Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://amstat.tandfonline.com/loi/uasa20

Bayesian Copula Density Deconvolution for Zero-Inflated Data in Nutritional Epidemiology

Abhra Sarkar, Debdeep Pati, Bani K. Mallick & Raymond J. Carroll

To cite this article: Abhra Sarkar, Debdeep Pati, Bani K. Mallick & Raymond J. Carroll (2020): Bayesian Copula Density Deconvolution for Zero-Inflated Data in Nutritional Epidemiology, Journal of the American Statistical Association, DOI: 10.1080/01621459.2020.1782220

To link to this article: https://doi.org/10.1080/01621459.2020.1782220

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2020, VOL. 00, NO. 0, 1–13: Applications and Case Studies
https://doi.org/10.1080/01621459.2020.1782220

Bayesian Copula Density Deconvolution for Zero-Inflated Data in Nutritional Epidemiology

Abhra Sarkar[a], Debdeep Pati[b], Bani K. Mallick[b], and Raymond J. Carroll[b,c]

[a]Department of Statistics and Data Sciences, The University of Texas at Austin, Austin, TX; [b]Department of Statistics, Texas A&M University, College Station, TX; [c]School of Mathematical and Physical Sciences, University of Technology Sydney, Broadway, NSW, Australia

ABSTRACT

Estimating the marginal and joint densities of the long-term average intakes of different dietary components is an important problem in nutritional epidemiology. Since these variables cannot be directly measured, data are usually collected in the form of 24-hr recalls of the intakes, which show marked patterns of conditional heteroscedasticity. Significantly compounding the challenges, the recalls for episodically consumed dietary components also include exact zeros. The problem of estimating the density of the latent long-time intakes from their observed measurement error contaminated proxies is then a problem of deconvolution of densities with zero-inflated data. We propose a Bayesian semiparametric solution to the problem, building on a novel hierarchical latent variable framework that translates the problem to one involving continuous surrogates only. Crucial to accommodating important aspects of the problem, we then design a copula based approach to model the involved joint distributions, adopting different modeling strategies for the marginals of the different dietary components. We design efficient Markov chain Monte Carlo algorithms for posterior inference and illustrate the efficacy of the proposed method through simulation experiments. Applied to our motivating nutritional epidemiology problems, compared to other approaches, our method provides more realistic estimates of the consumption patterns of episodically consumed dietary components. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.

ARTICLE HISTORY
Received April 2019
Accepted June 2020

KEYWORDS
Copula; Density deconvolution; Measurement error; Nutritional epidemiology; Zero inflated data

1. Introduction

1.1. Problem Statement

Dietary habits are important for our general health and well-being, having been known to play important roles in the etiology of many chronic diseases. Estimating the long-term average intakes of different dietary components $\mathbf{X}$ and their marginal and joint distributions is thus a fundamentally important problem in nutritional epidemiology.

The dietary component may be a nutrient, like sodium, vitamin A, etc., or a food group, like milk, whole grains, etc. In any case, by the very nature of the problem, $\mathbf{X}$ can never be observed directly. Data are thus often collected in the form of 24-hr recalls of the intakes. Many of the dietary components of interest are daily consumed. Examples include total grains, sodium, etc., the recalls for which are all continuous, comprising only strictly positive intakes. Compounding the challenge, interest may additionally lie in episodically consumed components whose long-term average intake is assumed to be strictly positive but the recalls are semicontinuous, comprising positive recalls for consumption days and exact zero recalls for nonconsumption days. Examples include milk, whole grains, etc.

Since dietary patterns often vary with energy levels, measured in total caloric intake, adjustments with energy provide a way of standardizing the dietary assessments. The recalls for energy are always continuous. From a statistical viewpoint, they can thus be treated just like the regular components, and hence, with some abuse, will be referred to as such.

When the recalls are recorded within a relatively short span of time, it may be assumed that the participants' dietary patterns $\mathbf{X}$ will not have changed significantly over this period. Treating the recalls $\mathbf{Y}$, like the ones shown in Table 1, to be surrogates for the latent $\mathbf{X}$, contaminated by measurement errors $\mathbf{U}$, the problem of estimating the joint and marginal distributions of $\mathbf{X}$ from the recalls $\mathbf{Y}$ then translates to a problem of multivariate deconvolution of densities with exact zero surrogates for some of the components.

Throughout we adopt the following generic notation for marginal, joint, and conditional densities, respectively. For random vectors $\mathbf{S}$ and $\mathbf{T}$, we denote the marginal density of $\mathbf{S}$, the joint density of $(\mathbf{S}, \mathbf{T})$, and the conditional density of $\mathbf{S}$ given $\mathbf{T}$, by the generic notation $f_{\mathbf{S}}$, $f_{\mathbf{S},\mathbf{T}}$, and $f_{\mathbf{S}|\mathbf{T}}$, respectively. Likewise, for univariate random variables S and T, the corresponding densities are denoted by f_S, $f_{S,T}$, and $f_{S|T}$, respectively. Additional summaries of the variables and notations used can be found in Table 2.

