plugged directly into (3.10). In related standard deconvolution problems, it is common to attenuate tail effects by a weight function, the Fourier transform of a function called a kernel, scaled by a smoothing parameter h; see e.g. Carroll and Hall (1988) and Stefanski and Carroll (1990). Using similar ideas, let

ϕ_K denote the Fourier transform of a real and symmetric kernel function K and let $h > 0$ be a bandwidth. Assuming that $\phi_U(t) \neq 0$ for all t, we estimate $g(x)$ by

$$\widehat{g}(x;h) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\frac{\widehat{\phi}_{\widetilde{W}_+}(t)}{\phi_U(t)}\,\phi_K(ht)\,\mathrm{d}t = \frac{1}{nJh}\sum_{j=1}^{n}\sum_{k=1}^{J} K_U\Big(\frac{x-\widetilde{W}_{jk}}{h};h\Big)\mathbb{I}(W_{jk}>0)\,, \quad (3.12)$$

where

$$K_U(x;h) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\frac{\phi_K(t)}{\phi_U(t/h)}\,\mathrm{d}t \;= \frac{1}{2\pi}\int_{-\infty}^{\infty}\cos(tx)\frac{\phi_K(t)}{\phi_U(t/h)}\,\mathrm{d}t\,. \quad (3.13)$$

Finally, we estimate $\boldsymbol{\beta} = (\beta_0,\beta_1)^T$ by the value $\widehat{\boldsymbol{\beta}} = (\widehat{\beta}_0,\widehat{\beta}_1)^T$ which satisfies simultaneously

$$\int H_{\widehat{\boldsymbol{\beta}}}(x)\widehat{g}(x;h_{\boldsymbol{\beta}})\,\mathrm{d}x \;= \widehat{p}_{W_+,W_+}\;,\quad \int xH_{\widehat{\boldsymbol{\beta}}}(x)\widehat{g}(x;h_{\boldsymbol{\beta}})\,\mathrm{d}x \;= \widehat{m}_+\,, \quad (3.14)$$

where $h_{\boldsymbol{\beta}}$ is a particular choice of h that will be discussed in Section 5.1.

Remark 3.2. *If we have $J \geqslant 3$ replicates per individual, instead of using m_+ at (3.7), we use $p_{W_+,W_+,W_+} \equiv \mathbb{P}(W_{ij}>0,W_{ik}>0,W_{i\ell}>0) = \int H^2_{\boldsymbol{\beta}}(x)g(x)\,\mathrm{d}x$, which, if $J \geqslant 3$, is easier to estimate than m_+. Indeed, we can estimate p_{W_+,W_+,W_+} by $\widehat{p}_{W_+,W_+,W_+} = c^{-1}\sum_{i=1}^{n}\sum_{j\neq k\neq \ell}\mathbb{I}(W_{ij}>0,W_{ik}>0,W_{i\ell}>0)$, where the second sum is for $1 \leqslant j,k,\ell \leqslant J$ and c is the number of terms in the quadruple sum. To estimate the integral, we replace g by $\hat{g}$. Note that $p_{W_+,W_+,W_+} = \int H^2_\beta(x)g(x)\,dx$ whereas $m_+ = \int xH_\beta(x)g(x)\,dx$. The latter integral involves a multiplicative x term which takes values between $-\infty$ and ∞, so that the error of estimation when estimating $g(x)$ and $H_\beta(x)$ (see (3.14)) are magnified by this multiplicative x term. On the other hand, in p_{W_+,W_+,W_+}, the multiplicative x term is replaced by $H_\beta(x)$, where the latter takes values in $[0,1]$ which attenuates the estimation error of $g(x)$ and $H_\beta(x)$. In Section 4 we develop theory for the version of our estimator of $\boldsymbol{\beta}$ that uses m_+, but the theory for the version of the estimator that uses $\widehat{p}_{W_+,W_+,W_+}$ is almost identical. In particular, as long as $J \geqslant 3$, the statements of our theorems are identical.*

Next we show how to estimate $\mathcal{I}_T(y) = \int_{-\infty}^{y} f_X(x)\,\mathrm{d}x$ at (3.4). By definition of g, we have

$f_X = H_{\boldsymbol{\beta}}^{-1} g$, which can be estimated by $\widehat{f}_X = H_{\widehat{\boldsymbol{\beta}}}^{-1} \widehat{g}(\cdot; h)$, with $\widehat{g}$ and $\widehat{\boldsymbol{\beta}}$ as above, and where $h > 0$ is a bandwidth that may differ from $h_{\boldsymbol{\beta}}$ at (3.14). Since H is a strictly increasing cumulative distribution function, when $\widehat{\beta}_1 < 0$, $\max_{x \in (-\infty, y]} H^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 x) = H_{\widehat{\boldsymbol{\beta}}}^{-1}(y) < \infty$ for all fixed y; moreover, $\widehat{f}_X$ is integrable under standard conditions, see e.g. Fan (1991a) and Section 4.2. In that case we can estimate $\mathcal{I}_T(y)$ by $\widehat{\mathcal{I}}_T(y) = \int_{-\infty}^{y} \widehat{f}_X(x)\,\mathrm{d}x$. When $\widehat{\beta}_1 > 0$, $H_{\widehat{\boldsymbol{\beta}}}^{-1}(x) = H^{-1}(\widehat{\beta}_0 + \widehat{\beta}_1 x) \to \infty$ as $x \to -\infty$, which can cause numerical problems unless we avoid the region where $H_{\widehat{\boldsymbol{\beta}}}^{-1}(x)$ is too large. To do this we can write $\mathcal{I}_T(y) = 1 - \int_{y}^{\infty} f_X(x)\,\mathrm{d}x$ which we can estimate by $\widehat{\mathcal{I}}_T(y) = 1 - \int_{y}^{\infty} \widehat{f}_X(x)\,\mathrm{d}x$. To summarise, for all $y \in \mathbb{R}$ we estimate $\mathcal{I}_T(y)$ by

$$\widehat{\mathcal{I}}_T(y) = \begin{cases} \int_{-\infty}^{y} \widehat{f}_X(x)\,\mathrm{d}x \;= \int_{-\infty}^{y} H_{\widehat{\boldsymbol{\beta}}}^{-1}(x) \widehat{g}(x; h)\,\mathrm{d}x & \text{when } \widehat{\beta}_1 \leqslant 0\,, \\ 1 - \int_{y}^{\infty} \widehat{f}_X(x)\,\mathrm{d}x \;= 1 - \int_{y}^{\infty} H_{\widehat{\boldsymbol{\beta}}}^{-1}(x) \widehat{g}(x; h)\,\mathrm{d}x & \text{when } \widehat{\beta}_1 > 0\,. \end{cases} \tag{3.15}$$

Finally, to estimate F_T at (3.3), it remains to construct estimators of the a_j's. Recall from Section 3.1 that these are defined implicitly through $A_{\boldsymbol{\beta}}$ in (3.2) as $A_{\boldsymbol{\beta}}(t) = \bigcup_{i=1}^{p} [a_{2i-1}, a_{2i}]$, where $a_1 < \ldots < a_p$. For each $t \in \mathbb{R}$ we can estimate $A_{\boldsymbol{\beta}}(t)$ by $A_{\widehat{\boldsymbol{\beta}}}(t) = \big\{x : H_{\widehat{\boldsymbol{\beta}}}(x)(\mathfrak{h}^{-1} * f_U)(x) \leqslant t\big\}$ with $\widehat{\boldsymbol{\beta}}$ as above. We deduce estimators $\widehat{a}_j$ of the a_j's by expressing $A_{\widehat{\boldsymbol{\beta}}}(t)$ as $A_{\widehat{\boldsymbol{\beta}}}(t) = \bigcup_{i=1}^{p} [\widehat{a}_{2i-1}, \widehat{a}_{2i}]$, where $\widehat{a}_1 < \ldots < \widehat{a}_{2p}$. Finally, taking $\widehat{\mathcal{I}}_T$ as in (3.15), for each $t \in \mathbb{R}$ we can estimate $F_T(t)$ by

$$\widehat{F}_T(t) = \sum_{i=1}^{2p} (-1)^i \widehat{\mathcal{I}}_T(\widehat{a}_i)\, \mathbb{I}(|\widehat{a}_i| < \infty) + \mathbb{I}(\widehat{a}_{2p} = \infty)\,. \tag{3.16}$$

3.3 When f_U is unknown

In practice the error density f_U is not always known, and in such cases it needs to be estimated from the data. If we have a parametric model for f_U, the unknown parameters can be estimated from the $\widetilde{W}_{jk}$'s, exploiting (2.2), and then we can replace ϕ_U in our estimators from Section 3 by the resulting estimator of ϕ_U. For example if the only unknown is the

error variance σ_U^2, and recalling the notation in (3.5), it can be estimated by

$$\widehat{\sigma}_U^2 = \frac{\sum_{j=1}^n \sum_{k<k'}^J (\widetilde{W}_{jk} - \widetilde{W}_{jk'})^2 \mathbb{I}(W_{jk} > 0, W_{jk'} > 0)}{2 \sum_{j=1}^n \sum_{k<k'}^J \mathbb{I}(W_{jk} > 0, W_{jk'} > 0)}.$$

Things are more complex if we have no parametric model for f_U. In standard deconvolution problems, Delaigle et al. (2008) suggested estimating ϕ_U via the empirical characteristic function of the difference of replicated data. Although our context is different, we can use similar ideas. By Lemma F.3 in the Supplementary Material, we have $\phi_U = \phi_-^{1/2}$, where, for all $u \in \mathbb{R}$, $\phi_-(u) = E\{e^{iu(\widetilde{W}_{jk} - \widetilde{W}_{jk'})} \mid W_{jk} > 0, W_{jk'} > 0\}$ can be estimated by

$$\widehat{\phi}_-(u) = \sum_{j=1}^n \sum_{k<k'}^J e^{iu(\widetilde{W}_{jk} - \widetilde{W}_{jk'})} \mathbb{I}(W_{jk} > 0, W_{jk'} > 0) \Big/ \sum_{j=1}^n \sum_{k<k'}^J \mathbb{I}(W_{jk} > 0, W_{jk'} > 0).$$

Since $\phi_U(u) > 0$ for all $u \in \mathbb{R}$, this suggests estimating $\phi_U(u)$ by

$$\widehat{\phi}_U(u) = \left| \frac{\sum_{j=1}^n \sum_{k<k'}^J \cos\{u(\widetilde{W}_{jk} - \widetilde{W}_{jk'})\} \mathbb{I}\left(W_{jk} > 0, W_{jk'} > 0\right)}{\sum_{j=1}^n \sum_{k<k'}^J \mathbb{I}(W_{jk} > 0, W_{jk'} > 0)} \right|^{1/2}. \tag{3.17}$$

As in Section 3.2, to estimate F_T we need to estimate $\boldsymbol{\beta}$, $\mathcal{I}_T$ and the a_j's. Proceeding as in that section, for $\boldsymbol{\beta}$, we need to solve (3.14), replacing there $\widehat{g}$ at (3.12) by a version that uses $\widehat{\phi}_U$ instead of ϕ_U. However this requires twice integrating $\widehat{\phi}_U$, which is unreliable in its tails and appears at the denominator of (3.12). This causes technical problems which can be alleviated by replacing ϕ_U by a tail-corrected $\widehat{\phi}_U$.

Specifically, in the tails we use a ridge function ρ which is such that it keeps $\widetilde{\phi}_U$ from getting too small. That is, we take $\widetilde{\phi}_U(u) = \widehat{\phi}_U(u)\mathbb{I}\{\widehat{\phi}_U(u) \geqslant \tau_n\} + \rho(u)\mathbb{I}\{\widehat{\phi}_U(u) < \tau_n\}$, where $\tau_n > 0$ is a threshold. Then, we estimate $\boldsymbol{\beta}$ by the value $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1)^T$ which satisfies (3.14), replacing there $\widehat{g}$ at (3.12) by a version that uses $\widetilde{\phi}_U$ instead of $\widehat{\phi}_U$.

To estimate $\mathcal{I}_T$, first recall from the explanations above (3.15) that, to avoid dividing by $H_{\boldsymbol{\beta}}$ when the latter is too close to zero, we use two different formulae depending on

whether $\beta_1 \leqslant 0$ or $\beta_1 > 0$. In the case where $\boldsymbol{\beta}$ is estimated, we could take the two-part estimator at (3.15) and replace $\widehat{\boldsymbol{\beta}}$ there by $\widetilde{\boldsymbol{\beta}}$ and ϕ_U in $\widehat{g}$ by $\widetilde{\phi}_U$ as above. However, numerical experiments suggest that integrating g in the expressions for $\mathcal{I}_T$ in Section 3.2 before estimating it improves the resulting estimator. This leads to

$$\mathcal{I}_T(y) = \begin{cases} H_{\boldsymbol{\beta}}^{-1}(y)G(y) - \int_{-\infty}^{y} G(x)\{H_{\boldsymbol{\beta}}^{-1}(x)\}' \,\mathrm{d}x & \text{when } \beta_1 \leqslant 0 \\ 1 - p_{W_+} + H_{\boldsymbol{\beta}}^{-1}(y)G(y) + \int_{y}^{\infty} G(x)\{H_{\boldsymbol{\beta}}^{-1}(x)\}' \,\mathrm{d}x & \text{when } \beta_1 > 0\,, \end{cases} \tag{3.18}$$

which we can then estimate by

$$\widetilde{\mathcal{I}}_T(y) = \begin{cases} H_{\widetilde{\boldsymbol{\beta}}}^{-1}(y)\widetilde{G}(y;h) - \int_{-\infty}^{y} \widetilde{G}(x;h)\{H_{\widetilde{\boldsymbol{\beta}}}^{-1}(x)\}' \,\mathrm{d}x & \text{when } \widetilde{\beta}_1 \leqslant 0 \\ 1 - \widehat{p}_{W_+} + H_{\widetilde{\boldsymbol{\beta}}}^{-1}(y)\widetilde{G}(y;h) + \int_{y}^{\infty} \widetilde{G}(x;h)\{H_{\widetilde{\boldsymbol{\beta}}}^{-1}(x)\}' \,\mathrm{d}x & \text{when } \widetilde{\beta}_1 > 0\,, \end{cases} \tag{3.19}$$

with $\widetilde{\boldsymbol{\beta}}$ as above, where

$$\widetilde{G}(x;h) = \frac{\widehat{p}_{W_+}}{2} - \frac{1}{2\pi}\int \phi_K(hu)\,\Im\left\{u^{-1}e^{-iux}\widehat{\phi}_{\widetilde{W}_+}(u)/\widetilde{\phi}_U(u)\right\}\,\mathrm{d}u\,, \tag{3.20}$$

$$\widehat{p}_{W_+} = \frac{1}{nJ}\sum_{j=1}^{n}\sum_{k=1}^{J}\mathbb{I}(W_{jk} > 0) \tag{3.21}$$

are estimators of $G(x) = \int_{-\infty}^{x} g(y)\,\mathrm{d}y$ and $p_{W_+} = \mathbb{P}(W_{ij} > 0)$, and where the bandwidth h may differ from $h_{\boldsymbol{\beta}}$.

Estimating the a_j's is simpler. We proceed as in Section 3.2, except that we replace $f_U(x)$ by a kernel estimator $\widehat{f}_U(x) = (2\pi)^{-1}\int e^{-iux}\phi_K(h_U u)\widehat{\phi}_U(u)\,\mathrm{d}u$, with $\widehat{\phi}_U$ at (3.17), K a kernel function and $h_U > 0$ a bandwidth. In other words, for all $t \in \mathbb{R}$ we estimate $A_{\boldsymbol{\beta}}(t)$ by $\widetilde{A}_{\widetilde{\boldsymbol{\beta}}}(t) = \{x : H_{\widetilde{\boldsymbol{\beta}}}(x)(\mathfrak{h}^{-1} * \widehat{f}_U)(x) \leqslant t\}$ with $\widetilde{\boldsymbol{\beta}}$ as above, and express $\widetilde{A}_{\widetilde{\boldsymbol{\beta}}}(t)$ as $\widetilde{A}_{\widetilde{\boldsymbol{\beta}}}(t) = \bigcup_{i=1}^{p}[\widetilde{a}_{2i-1}, \widetilde{a}_{2i}]$ where $\widetilde{a}_1 < \ldots < \widetilde{a}_{2p}$ are our estimators of the a_j's. In some cases, it might be needed to truncate the convolution integral in $\mathfrak{h}^{-1} * \widehat{f}_U$ to ensure that the latter is finite with probability one. This is always done in practice, where integrals are approximated

by finite sums.

Finally, following (3.3), we estimate $F_T(t)$ by

$$\widetilde{F}_T(t) = \sum_{i=1}^{2p} (-1)^i \widetilde{\mathcal{I}}_T(\widetilde{a}_i)\, \mathbb{I}(|\widetilde{a}_i| < \infty) + \mathbb{I}(\widetilde{a}_{2p} = \infty)\,. \tag{3.22}$$

4 Theory

4.1 Cases considered in the theory

In the nonparametric deconvolution literature, it is well known that the smoothness of the error distribution has a major impact on the rate of convergence of estimators, which depends on the tail behaviour of ϕ_U (Stefanski and Carroll, 1990; Fan, 1991b). It is standard to distinguish between two main classes of errors: ordinary smooth errors, which are such that ϕ_U decays to zero polynomially fast in its tails, and supersmooth errors, where ϕ_U decays exponentially fast. Proving theoretical results for the two cases requires similar but different arguments. Moreover, the supersmooth error case is the least interesting theoretically because in that case, nonparametric estimators converge at a very slow logarithmic rate: this is true also for our estimator. More interesting results can be derived in the ordinary smooth error case. Therefore, since our technical arguments are already very long, we present theory for the ordinary smooth case only.

Remark 4.1. *Despite the logarithmic convergence rates in the supersmooth error case, it is well known in the deconvolution literature that nonparametric estimators usually work well for those errors too. Indeed, various factors such as the size of the error variance can influence on performance of estimators, so that in the supersmooth error case, practical performance is often better than predicted by the standard asymptotics. See for example Delaigle (2008) who uses a double asymptotic approach taking both sample size and the magnitude of the error variance into account. In particular, even in the supersmooth case, nonparametric*

estimators typically perform better than parametric estimators unless we have a rough idea of a parametric model that is not too far from the true curve to estimate. In our numerical work we will consider both ordinary smooth and supersmooth errors and we will see that even when the errors are supersmooth (for example, normally distributed), our estimator performs well despite the theoretical logarithmic rates, and significantly outperforms a parametric estimator based on incorrect parametric assumptions.

For the same reason, we derive theory only for the case where $\beta_1 > 0$. This is the interesting case in practice, because larger X_i should lead to a higher probability of consumption. The only difference between that case and the case where $\beta_1 < 0$ is that, when $\beta_1 < 0$, we need to adjust some of the arguments in the proof of Proposition C.2 in the Supplementary Material Section C.2 for taking into account the fact that, when $\beta_1 < 0$, the equation $H_{\boldsymbol{\beta}}(x)\mathfrak{h}^{-1} * f_U(x) = t$ can have multiple solutions. This can be done using relatively standard but long arguments. All our other results in the Supplementary Material hold also when $\beta_1 < 0$.

In Section 4.3 where we derive theory in the case where f_U is unknown, the main challenge is to track the impact that replacing ϕ_U by $\widetilde{\phi}_U$ has on properties of our estimator. It can be proved that estimating $\boldsymbol{\beta}$ only has second order effects on our results. Therefore, and again given that our technical arguments are already very long, in the case where f_U is unknown we prove our results under the assumption that $\boldsymbol{\beta}$ is known.

Finally, as usual with nonparametric curve estimation problems, we can use either finite order kernels, which have a finite number of non-vanishing moments, or infinite order kernels such as the sinc kernel defined by $\phi_K(t) = \mathbb{I}[-1, 1](t)$. Infinite order kernels have better theoretical properties, but in practice they tend to produce wiggly estimators, and so in our work we use finite order kernels. We note that the proofs for the sinc kernel would be even easier than our proofs.

4.2 Theory when f_U is known

We will need the following assumptions.

Assumption A

(A1) f_U is a symmetric density and $\phi_U(u) > 0$ for all u. Moreover, ϕ_U has three continuous derivatives and there exist finite constants $c_U > 0$ and $\alpha > 1$ such that $\lim_{|u|\to\infty} |u|^\alpha \phi_U(u) = c_U^{-1}$ and $\lim_{|u|\to\infty} |u|^{\alpha+1}\phi_U'(u) = -\alpha c_U^{-1}$, where ϕ_U' is the derivative of ϕ_U.

(A2) K is real, continuous and symmetric and is such that ϕ_K vanishes outside $(-1, 1)$ and has $m + 3$ continuous and bounded derivatives, for some positive integer m. Also, $\phi_K(0) = 1$ and $\phi_K(u) = 1 + O(|u|^{m+1})$ as $u \to 0$.