CONTACT Abhra Sarkar abhra.sarkar@utexas.edu Department of Statistics and Data Sciences, The University of Texas at Austin, 2317 Speedway D9800, Austin, TX 78712-1823.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

These materials were reviewed for reproducibility.

1.2. The EATS Dataset and Its Prominent Features

The main motivation behind the research being reported here comes from the Eating at America's Table Study (EATS) (Subar et al. 2001), a large scale epidemiological study conducted by the National Cancer Institute in which $i = 1, \ldots, n = 965$ participants were interviewed $j = 1, \ldots, m_i = 4$ times over the course of a year and their 24-hr dietary recalls were recorded.

Data on many different dietary components were recorded in the EATS study, including episodic components milk and whole grains, whose recalls involved approximately 21% and 37% exact zeros, respectively. Table 1 shows the general structure of this dataset for one regularly consumed and one episodically consumed dietary component.

Patterns of conditional heteroscedasticity are also generally very prominent in dietary recall data. See, for example, the right panels of Figure 1 which shows the plot of subject-specific means $\overline{Y}_{\ell,i} = \sum_{j=1}^{m_i} Y_{\ell,i,j}/4$ versus subject-specific variances $S^2_{Y,\ell,i} = \sum_{j=1}^{m_i}(Y_{\ell,i,j} - \overline{Y}_{\ell,i})^2/3$ for the 24-hr recalls of sodium and energy, which provide crude estimates of the underlying true intakes $X_{\ell,i}$ and the conditional measurement error variances $\text{var}(U_{\ell,i,j}|X_{\ell,i})$, respectively, suggesting strongly that $\text{var}(U|X)$ increases as X increases. Similar observation can also be made for positive recalls of episodic components from the middle panels of Figure 2.

Table 1. The general structure of the EATS dataset showing the recalls for one episodically consumed and one regularly consumed dietary component.

Subject	24-hr recalls							
	Episodic component				Regular component			
1	$Y_{e,1,1}$	$Y_{e,1,2}$	$Y_{e,1,3}$	$Y_{e,1,4}$	$Y_{r,1,1}$	$Y_{r,1,2}$	$Y_{r,1,3}$	$Y_{r,1,4}$
2	0	$Y_{e,2,2}$	$Y_{e,2,3}$	$Y_{e,2,4}$	$Y_{r,1,1}$	$Y_{r,2,2}$	$Y_{r,2,3}$	$Y_{r,2,4}$
3	$Y_{e,3,1}$	0	$Y_{e,3,3}$	$Y_{e,3,4}$	$Y_{r,3,1}$	$Y_{r,3,2}$	$Y_{r,3,3}$	$Y_{r,3,4}$
4	0	$Y_{e,4,2}$	$Y_{e,4,3}$	0	$Y_{r,4,1}$	$Y_{r,4,2}$	$Y_{r,4,3}$	$Y_{r,4,4}$
...	...	...	...	...	...	...	...	...
n	$Y_{e,n,1}$	$Y_{e,n,2}$	0	0	$Y_{r,n,1}$	$Y_{r,n,2}$	$Y_{r,n,3}$	$Y_{r,n,4}$

NOTE: Here, $Y_{\ell,i,j}$ is the reported intake for the jth recall of the ith individual for the ℓth dietary component.

As can be seen from the right and middle panels in Figures 1 and 2, respectively, for both regular and episodic components, the variability of the positive recalls naturally decreases to zero as the average intake on consumption days decreases to zero. For all regularly consumed components, the histograms of the recalls are mildly right skewed bell shaped. The histograms for the episodically consumed components are, however, reflected J-shaped—the frequencies of the bins start with their largest value at the left end and then rapidly decrease as we move to the right. These imply that, for regularly consumed components, the distributions of the true long-term average intakes smooth out near both ends, whereas, for episodically consumed components, the distributions of the true long-term average intakes have discontinuities at zero. The right panels of Figure 2 also show that, as expected, individuals consuming an episodic component in smaller amounts also consume it less often on average.

1.3. Existing Methods and Their Limitations

The literature on univariate density deconvolution for continuous surrogates, in which context we denote the variable of interest by X and the measurement errors by U, is massive. The early literature, however, focused on scenarios with restrictive assumptions, such as known measurement error distribution, homoscedasticity of the errors, their independence from X, etc., which are all highly unrealistic, especially in nutritional epidemiology applications like ours. Reviews of these early methods can be found in Carroll et al. (2006) and Buonaccorsi (2010). We cite below some relatively recent ideas that are directly relevant to our proposed solution.