(A3) H is twice continuously differentiable, $H(x) > 0$ for all $x \in \mathbb{R}$, H' is bounded, $\int |x|H'(x)\,\mathrm{d}x < \infty$ and $\int |xH''(x)|\,\mathrm{d}x < \infty$. Also, $\phi_{H'}$ is continuously differentiable, and there exist two constants $D, \vartheta > 1$ such that for all $u \in \mathbb{R}$, $\max\{|\phi_{H'}(u)|, |\phi_{H'}'(u)|\} \leqslant D\min(1, |u|^{-\vartheta})$.

(A4) $f_{\widetilde{W}_{jk}|W_{jk}>0}$ is bounded and there exists $\delta > 0$ such that $\int x^{4+\delta} f_{\widetilde{W}_{jk}|W_{jk}>0}(x)\,\mathrm{d}x < \infty$.

(A5) $g = f_X H_{\boldsymbol{\beta}}$ is continuous, $\int x^2 g(x)\mathrm{d}x < \infty$, and there exist constants $\gamma > 0$ and $C_g > 0$ such that for all $u \in \mathbb{R}$, $|\phi_g(u)| \leqslant C_g(1 + |u|)^{-\gamma}$. Also, g has m bounded derivatives and $g^{(m)}$ is Lipschitz continuous, with m in (A2). In addition, there exists $c_g > 0$ such that for all $x \in \mathbb{R}$ and $j \leqslant m$ $|xg^{(j)}(x)| < c_g$, and the function $xg^{(m)}(x)$ is Lipschitz continuous.

(A6) $2 \leqslant J < \infty$.

(A7) $p_{W_+} > 0$ and $p_{W_+,W_+} > 0$, with p_{W_+} as on page 12 and p_{W_+,W_+} as at (3.6).

(A8) For $\widetilde{h} = h$ and $\widetilde{h} = h_{\boldsymbol{\beta}}$, as $n \to \infty$, $\widetilde{h} \to 0$, $n\widetilde{h} \to \infty$, $\widetilde{h}|\log\widetilde{h}|^2 \to 0$, $\widetilde{h}^{\alpha-1/2}(\log n)^{1/2} \to 0$ and $n^{1/2}(\log n)^{-1}\widetilde{h}^{\alpha+1/2} = \infty$ with α as in (A1).

(A9) There exists a constant $c_{\boldsymbol{\beta}} > 0$ such that $\beta_0, \beta_1, \widehat{\beta}_0, \widehat{\beta}_1 \in (-c_{\boldsymbol{\beta}}, c_{\boldsymbol{\beta}})$. Moreover, for any $\epsilon > 0$ and $\bar{\boldsymbol{\beta}} \in \mathbb{R}^2$ with $\|\bar{\boldsymbol{\beta}}\|_\infty < c_{\boldsymbol{\beta}}$, $\inf_{|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}|>\epsilon}\{\int H_{\bar{\boldsymbol{\beta}}}(x)g(x)\,\mathrm{d}x - p_{W_+,W_+}\}^2 + \{\int xH_{\bar{\boldsymbol{\beta}}}(x)g(x)\,\mathrm{d}x - m_+\}^2 > 0 = \{\int H_{\boldsymbol{\beta}}(x)g(x)\,\mathrm{d}x - p_{W_+,W_+}\}^2 + \{\int xH_{\boldsymbol{\beta}}(x)g(x)\,\mathrm{d}x - m_+\}^2$.

Assumption (A1) is similar to the one used in Fan (1991a) and Masry (1993) to prove asymptotic normality of a deconvolution kernel density estimator in the ordinary smooth error case. The only difference is that we assume that $\alpha > 1$, instead of $\alpha \geqslant 0$ or $\alpha \geqslant 1$, and that ϕ_U has three (instead of two) continuous derivatives, which we need for proving the consistency of $\widehat{\boldsymbol{\beta}}$. Those conditions are satisfied by many distributions. Assumption (A2) is

fairly standard in the deconvolution literature; see e.g. Fan (1991a,b) and Masry (1993a). It is not very restrictive since we can choose the kernel. Assumption (A3) is not very restrictive in our context either and is satisfied, for example, by the logistic and the normal distributions. Assumption (A4) is used to obtain the rate of convergence of $\widehat{\boldsymbol{\beta}}$; it is satisfied by any distribution with finite first five moments. The conditions imposed on g in Assumption (A5) are similar to those imposed on f_X in Fan (1991a,b) and Masry (1993a), except for the condition on $xg^{(j)}(x)$, which is used to control the asymptotic bias of $\widehat{\boldsymbol{\beta}}$. Assumption (A6) just states that we need at least $J = 2$ replicates, which is necessary to identify β_0 and β_1 and estimate the error density; see also Delaigle et al. (2008). Assumption (A7) cannot be avoided if we observe some nonzero data. Assumption (A8) is of the same type as conditions usually imposed in the deconvolution literature.

Assumption (A9) is rather technical. Conditions such as $\beta_0, \beta_1, \widehat{\beta}_0, \widehat{\beta}_1 \in (-c_{\boldsymbol{\beta}}, c_{\boldsymbol{\beta}})$ are often used in parametric estimation problems, where they are needed to establish convergence rates. The condition $\inf_{|\boldsymbol{\beta}-\bar{\boldsymbol{\beta}}|>\epsilon}\{\int H_{\bar{\boldsymbol{\beta}}}(x)g(x)\,\mathrm{d}x - p_{W_+,W_+}\}^2 + \{\int xH_{\bar{\boldsymbol{\beta}}}(x)g(x)\,\mathrm{d}x - m_+\}^2 > 0 = \{\int H_{\boldsymbol{\beta}}(x)g(x)\,\mathrm{d}x - p_{W_+,W_+}\}^2 + \{\int xH_{\boldsymbol{\beta}}(x)g(x)\,\mathrm{d}x - m_+\}^2$ guarantees that the system of equations formed by (3.6) and (3.9) admits a unique solution in the vicinity of the true value of $\boldsymbol{\beta}$, see for instance van der Vaart (1998), Chapter 5. In practice, since H is known, this assumption can be verified numerically by checking that it is satisfied for different values of $\boldsymbol{\beta}$.

Theorem 4.2 below establishes the asymptotic behaviour of $\widehat{F}_T$ in the case $\beta_1 > 0$. Its proof is given in the Supplementary Material Section C, and uses arguments of a similar type as those used by Hall and Lahiri (2008), Dattner et al. (2011), Dattner and Reiser (2013), Delaigle and Hall (2015) and Datta et al. (2018). As mentioned in Section 4.1, a similar result can be established in the case that $\beta_1 < 0$, using standard but long arguments. In that case, the expression of $\widehat{F}_T(t) - F_T(t)$ at (4.1) would contain a sum of correlated normally distributed terms. Specifically, instead of $\xi(t)$ in Theorem 4.2 below, there would be a sum of, say, $\xi_k(t)$'s, plus a sum of deterministic terms of order h^{m+1} similar to the third and fourth terms on the right hand side of (4.1).

Theorem 4.2. *Under Conditions (A1) to (A9), if $\beta_1 > 0$ and $\mathfrak{h}$ is strictly increasing and*

continuous, for any $t \in \mathbb{R}_+$ *such that* $\sup_{x\in\mathbb{R}} \mathfrak{h}^{-1} * f_U(x) > t$, *as* $n \to \infty$ *we have*

$$\begin{aligned}\widehat{F}_T(t) - F_T(t) =& n^{-1/2}h^{-\alpha+1/2}\{\xi(t) + o_{\mathbb{P}}(1)\} + (n^{-1/2}h_{\boldsymbol{\beta}}^{-\alpha+1}\log n + h_{\boldsymbol{\beta}}^{m+1})\zeta_n(t,\boldsymbol{\beta}) \\ &+ h^{m+1}\frac{\mu_{K,m+1}}{(m+1)!}\Big[H_{\boldsymbol{\beta}}^{-1}\{x_{\boldsymbol{\beta}}(t)\}g^{(m)}\{x_{\boldsymbol{\beta}}(t)\} \\ &+ \int_{x_{\boldsymbol{\beta}}(t)}^{\infty} g^{(m)}(x)\{H_{\boldsymbol{\beta}}^{-1}(x)\}'\,\mathrm{d}x\Big] + o(h^{m+1})\,, \end{aligned} \tag{4.1}$$

where $\zeta_n(t,\boldsymbol{\beta}) = O_{a.s}(1)$, $\xi(t)$ *is a normal random variable with zero mean and variance* $\sigma^2\{x_{\boldsymbol{\beta}}(t)\} = p_{W_+}J^{-1}f_{\widetilde{W}_{jk}|W_{jk}>0}\{x_{\boldsymbol{\beta}}(t)\}H_{\boldsymbol{\beta}}^{-2}\{x_{\boldsymbol{\beta}}(t)\}\int_{-\infty}^{\infty}\mathcal{K}^2(x)\,\mathrm{d}x$, $x_{\boldsymbol{\beta}}(t)$ *is the unique solution to the equation* $H_{\boldsymbol{\beta}}(x)\mathfrak{h}^{-1} * f_U(x) = t$, $\mathcal{K}(x) = c_U\pi^{-1}\int_0^1 u^{\alpha-1}\phi_K(u)\sin(ux)\,\mathrm{d}u$ *and* $\mu_{K,m+1} = \int u^{m+1}K(u)\,\mathrm{d}u$.

It can be seen from our proof of the theorem that the second term on the right hand side of (4.1) comes only from the estimation of $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$ (if $\boldsymbol{\beta}$ is known and does not need to be estimated then this term disappears). In addition to satisfying Condition (A8), if $h_{\boldsymbol{\beta}}$ is chosen so that $h_{\boldsymbol{\beta}}/h \to 0$ and $h_{\boldsymbol{\beta}}^{1/2}\log n \to 0$ as $n \to \infty$, then this term is asymptotically negligible compared to the others, so that estimating $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}}$ has no impact on the asymptotic behaviour of $\widehat{F}_T(t) - F_T(t)$. The fastest convergence rate of our estimator is then obtained by taking the first and third terms of the same size, that is, by taking $h \asymp n^{-1/(2m+2\alpha+1)}$, which leads to $\widehat{F}_T(t) = F_T(t) + O_P\{n^{-(m+1)/(2m+2\alpha+1)}\}$.

Remark 4.3. *Theorem 4.2 provides a pointwise consistency rate for* $\widehat{F}_T$ *and opens the question to whether a uniform consistency rate could be obtained in a setting similar to ours. Although providing a rigorous answer would required a fair amount of additional pages of technical computations, elements from our proofs combined with arguments presented in the proof of Theorem 2.2 in Masry (1993) suggest it could be possible to do so, and to show that for any compact* $\mathcal{C} \subset \mathbb{R}$ *satisfying* $\sup_{x\in\mathbb{R}} \mathfrak{h}^{-1} * f_U(x) > \sup_{t\in\mathcal{C}} t$, *it holds as* $n \to \infty$ *that* $\sup_{t\in\mathcal{C}} |\widehat{F}_T - F_T|$ *is almost surely of order* $n^{-1/2}h^{-\alpha+1/2}(\log n)^{1/2} + n^{-1/2}h_{\boldsymbol{\beta}}^{-\alpha+1}\log n + h_{\boldsymbol{\beta}}^{m+1} + h^{m+1}$. *Theoretical arguments supporting this remark are discussed in the Supplementary Material Section D.*

4.3 Theory when f_U is unknown

In the case where f_U is unknown, the large sample behaviour of the estimator $\widetilde{F}_T$ defined in Section 3.3 can be investigated by expressing $\widetilde{F}_T - F_T$ as

$$\widetilde{F}_T - F_T = (\widetilde{F}_T - \widetilde{F}_T^0) + (\widetilde{F}_T^0 - \widehat{F}_T^0) + (\widehat{F}_T^0 - F_T)\,, \tag{4.2}$$

where $\widetilde{F}_T^0$ and $\widehat{F}_T^0$ are versions of, respectively, $\widetilde{F}_T$ and $\widehat{F}_T$ where $\boldsymbol{\beta}$ and $A_{\boldsymbol{\beta}}$ are known, see their definitions in the Supplementary Material Section E. The first term on the right hand side of (4.2) comes from the effect that estimating $\boldsymbol{\beta}$ has on asymptotic properties of our estimator; the second term reflects the impact of estimating ϕ_U by $\widetilde{\phi}_U$ in a scenario where $\boldsymbol{\beta}$ and $A_{\boldsymbol{\beta}}$ would be known, and the third term is a version of the left hand side of (4.1) where $\boldsymbol{\beta}$ is known. Hence, roughly speaking, when studying the behaviour of $\widetilde{F}_T$, only the second term on the right hand side of (4.2) is significantly different from terms that appear when studying $\widehat{F}_T$ in Theorem 4.2. Therefore, as our technical arguments are extremely long, we provide rigourous results only for the second term, see Theorem 4.4 below.

For the third term, it can be readily deduced from Theorem 4.2 that $\widehat{F}_T^0 - F_T = O_{\mathbb{P}}(n^{-1/2}h^{\alpha-1/2} + h^{m+1})$. Regarding the first term, by combining arguments similar to those used in our proof of Theorems 4.2 and 4.4, it can be proved that $\widetilde{F}_T(t) - \widetilde{F}_T^0(t) = O_{\mathbb{P}}\{a_n\,(1 + b_n)\}$, where $a_n = n^{-1/2}\log n\, h_{\boldsymbol{\beta}}^{-\alpha+1} + h_{\boldsymbol{\beta}}^{m+1}$, which also appears in Theorem 4.2, is a factor induced by the estimation of $\boldsymbol{\beta}$ by $\widetilde{\boldsymbol{\beta}}$, and $b_n = n^{-1/2}(\log n)^{1/2}h^{-3\alpha} + h^{\gamma-2\alpha} + \mathbb{I}(2\alpha = \gamma)|\log h|$ arises because of the estimation of ϕ_U by $\widetilde{\phi}_U$, see also Theorem 4.4 below. Finally, to study the second term on the right hand side of (4.1), we need the following additional conditions.

Assumption B

(B1) There exist two constants $R > 0$ and $\delta \in (0, 1)$ such that, for all $x > R$, $\mathbb{P}(|U_{11} - U_{12}| > x) \leqslant (\log x)^{-1/\delta}$.

(B2) $\lim_{n\to\infty} h \log n \,< \infty$ and $\lim_{n\to\infty} n^{1/2}(\log n)^{-1/2}h^{2\alpha} = \infty$, with α as in (A1).

(B3) $0 < \tau_n < h^{\alpha+\delta'}$ for some $\delta' > 0$, with α as in (A1).

Assumption (B1) is used to obtain an almost sure bound on the difference between $\widetilde{\phi}_U$ and ϕ_U, and is satisfied whenever $\mathrm{E}(|U|^\delta) < \infty$ for some $\delta > 0$, which is arguably not very

restrictive. Assumption (B2) is similar to Assumption (iv) in Theorem 3.1 of Delaigle et al. (2008).

The next theorem establishes asymptotic properties for the second term on the right hand side of (4.1). See the Supplementary Material Section E for a proof.

Theorem 4.4. *Assume that $\beta_1 \neq 0$. Under Conditions (A1) to (A8) and (B1) to (B3), there exists a constant $\eta > 0$ such that for any $t \in \mathbb{R}_+$ and sufficiently large n,*

$$\begin{aligned}\left|\widetilde{F}_T^0(t) - \widehat{F}_T^0(t)\right| \leqslant \eta \sum_{j=1}^{2p} \mathbb{I}(|a_j| < \infty) H_{\boldsymbol{\beta}}^{-2}(a_j) n^{-1/2} (\log n)^{1/2} \\ \times \left\{1 + n^{-1/2}(\log n)^{1/2} h^{-3\alpha} + h^{\gamma-2\alpha} + \mathbb{I}(2\alpha = \gamma)|\log h|\right\} \ a.s.\end{aligned}$$

Combining this theorem with the discussion provided above, we deduce that

$$\widetilde{F}_T(t) - F_T(t) = O_{\mathbb{P}}\Big[\big\{a_n + n^{-1/2}(\log n)^{1/2}\big\}(1 + b_n)\Big] + O_{\mathbb{P}}(n^{-1/2}h^{\alpha-1/2} + h^{m+1})$$

where $a_n = n^{-1/2} \log n \, h_{\boldsymbol{\beta}}^{-\alpha+1} + h_{\boldsymbol{\beta}}^{m+1}$ and $b_n = n^{-1/2}(\log n)^{1/2} h^{-3\alpha} + h^{\gamma-2\alpha} + \mathbb{I}(2\alpha = \gamma)|\log h|$. For comparison, in the case where f_U is known in Theorem 4.2 we established that $\widehat{F}_T(t) - F_T(t) = O_{\mathbb{P}}(a_n) + O_{\mathbb{P}}\big(n^{-1/2}h^{-\alpha+1/2} + h^{m+1}\big)$. Thus, estimating f_U causes extra error terms of order $n^{-1/2}(\log n)^{1/2}$ and b_n.

5 Numerical aspects

5.1 Bandwidth selection when f_U is known

In the case where f_U is known, we need to choose two bandwidths, h and $h_{\boldsymbol{\beta}}$. The bandwidth $h_{\boldsymbol{\beta}}$ is less important than h because it is used in auxiliary step, where we compute an estimator of g used only to estimate $\boldsymbol{\beta}$. Now g and the density $f_{\widetilde{W}_{jk}|W_{jk}>0}$ of the data $\widetilde{W}_{ij}$ for which $W_{ij} > 0$ are related via the equation

$$f_{\widetilde{W}_{jk}|W_{jk}>0}(v) = \int f_{\widetilde{W}_{ij}|W_{ij}>0, X_i=x}(v) H_{\boldsymbol{\beta}}(x) f_X(x) \, \mathrm{d}x / p_{W_+} = \int f_U(v-x) g(x) \, \mathrm{d}x / p_{W_+} \, .$$

Therefore, the density obtained by deconvolving $f_{\widetilde{W}_{jk}|W_{jk}>0}$ from f_U equals g/p_{W_+}, which suggests taking $h_{\boldsymbol{\beta}}$ equal to the deconvolution plug-in bandwidth of Delaigle and Gijbels (2002, 2004) computed from the $\widetilde{W}_{ij}$'s for which $W_{ij} > 0$. Following the discussion under Theorem 4.2, we could multiply this bandwidth by a negative power of n so that $h_{\boldsymbol{\beta}} = o(h)$, but we found this to be unnecessary in practice.

Once we have obtained our estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ computed using $h_{\boldsymbol{\beta}}$, we need to compute h used by our estimator $\widehat{F}_T$ of F_T. We propose to use a SIMEX (Simulation Extrapolation) procedure of the type proposed by Delaigle and Hall (2008), as follows. First, if we knew F_T, we could choose h so as to minimise $D(\widehat{F}_T, F_T) \equiv \int |\widehat{F}_T(t) - F_T(t)|\, dF_T(t)$. Instead, the SIMEX approach consists in simulating two levels of data, SIMEX 1 and SIMEX 2, that are even more contaminated than the original data, and extrapolate from there the bandwidth that minimises $D(\widehat{F}_T, F_T)$.

At the SIMEX k level, for $k = 1, 2$, we create data $(W_{k,ij}, \widetilde{W}_{k,ij}, T_{k,i})$ which contain k additional levels of measurement errors compared to the $(W_{ij}, \widetilde{W}_{ij}, T_i)$'s. Using the notation $W_{0,ij} = W_{ij}$, at the SIMEX k level ($k = 1, 2$), for $i = 1, \ldots, n$ and $j = 1, \ldots, J$ we proceed as follows. First, generate $U_{k,ij} \sim f_U$. If $W_{k-1,i1} > 0$, take $T_{k,i} = H_{\boldsymbol{\beta}}(W_{k-1,i1})\,(\mathfrak{h}^{-1} * f_U)(W_{k-1,i1})$ and $\widetilde{W}_{k,ij} = \widetilde{W}_{k-1,i1} + U_{k,ij}$ and $W_{k,ij} = \mathfrak{h}^{-1}(\widetilde{W}_{k,ij})$; do not define $T_{k,i}$, $\widetilde{W}_{k,ij}$ and $W_{k,ij}$ otherwise. Relabel these data as $T_{k,1}, \ldots, T_{k,n_k}$, $\widetilde{W}_{k,1j}, \ldots, \widetilde{W}_{k,n_kj}$ and $W_{k,1j}, \ldots, W_{k,n_kj}$, where n_k is the number of nonzero $W_{k-1,i1}$'s. Then, with probability $1 - H_{\widehat{\boldsymbol{\beta}}}(\widetilde{W}_{k-1,i1})$, set $W_{k,ij}$ to zero.