Bayesian frameworks can accommodate measurement errors through natural hierarchies, providing powerful tools for solving complex deconvolution problems, including scenarios when the measurement errors can be conditionally heteroscedastic. Taking such a route, Staudenmayer, Ruppert, and Buonaccorsi (2008) assumed the measurement errors to be normally distributed but allowed the variability of U to depend on X,

Table 2. Variables representing the data and other random variables in our model.

Notation	Description
q	Number of episodically consumed components.
p	Number of regularly consumed components.
$Y_{\ell,i,j}$	Observed recall of the ℓth dietary component for the ith individual on the jth sampling occasion—binary for $\ell = 1, \ldots, q$, zero if the component was not consumed, one otherwise; zero or positive continuous for $\ell = q+1, \ldots, 2q$, representing the reported intakes, zero when the component was not consumed, positive continuous otherwise; positive continuous for $\ell = 2q+1, \ldots, 2q+p$, representing the reported intakes.
$W_{\ell,i,j}$	Proxy recall of the ℓth dietary component for the ith individual on the jth sampling occasion—always continuous; latent for $\ell = 1, \ldots, q$, negative if $Y_{\ell,i,j} = 0$, positive if $Y_{\ell,i,j} = 1$; latent or observed for $\ell = q+1, \ldots, 2q$, latent when the component was not consumed, observed and equals $Y_{\ell,i,j}$ when a positive recall was recorded; positive for $\ell = 2q+1, \ldots, 2q+p$, equaling $Y_{\ell,i,j}$, the reported positive intake.
$X_{\ell,i}$	Long-term daily average intake of the ℓth dietary component for the ith individual, consumption and nonconsumption days combined. Strictly positive and continuous. For $\ell = 1, \ldots, q+p$, the observed recalls $Y_{q+\ell,i,j}$ are unbiased for $X_{\ell,i}$.
$X^+_{\ell,i}$	Long-term daily average intake of the ℓth dietary component for the ith individual, on consumption days only. Strictly positive and continuous. For $\ell = 1, \ldots, q+p$, the proxy recalls $W_{q+\ell,i,j}$'s are unbiased for the $X^+_{\ell,i}$'s.
$P_\ell(X_{\ell,i})$	Probability of reporting positive consumption on the ℓth dietary component by the ith individual on any sampling occasion.
$\widetilde{X}_{\ell,i}$	Functions of $X_{\ell,i}$, $X^+_{\ell,i}$ and $P_\ell(X_{\ell,i})$ such that $W_{\ell,i,j}$ is unbiased for $\widetilde{X}_{\ell,i}$.
$U_{\ell,i,j}$	Measurement errors or pseudo-errors contaminating $\widetilde{X}_{\ell,i}$ to generate $W_{\ell,i,j}$. The $U_{\ell,i,j}$'s are all unbiased for zero. For $\ell = q+1, \ldots, 2q+p$, variability of $U_{\ell,i,j}$ depends on the associated $\widetilde{X}_{\ell,i}$.
$s^2_\ell(\widetilde{X}_{\ell,i})$	Variance function explaining how the conditional variability of $U_{\ell,i,j}$ depends on the associated $\widetilde{X}_{\ell,i}$ for $\ell = q+1, \ldots, 2q+p$.
$\epsilon_{\ell,i,j}$	Scaled measurement error or pseudo-error obtained by scaling $U_{\ell,i,j}$ by $s_\ell(\widetilde{X}_{\ell,i})$. The $\epsilon_{\ell,i,j}$'s are unbiased for zero, homoscedastic, and independent of $\widetilde{X}_{\ell,i}$.
$Z_{\ell,i}$	Long-term daily average normalized intake of the ℓth dietary component for the ith individual, normalized by energy.