For $k = 1, 2$, in SIMEX, the distribution F_{T_k} of the $T_{k,i}$ plays the role of F_T, and we observe the $T_{k,i}$. Therefore, in addition to computing our estimator $\widehat{F}_{T_k}$ of F_{T_k} using the method from Section 3.2 applied to the $W_{k,ij}$, we can also compute $\widehat{F}_{T_k,\mathrm{emp}}$, the empirical distribution function of the $T_{k,i}$'s. The latter is a better estimator than the former, so that we can reasonably approximate $D(\widehat{F}_{T_k}, F_{T_k}) = \int |\widehat{F}_{T_k}(t) - F_{T_k}(t)|\, \mathrm{d}F_{T_k}(t)$ by $\widehat{D}_k(\widehat{F}_{T_k}, \widehat{F}_{T_k,\mathrm{emp}}) = n^{-1} \sum_i |\widehat{F}_{T_k,\mathrm{emp}}(T_{k,i}) - \widehat{F}_{T_k}(T_{k,i})|$. Therefore, we can choose the bandwidth h_k for estimating F_{T_k} by minimising $\widehat{D}_k$.

As the SIMEX data were constructed from the original data using the same measurement error structure as that which relates the $\widetilde{W}_{ij}$'s to the X_i's, then paraphrasing Delaigle and Hall (2008), $\widetilde{W}_{2,ij}$ measures $\widetilde{W}_{1,ij}$ and $\widetilde{W}_{1,ij}$ measures $\widetilde{W}_{ij}$ in the same way as $\widetilde{W}_{ij}$ measures X_i. As in Delaigle and Hall (2008), this suggests that the relationship between h_2 and h_1

mimics that between h_1 and h, in the sense that $h_2/h_1 \approx h_1/h$. This motivates us to choose our bandwidth for computing $\widehat{F}_T$ as $h = h_1^2/h_2$. As pointed by Delaigle and Hall (2008), this approach is too variable because the bandwidths h_k depends on the particular SIMEX sample that has been generated. Like them, to stabilise the procedure, at both SIMEX levels we generate several, B say, SIMEX samples, and choose h_k that minimises the average of the resulting B distances $\widehat{D}_k(\widehat{F}_{T_k}, \widehat{F}_{T_k,\text{emp}})$. In our simulations we took $B = 20$.

5.2 Implementation when f_U is unknown

In the case where f_U is unknown and is estimated nonparametrically as in Section 3.3, we need to select three additional parameters: a ridge function ρ, a threshold τ_n and the bandwidth h_U used to compute $\widehat{f}_U$. The ridge and τ_n are needed only to avoid using $\widehat{\phi}_U$ when it is too close to zero. In standard deconvolution problems considered in Delaigle et al. (2008), Delaigle and Meister (2008) and Delaigle and Hall (2016), these authors argued that we can take ρ equal to the characteristic function of a Laplace random variable with variance equal to the empirical variance of U, and we follow their recommendation. In their case they take τ_n equal to $\widehat{\phi}_U(t^*)$, where t^* is the smallest $t > 0$ at which $\widehat{\phi}_U(t)$ has a local minimum, but often this t^* is too large. We refine their approach by taking t^* equal to the smallest $t > 0$ at which $\widehat{\phi}_U(t)$ reaches its largest local maximum. This intuition is that outside its main body, the wiggles of $\widehat{\phi}_U$ correspond to pure noise, and anything smaller that the largest of those wiggles should correspond to noise.

To choose h_U, recall that this bandwidth is used by our estimator $\widehat{f}_U$ computed from data $\widetilde{U}_i \sim f_U * f_U$, where the $\widetilde{U}_i$'s denote the sample of $\widetilde{W}_{jk} - \widetilde{W}_{jk'} \mid W_{jk} > 0, W_{jk'} > 0$'s. Since our goal is to estimate $\mathfrak{h}^{-1} * f_U$, if we knew f_U we would choose h_U that minimises the integrated squared error ISE $= \int\{\mathfrak{h}^{-1} * \widehat{f}_U(x) - \mathfrak{h}^{-1} * f_U(x)\}^2 \,\mathrm{d}x$, but we do not know f_U and so instead, we use a SIMEX procedure.

For this, consider estimating $f_{U,1} = f_U * f_U$ and $f_{U,2} = f_{U,1} * f_{U,1}$, using the version of our estimator $\widehat{f}_U$ applied to data $\widetilde{U}_{i,2} \sim f_{U,2}$ and $\widetilde{U}_{i,3} \sim f_{U,2} * f_{U,2}$. Here, $\widetilde{U}_{i,2}$ and $\widetilde{U}_{i,3}$ can be obtained by taking the sum of, respectively, four and eight independent $\widetilde{U}_i$'s obtained by drawing randomly with replacement from the $\widetilde{U}_i$'s. We can also construct error-free data $\widetilde{U}_{i,2} \sim f_{U,2}$ and $\widetilde{U}_{i,1} \sim f_{U,1}$, where the latter are obtained by taking the sum

of two independent $\widetilde{U}_i$'s. Using those error-free data, we can also compute the standard kernel density estimators $\widehat{f}_{U,1,EF}$ and $\widehat{f}_{U,2,EF}$ of $f_{U,1}$ and $f_{U,2}$. Since these converge faster to $f_{U,1}$ and $f_{U,2}$ than our deconvolution estimator, this suggests that, for $k = 1, 2$, we can choose the bandwidth $h_{U,k}$ used to compute $\widehat{f}_{U,k}$ by the value that minimises $\mathrm{ISE}_k = \int\{\mathfrak{h}^{-1} * \widehat{f}_{U,k}(x) - \mathfrak{h}^{-1} * \widehat{f}_{U,k,EF}(x)\}^2 \,\mathrm{d}x$. Then, since the relation between f_{U_1} and f_{U_2} mimics that between f_U and f_{U_1}, this motivates us to assume that $h_U/h_{U,1} \approx h_{U,1}/h_{U,2}$ and take $h_U = h_{U,1}^2/h_{U,2}$. As in Section 5.1, we reduce the variability of the procedure by generating $B = 20$ such samples and taking $h_{U,k}$ that minimises the average of the resulting B ISE_k's.

We also need to choose h as in Section 5.1, but here we cannot use exactly the same SIMEX approach as there since f_U is unknown. To overcome this difficulty, for $k = 1, 2$, instead of generating $U_{k,ij} \sim f_U$, we generate $U_{1,ij}$ and $U_{2,ij}$ by drawing with replacement from, respectively, the $(W_{i1} - W_{i2})/\sqrt{2}$'s for which W_{i1} and W_{i2} are non zero, and the $(W_{i1} - W_{i2} + W_{j1} - W_{j2})/2$'s for which W_{i1},W_{i2}, W_{j1} and W_{j2} are non zero, which is an approximate way of taking $U_{1,ij} \sim f_U * f_U(\sqrt{2}\cdot)$ and $U_{2,ij} \sim f_U * f_U * f_U * f_U(2\cdot)$. Since U_{ij} and the $U_{k,ij}$'s all have the same variance, then the relationship between f_U and $f_U * f_U(\sqrt{2}\cdot)$ is mimicked by that between $f_U * f_U(\sqrt{2}\cdot)$ and $f_U * f_U * f_U * f_U(2\cdot)$. Alternatively we could generate data from $\widehat{f}_U$ but this is time consuming.

Finally, we note that in most cases the estimator $\widetilde{F}_T$ of F_T obtained from the estimator of I_T at (3.19) works better than the one obtained by the formula at (3.15) adapted to the unknown error case as discussed in section 3.3, call it $\widetilde{F}_{T,2}$. However in some cases $\widetilde{F}_T(t)$ gets larger than 1 for some t's, and in that case we use $\widetilde{F}_{T,2}$, unless it too gets larger than 1 and does so for smaller values of t than $\widetilde{F}_T(t)$, in which case we replace $\widetilde{F}_T(t) > 1$ by 1. We found this approach worked better than simply replacing $\widetilde{F}_T(t) > 1$ by 1. Likewise, in some cases $\widetilde{F}_T(t)$ stays away from 1 as t increases, and in that case to we use $\widetilde{F}_{T,2}$.

5.3 Monotonizing the estimator

While our estimators of F_T are consistent, as usual in deconvolution problems, in finite sample they are not guaranteed to be non decreasing functions of t. (In finite samples, standard deconvolution kernel density estimators are not guaranteed to be positive everywhere and so their corresponding distribution function is not guaranteed to be monotone). They can be

made monotone using procedures that exist in the literature. For example, applied to our context, the technique of Dette et al. (2006) monotonizes, on an interval $[a, b]$, an estimator $\check{F}_T$ (for example $\widetilde{F}_T$ or $\widehat{F}_T$) of F_T as follows.

Let $V = F_T(U)$, where $U \sim U[a, b]$ and let f_V, F_V and $F_T^{(-1)}$ denote, respectively, the density of V, the distribution function of V and the inverse of F_T. Then for $v \in [F_T(a), F_T(b)]$, we can write $F_T^{(-1)}(v) = a + (b - a) \int_{F_T(a)}^{v} f_V(x)\, dx$. Thus, to obtain an increasing estimator of $F_T^{(-1)}(v)$ for $v \in [F_T(a), F_T(b)]$, we can take

$$\check{F}_T^{(-1)}(v) = a + (b - a) \int_{\check{F}_T(a)}^{v} \hat{f}_V(x)\, dx\,,$$

where $\hat{f}_V$ is a positive estimator of f_V constructed from a sample $V_1, \ldots, V_N$, with $V_j = \check{F}_T(U_j)$ and $U_j \sim U[a, b]$, for $j = 1, \ldots, N$. We obtain an increasing estimator of F_T on $[a, b]$ by numerically inverting $\check{F}_T^{(-1)}$.

Since f_V is compactly supported and may have jumps at a and b, instead of taking $\hat{f}_V$ to be a standard kernel density estimator, we use the probit transformed version of Geenens (2014) which is designed for this type of density. We used the R code provided by the author, with the least square cross-validation bandwidth suggested there. For a and b, we take $[a, b]$ to be the interval where we are seeking to estimate F (in our case this is the interval used in the figures). However, in the case where f_U is unknown, $\widetilde{F}_T$ is sometimes flat on an interval near zero (recall that $T \geqslant 0$), partly because we set $\widetilde{F}_T$ to zero if it takes negative values. In that case, to avoid introducing significant bias using the monotization procedure, we take a to be the smallest number such that $\widetilde{F}_T$ is not flat.

5.4 Simulations

We applied our method to data from the following four models, where in each case we took H to be the logistic function and $\mathfrak{h}$ to be the log transform, as these are the commonly used in applications:

(i) $X_i \sim \chi^2(10)$, $U_{ij} \sim N(0, \sigma^2)$ and (a): $\boldsymbol{\beta} = (-5, 1.5)^T$ or (b): $\boldsymbol{\beta} = (-5, 1)^T$;

(ii) $X_i \sim N(-2, 2)$, $U_{ij} \sim N(0, \sigma^2)$, and (a): $\boldsymbol{\beta} = (3, 0.3)^T$ or (b): $\boldsymbol{\beta} = (1.6, 0.3)^T$;

(iii) $X_i \sim N(-2, 2)$, $U_{ij} \sim \text{Laplace}(\sigma)$, and (a): $\boldsymbol{\beta} = (3, 0.3)^T$ or (b): $\boldsymbol{\beta} = (1.6, 0.3)^T$;

Table 1: Simulation results. For each estimator $\widehat{F}_T$ and $\widetilde{F}_T$, each model (denoted by M in the table) and each noise to signal ratio NSR, the numbers show $10^3\times$ median [first decile, ninth decile] of 200 values of the IWAD.

		f_U known, estimator $\widehat{F}_T$			f_U unknown, estimator $\widetilde{F}_T$		
M	NSR	$n = 100$	$n = 250$	$n = 500$	$n = 100$	$n = 250$	$n = 500$
(i)(a)	10%	28.6[14.7,52.9]	18.3[9.9,42.2]	14.3[8.2,27.7]	37.3[16.2,98.0]	25.3[12.3,49.2]	19.2[10.1,37.3]
	25%	34.5[16.1,65.4]	23.4[13.6,35.7]	19.5[10.6,29.0]	46.6[20.9,95.0]	32.7[14.2,68.3]	24.3[12.8,47.7]
(i)(b)	10%	30.6[16.6,55.7]	20.1[10.2,43.0]	16.2[9.4,46.6]	50.8[23.6,106]	48.5[17.7,102]	37.8[14.8,77.7]
	25%	32.4[16.5,60.2]	25.5[14.5,39.9]	18.9[12.3,28.5]	53.9[25.5,113]	36.1[16.2,85.3]	29.8[14.7,65.1]
(ii)(a)	10%	31.0[16.0,64.6]	19.6[9.2,61.7]	12.4[6.9,25.4]	33.1[16.8,68.1]	23.1[11.9,45.1]	16.1[7.6,38.5]
	25%	31.0[16.7,62.2]	20.4[11.1,38.6]	14.4[7.5,32.8]	37.3[17.8,82.2]	23.6[12.4,56.0]	18.8[9.5,40.5]
(ii)(b)	10%	32.6[14.1,58.0]	19.0[8.9,46.3]	15.5[7.8,28.6]	38.4[20.7,72.4]	25.8[12.9,50.1]	19.9[9.7,38.6]
	25%	34.2[18.2,64.6]	23.5[12.8,40.1]	18.8[9.6,41.0]	37.9[19.1,82.8]	25.1[10.8,53.1]	20.9[10.5,40.5]
(iii)(a)	10%	28.5[13.9,57.2]	17.6[9.4,32.3]	13.6[7.8,22.5]	32.4[15.2,64.0]	20.4[11.4,37.7]	14.8[8.2,25.8]
	25%	27.7[14.5,52.6]	19.3[9.8,33.5]	13.9[7.0,25.1]	35.1[18.5,73.6]	22.2[10.8,49.9]	16.4[9.0,35.2]
(iii)(b)	10%	31.9[16.9,61.4]	20.6[10.8,35.5]	14.5[8.2,25.2]	34.9[19.7,72.1]	27.8[12.4,53.5]	22.9[9.9,47.7]
	25%	34.8[16.9,60.0]	21.6[11.0,37.4]	14.6[9.3,25.9]	36.7[17.7,76.2]	26.7[12.6,53.1]	21.2[10.7,44.2]
(iv)(a)	10%	34.3[21.2,57.8]	21.9[13.2,39.7]	16.7[9.5,29.6]	33.8[17.8,74.5]	22.6[13.0,42.0]	19.1[10.6,33.2]
	25%	36.9[23.7,61.9]	27.1[19.9,39.9]	22.3[16.9,30.0]	47.2[26.0,89.9]	33.9[18.8,58.7]	25.2[14.4,44.2]
(iv)(b)	10%	31.8[19.3,57.0]	23.5[13.1,46.4]	18.5[11.5,37.2]	37.8[21.2,66.4]	23.9[13.9,40.1]	19.2[11.4,33.4]
	25%	37.3[24.2,60.0]	27.7[19.6,41.5]	24.2[16.6,32.5]	45.2[23.5,108]	33.4[22.0,56.7]	26.5[16.1,40.2]

(iv) $X_i \sim 0.3N(-3,1) + 0.7N(3,1)$, $U_{ij} \sim N(0,\sigma^2)$, and (a): $\boldsymbol{\beta} = (3, 0.7)^T$ or (b): $\boldsymbol{\beta} = (2, 0.7)^T$. For each model, $\mathbb{P}(W_{ij} > 0)$ is larger in case (a) than in case (b), so that we can expect F_T to be easier to estimate in case (a) than in case (b) as we effectively have more data to compute our semiparametric estimators.

In each case, for $j = 1, 2$, and $i = 1, \ldots, n$, we set W_{ij} to zero with probability $H_{\boldsymbol{\beta}}(X_i)$. For $W_{ij} \neq 0$ we took $\widetilde{W}_{ij} = X_i + U_{ij}$, where U_{ij} was independent of X_i, and where we took the parameter σ from the distribution of U_{ij} such that the noise to signal ratio NSR = $\mathrm{var}(U)/\mathrm{var}(X)$ was equal to 10% or 25%. For each configuration, we generated 200 samples of size n equal to 100, 250 or 500.

In each case, we applied our method assuming that the error density was known, where we used the estimator $\widehat{F}_T$ at (3.16), or assuming that the error density was unknown, where we used the estimator $\widetilde{F}_T$ at (3.22) combined with the monotonization procedure from section 5.3. We also computed a naive estimator $\widehat{F}_{T,\mathrm{naive}}$ of F_T obtained by computing the empirical characteristic function of the $\mathfrak{h}(\bar{W}_i)$'s, where $\bar{W}_i = (W_{i1} + W_{i2})/2$, pretending that, when averaged, the W_{ij}'s can be treated as the X_i's. Finally, we computed a parametric

Table 2: Simulation results for $\widehat{F}_{T,\text{naive}}$ and for the parametric estimator which assumes that $\widetilde{W}|W>0 \sim N(\mu_W, \sigma_W^2)$. For each model (denoted by M in the table) and each noise to signal ratio NSR, the numbers show $10^3\times$ median [first decile, ninth decile] of 200 values of the IWAD.

		$\widehat{F}_{T,\text{naive}}$			Parametric estimator		
M	NSR	$n=100$	$n=250$	$n=500$	$n=100$	$n=250$	$n=500$
(i)(a)	10%	65.2[31.7,107]	67.1[44.0,91.2]	66.5[53.6,84.9]	38.7[24.0,69.5]	35.1[23.6,53.4]	34.7[26.6,47.8]
	25%	152[111,195]	156[133,180]	156[141,173]	40.2[23.5,7.0]	35.3[23.6,54.4]	34.9[26.3,48.2]
(i)(b)	10%	72.4[38.8,114]	74.5[50.4,99.5]	73.9[59.3,92.0]	41.7[25.6,66.7]	36.7[26.5,54.2]	36.7[28.7,48.5]
	25%	150[111,194]	155[132,179]	155[139,173]	41.1[26.1,66.8]	36.6[26.3,53.8]	37.0[27.9,48.8]
(ii)(a)	10%	49.4[28.1,83.6]	46.8[28.8,63.5]	45.3[31.3,58.3]	23.6[8.8,42.9]	14.8[5.8,31.6]	10.0[4.2,20.7]
	25%	35.8[21.4,66.1]	30.7[16.9,46.1]	27.4[18.4,38.4]	25.1[9.9,43.6]	15.8[6.3,31.8]	10.0[4.5,20.7]
(ii)(b)	10%	160[134,186]	158[141,175]	158[147,168]	28.9[11.1,53.2]	16.5[7.5,35.2]	12.4[5.1,23.6]
	25%	139[112,167]	137[120,154]	137[126,147]	30.3[11.1,54.3]	17.4[7.7,34.4]	12.2[4.7,23.7]
(iii)(a)	10%	51.0[28.5,75.3]	44.7[31.5,65.4]	44.5[33.2,57.2]	21.6[9.5,47.3]	14.2[5.1,30.7]	9.6[3.9,20.8]
	25%	33.5[21.0,53.0]	26.4[17.4,43.6]	25.0[17.4,35.0]	22.8[9.2,48.8]	15.1[5.2,31.5]	9.7[4.0,20.4]
(iii)(b)	10%	162[135,184]	157[142,173]	157[146,168]	24.5[10.9,52.5]	17.2[7.2,31.9]	11.6[5.3,24.7]
	25%	138[109,160]	132[117,149]	133[123,143]	24.6[12.1,53.4]	17.9[7.8,34.1]	13.0[5.1,25.4]
(iv)(a)	10%	54.0[35.1,78.6]	53.1[40.8,68.9]	54.5[42.5,67.1]	42.5[31.5,67.6]	42.1[33.8,54.5]	40.1[34.4,48.2]
	25%	86.7[60.2,112]	85.2[70.7,103]	85.6[72.8,97.4]	43.5[32.5,64.8]	40.8[33.7,53.5]	39.7[34.4,47.8]
(iv)(b)	10%	69.9[52.9,96.4]	71.8[58.2,88.1]	72.2[60.3,84.9]	40.0[27.0,66.3]	36.5[27.4,49.9]	33.9[27.9,42.2]
	25%	101[76.6,126]	102[88.7,118]	103[91.2,115]	41.7[28.1,64.6]	36.4[27.9,49.5]	33.6[27.9,42.0]

estimator of F_T where in (3.10) we estimated ϕ_{W_+} by p_{W_+} times the characteristic function of $\widetilde{W}|W>0$ (see equation above (3.10)) assuming that $\widetilde{W}|W>0 \sim N(\mu_W, \sigma_W^2)$ with μ_W and σ_W estimated by the empirical mean and variance, and with f_U correctly specified. This parametric assumption is correct for models (ii) and (iii) but it is incorrect for models (i) and (iv).

To assess the performance of our procedures, for each of the 200 generated samples for each combination of model, n, NSR and $\boldsymbol{\beta}$, and for each of $\widehat{F}_T$, $\widehat{F}_{T,\text{naive}}$ and $\widetilde{F}_T$ denoted here generically by $\widehat{F}$, we computed the integrated weighted absolute deviation $\text{IWAD}(\widehat{F}) = \int |\widehat{F}(t) - F_T(t)| f_T(t)\, dt$. For each estimator and each configuration, we obtained 200 values of IWAD of which we computed the first, fifth and ninth deciles. The results, reported in Table 1 for $\widehat{F}_T$ and $\widetilde{F}_T$ and in Table 2 for $\widehat{F}_{T,\text{naive}}$ and the parametric estimator, show clearly that the naive estimator performs very poorly, and so will not be considered for the rest of this section. Unsurprisingly, the results also show that in models (ii) and (iii) where the parametric normality assumption is correct, the parametric estimator outperforms our estimator, although the latter performs reasonably well. However, in models (i) and (iv) where this parametric assumption is incorrect, our semiparametric estimator performs

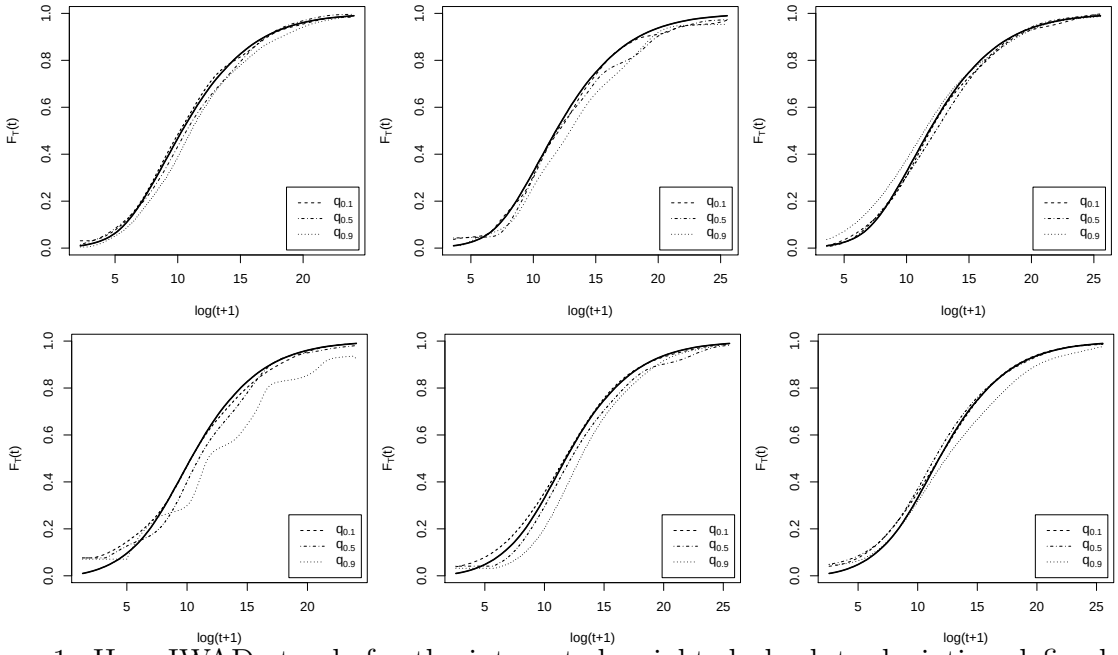

Figure 1: Here IWAD stands for the integrated weighted absolute deviation defined in the text, and NSR stands for the noise to signal ratio defined in the text. Estimated curves corresponding to the first, fifth and ninth deciles of the IWAD of the estimator $\widetilde{F}_T$ at (3.22) (first two columns) or $\widehat{F}_T$ at (3.16) (third column) computed from 200 samples of size $n = 250$ from model (i)(a) (first row) and (i)(b) (second row), when the error density is unknown and NSR=10% (first column) or NSR=25% (second column), or when the error density is known and NSR=25% (third column). The solid line shows the true F_T.

significantly better than the parametric one, even though in those cases the measurement error is normally distributed.

To assess visually the quality of our procedures, for each configuration and each of $\widehat{F}_T$ and $\widetilde{F}_T$, we also plotted the estimated curves corresponding to the three samples that gave the first, fifth and ninth deciles of the IWAD; in the graphs we refer to them as $q_{0.1}$, $q_{0.5}$ and $q_{0.9}$. In the figures, to increase visibility, we plot $\log(t+1)$ versus $F_T(t)$. In Figure 1 we show these curves for model (i)(a) and (i)(b) and $n = 250$ in the case where the error density is known and we use the estimator $\widehat{F}_T$, and in case the error density is unknown and we use the estimator $\widetilde{F}_T$. Together with the table, this figure illustrates without any surprise that estimating F_T is easier when f_U is known than when it needs to be estimated, but that our estimation procedure works well even when f_U needs to be estimated. Together with the table, this figure also illustrates the fact that estimating F_T is easier when the proportion of W_{ij}'s taking the value zero is lower and when the NRS is lower.

In Figure 2 we compare the estimated curves for models (ii)(b) and (iii)(b) and for

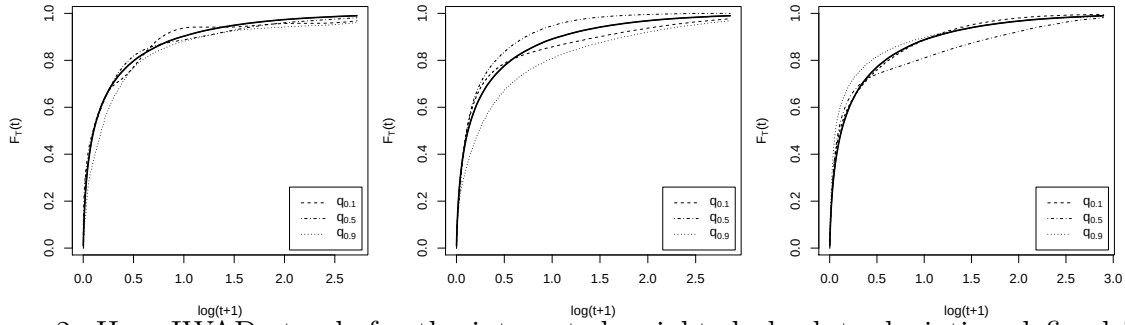

Figure 2: Here IWAD stands for the integrated weighted absolute deviation defined in the text, and NSR stands for the noise to signal ratio defined in the text. Estimated curves corresponding to the first, fifth and ninth deciles of the IWAD of the estimator $\widetilde{F}_T$ at (3.22) computed from 200 samples from model (ii)(b) when $n = 100$ and NSR=10% (first column) or $n = 100$ and NSR=25% (second column) or model (iii)(b) when $n = 100$ and NSR=25% (third column). The solid line shows the true F_T.

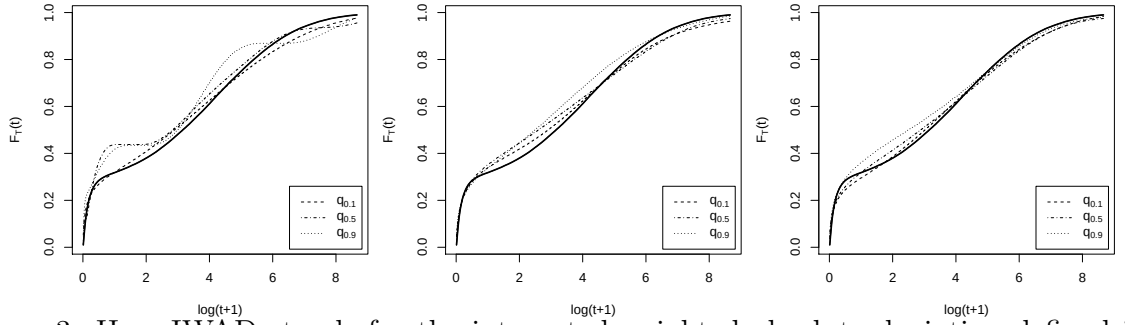

Figure 3: Here IWAD stands for the integrated weighted absolute deviation defined in the text. Estimated curves corresponding to the first, fifth and ninth deciles of the IWAD of the estimator $\widehat{F}_T$ at (3.16) computed from 200 samples from model (iv)(a) when NSR=25% (first column) and $n = 100$ (first column), $n = 250$ (second column) or $n = 500$ (third column). The solid line shows the true F_T.

NSR=10% or 25%. Together with the table, the figure illustrates the fact that estimating F_T is easier when the NSR is smaller and when the error density is Laplace, model (iii), than when it is normal, model (ii), as expected by the theory (normal errors are supersmooth and cause slower convergence rates). Finally, in Figure 3 we show the estimator $\widehat{F}_T$ for model (iv)(a) and NSR=25%, for samples of sizes $n = 100$, 250 and 500. Together with the table, this figure illustrates the fact that our estimators improve as sample size n increases.

5.5 Application

We applied our methodology to data from the Eating at America's Table Study (EATS, Subar et al., 2001). In this study, $n = 965$ participants reported their alcohol intake and

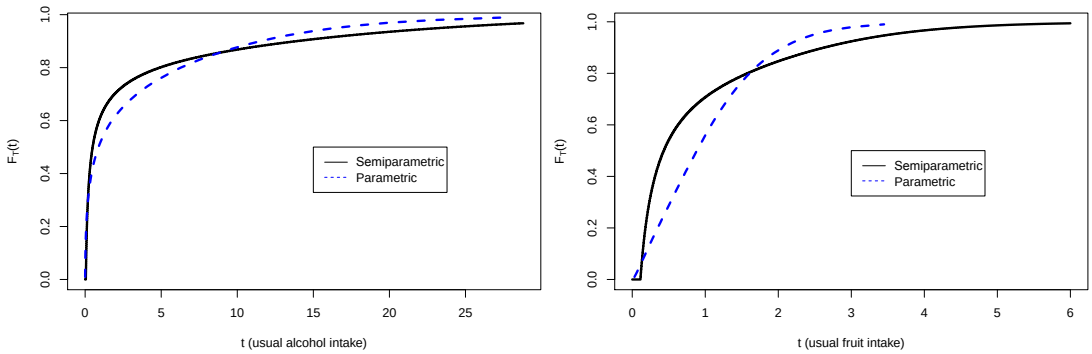

Figure 4: Comparison of our semiparametric estimator $\widetilde{F}_T$ at (3.22) and a fully parametric estimator when T is the usual alcohol (left) or fruit intake (right).

also their total fruit consumption from a 24 hour recall (24HR) on $J = 4$ different and widely spaced days. We took $\mathfrak{h}$ to be the log transform, and H to be the logistic distribution function. A significant fraction of the population eats no fruit on any given day, and an even larger proportion has no alcohol intake on any given day.

We compared our semiparametric estimator of F_T for these data with the fully parametric estimator described in Tooze et al. (2010) and implemented in a SAS program available at `https://epi.grants.cancer.gov/diet/usualintakes/method.html` and written by the U.S. National Cancer Institute (NCI). This estimator relies heavily on the transformed data and the errors to be both normally distributed. To make the data closer to normal, we followed common practice in nutritional epidemiology of eliminating or censoring implausibly small values of the variables. In our analysis, we replaced any such implausibly small intakes by zero: (a) < 0.7 grams of alcohol, the equivalent of < 0.6 ounces of a standard US 5% 12 ounce bottle of beer; and (b) < 0.30 standard servings of fruit, the equivalent of $< 1/3$ of a medium-sized apple. This resulted in more reasonable parametric estimators.

Although the parametric estimator assumes that the error distribution is normally distributed, in this example, this distribution is actually unknown and so when computing our estimator, we used the estimator $\widetilde{F}_T$ from Section 3.3. Because more than 2 replicates per individual are available, to estimate $\boldsymbol{\beta}$ we consider a version of the procedure described in Remark 3.2 adapted to the case where f_U is unknown. Specifically, we estimate $\boldsymbol{\beta}$ by the value $\widetilde{\boldsymbol{\beta}} = (\widetilde{\beta}_0, \widetilde{\beta}_1)^T$ which satisfies $\widehat{p}_{W_+,W_+} = \int H^2_{\boldsymbol{\beta}}(x)\widetilde{g}(x)\,\mathrm{d}x$ and $\widehat{p}_{W_+,W_+,W_+} = \int H^2_{\boldsymbol{\beta}}(x)\widetilde{g}(x)\,\mathrm{d}x$, where $\widehat{p}_{W_+,W_+}$ and $\widehat{p}_{W_+,W_+,W_+}$ are defined respectively at (3.8) and in Remark 3.2, and where $\widetilde{g}$ is a version of $\widehat{g}$ at (3.12) that uses $\widetilde{\phi}_U$ instead of $\widehat{\phi}_U$ (see Section 3.3).

The resulting estimators of F_T are shown in Figure 4. In the case of alcohol consumption, our semiparametric estimator and the existing parametric estimator gave similar results, which suggests that in that case the normality assumptions are reasonable (a q-q plot analysis of these data confirmed this: it indicated only moderate departure from normality). However, for the case of fruit consumption, the two estimators differ significantly, suggesting that in that case the normality assumption is less reasonable (indeed a q-q plot analysis of the data indicated more pronounced departure from normality).

Supplementary Material

The Supplementary Material includes technical details and Matlab code for computing our estimator. We do not have permission to distribute the EATS data set analyzed in this paper, but it can obtained by the National Cancer Institute (NCI, http://www.cancer.gov/) by arranging a Data Transfer Agreement with that institution.

Acknowledgment

Delaigle's work was supported by a discovery project (DP170102434) from the Australian Research Council. Camirand's work was supported by a grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) and by a discovery project (DP170102434) from the Australian Research Council. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030).

References

Carroll, R.J. (2014). Estimating the distribution of dietary consumption patterns. *Stat. Sci.*, **29**, 2–8.

Carroll, R.J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.*, **83**, 1184–1186.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006). *Measurement error in nonlinear models*, 2nd Edn. Chapman and Hall CRC Press, Boca Raton.

Datta, G., Delaigle, A., Hall, P. and Wang, L. (2018). Semi-parametric prediction intervals in small areas when auxiliary data are measured with error. *Statist. Sinica*, **28**, 2309–2335.

Dattner, I., Goldenshluger, A. and Juditsky, A. (2011). On deconvolution of distribution functions. *Ann. Statist.*, **39**, 2477–2501.

Dattner, I. and Reiser, B. (2013). Estimation of distribution functions in measurement error models. *J. Stat. Plan. Infer.*, **143**, 479–493.

Delaigle, A. (2008). An alternative view of the deconvolution problem. *Statist. Sinica*, **18**, 1025–1045.

Delaigle, A. and Gijbels, I. (2002). Estimation of integrated squared density derivatives from a contaminated sample. *J. Roy. Statist. Soc.* Series B, **64**, 869–886.

Delaigle, A. and Gijbels, I. (2004). Practical bandwidth selection in deconvolution kernel density estimation. *Comp. Statist. Data Anal.*, **45**, 249–267.

Delaigle, A. and Hall, P. (2008). Using SIMEX for smoothing-parameter choice in errors-in-variables problems. *J. Amer. Statist. Assoc.*, **103**, 280–287.

Delaigle, A. and Hall, P. (2013). Methodology for nonparametric deconvolution when the error distribution is unknown. *J. Roy. Statist. Soc.* Series B, **78**, 231–252.

Delaigle, A. and Hall, P. (2015). Nonparametric methods for group testing data, taking dilution into account. *Biometrika*, **102**, 871–887.

Delaigle, A., Hall, P. and Meister, A. (2008). On deconvolution with repeated measurements. *Ann. Statist.*, **36**, 665–685.

Delaigle, A. and Meister, A. (2007). Nonparametric regression estimation in the heteroscedastic errors-in-variables problem. *J. Amer. Statist. Assoc.*, **102**, 1416–1426.

Dette, H., Neumeyer, N. and Pilz, K.F. (2006). A simple nonparametric estimator of a strictly monotone regression function. *Bernoulli*, **12**, 469–490.

Diggle, P. J. and Hall, P. (1993). A Fourier approach to non-parametric deconvolution of a density estimate. *J. Roy. Statist. Soc.* Series B, **55**, 523–531.

Dwyer, J., Picciano, M.F., Raiten, D.J, and others (2003). Collection of food and dietary supplement intake data: what we eat in America–NHANES. *J. of Nutrition*, 133, 590S-600S.

Fan, J. (1991a). Asymptotic normality for deconvolution kernel density estimators. *Sankhya*, Series A, **53**, 97–110.

Fan, J. (1991b). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.*, **19**, 1257–1272.

Fan, J. (1993). Adaptively local one-dimensional subproblems with application to a deconvolution problem. *Ann. Statist.*, **21**, 600–610.

Fletcher, D., MacKenzie, D. and Villouta, E. (2005). Modelling skewed data with many zeros: a simple approach combining ordinary and logistic regression. *Environmental and Ecological Statistics*, **12**, 45–54.

Geenens, G. (2014), Probit transformation for kernel density estimation on the unit interval. *J. Amer. Statist. Assoc.*, **109**, 346–358.

Guenther, P.M., Reedy, J., Krebs-Smith, S.M. and Reeve, B.B. (2008a). Evaluation of the Healthy Eating Index-2005. *J. Am. Dietetic Assoc.*, **108**, 1854–1864.

Guenther, P.M., Reedy, J. and Krebs-Smith, S.M. (2008b). Development of the Healthy Eating Index-2005. *J. Am. Dietetic Assoc.*, **108**, 1896–1901.

Guenther, P.M., Kirkpatrick, S.L., Reedy, J., Krebs-Smith, S.M., Buckman, D.W., Dodd, K.W. Casavale, K.O. and Carroll, R.J. (2014). Healthy Eating Index-2010 is a valid and reliable measure of diet quality according to the 2010 Dietary Guidelines for Americans. *J. of Nutrition*, **144**, 399–407.

Hall, P. and Lahiri, S. (2008). Estimation of distributions, moments and quantiles in deconvolution problems. *Ann. Statist.*, **36**, 2110–2134.

Keogh, R.H. and White, I.R. (2011). Allowing for never and episodic consumers when correcting for error in food record measurements of dietary intake. *Biostatistics*, **12**, 624-636.

Kipnis, V., Midthune, D., Buckman, D. W., Dodd, K. W., Guenther, P. M., Krebs-Smith, S. M., Subar, A. F., Tooze, J. A., Carroll, R .J. and Freedman, L. S. (2009). Modeling data with excess zeros and measurement error: Application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, **65**, 1003–1010.

Li, L., Shao, J., and Palta, M. (2005). A longitudinal measurement error model with a semicontinuous covariate. *Biometrics*, **61**, 824–830.

Li, T. and Vuong, Q. (1998). Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivar. Anal.*, **65**, 139–165.

Masry, E. (1993). Strong consistency and rates for deconvolution of multivariate densities of stationary processes. *Stochastic Process. Appl.*, **47**, 53–74

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

Subar, A.F., Thompson, F. E., Kipnis, V., Midthune, D., Hurwitz, P., McNutt, S., McIntosh, A. and Rosenfeld, S. (2001). Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study. *Am. J. of Epid.*, **154**, 1089–1099.

Tian L. Inferences on the mean of zero-inflated lognormal data: the generalized variable approach. (2005). *Statistics in Medicine* .**24**, 3223–3232.

Tooze, J.A., Grunwald, G.K., and Jones, R.H. (2002). Analysis of repeated measures data clumping at zero. *Statist. Methods in Med. Res.*, **11**, 341–355.

Tooze, J.A., Midthune, D., Dodd, K. W., Freedman, L.S., Krebs-Smith, S.M., Subar, A.F., Guenther, P.M., Carroll, R.J. and Kipnis, V. (2006). A new statistical method for estimating the usual intake of episodically consumed foods with application to their distribution. *J. Am. Diet. Assoc.*, **106**, 1575–1587.

Tooze, J.A., Kipnis, V., Buckman, D.W., Carroll, R.J., Freedman, L.S., Guenther, P.M., Krebs-Smith, S.M., Subar, A.F. and Dodd, K.W. (2010). A mixed-effects model approach for estimating the distribution of usual intake of nutrients: the NCI method. *Statistics in Medicine*, **29**, 2857–2868.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Zhang, S. Midthune, D., Pérez, A, Buckman, D.W., Kipnis, V., Freedman, L.S., Dodd, K.W., Krebs-Smith, S.M. and Carroll, R.J. (2011a). Fitting a bivariate measurement error model for episodically consumed dietary components. *Int. J. of Biostat.*, **7**, Issue 1, Article 1.

Zhang, S., Midthune, D., Guenther, P.M., Krebs-Smith, S.M., Kipnis, V., Dodd, K.W., Buckman, D.W., Tooze, J.A., Freedman, L.S. and Carroll, R.J. (2011b). A new multivariate measurement error model with zero-inflated dietary data, and its application to dietary assessment. *Ann. of Applied Stat.*, **5**, 1456–1487.

Zhou, X. and Tu, W. (1999). Comparison of several independent population means when their samples contain log-normal and possibly zero observations. *Biometrics*. **55**, 645–651.