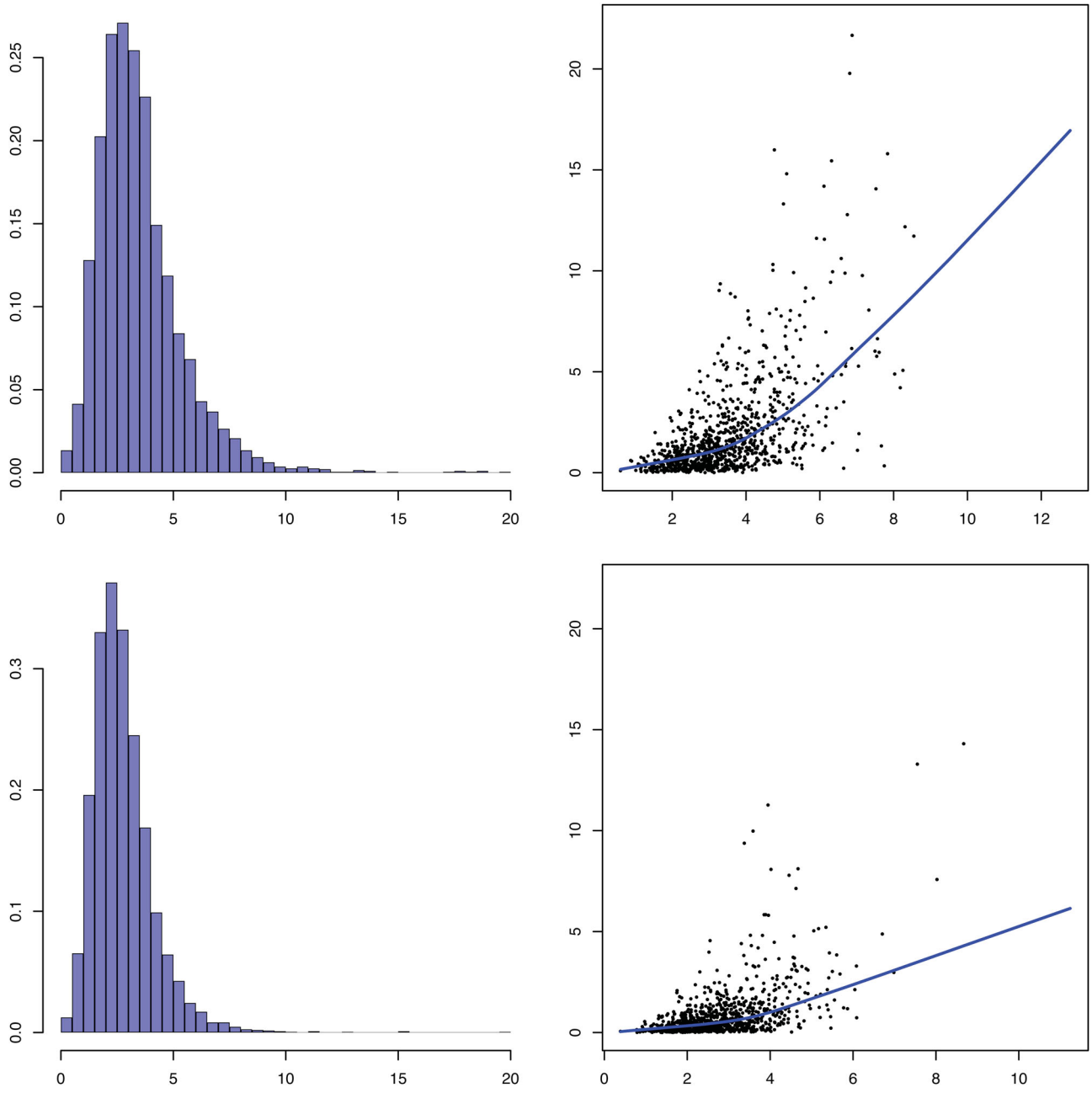

Figure 1. Exploratory plots for sodium (top row) and energy (bottom row). Left panels: Histogram of recalls $Y_{\ell,i,j}$; right panels: subject-specific means $\overline{Y}_{\ell,i}$ versus variances $S^2_{Y,\ell,i}$.

utilizing mixtures of B-splines to estimate f_X as well the conditional variability var$(U|X)$. Sarkar et al. (2014) relaxed the assumption of normality of U, employing flexible mixtures of normals (Escobar and West 1995; Frühwirth-Schnatter 2006) to model both f_X and $f_{U|X}$. Sarkar et al. (2018) extended the methods to multivariate settings, modeling $f_{\mathbf{X}}$ and $f_{\mathbf{U}|\mathbf{X}}$ using mixtures of multivariate normals.

While Staudenmayer, Ruppert, and Buonaccorsi (2008) and Sarkar et al. (2014, 2018) provided progressively flexible frameworks for univariate and multivariate deconvolution with continuously measured surrogates, they cannot directly handle multivariate zero-inflated dietary recall data. There are several restrictive aspects of their approaches that also do not allow them to be straightforwardly extended to deconvolution problems with zero-inflated surrogates, as we outline shortly while describing our proposed approach.

The problem of estimating long-term nutritional intakes of a single episodic dietary component from zero-inflated recall data has previously been considered in Tooze, Grunwald, and Jones (2002), Tooze et al. (2006), Kipnis et al. (2009), and Zhang, Krebs-Smith, et al. (2011). The work was extended to multivariate settings with both episodic and regular components in Zhang, Midthune, et al. (2011). These approaches all worked with component-wise Box–Cox transformed (Box and Cox 1964) positive recalls which were then assumed to decompose into a subject specific random effect component and an error or pseudo-error component. Assumed independent and homoscedastic, these components were then both

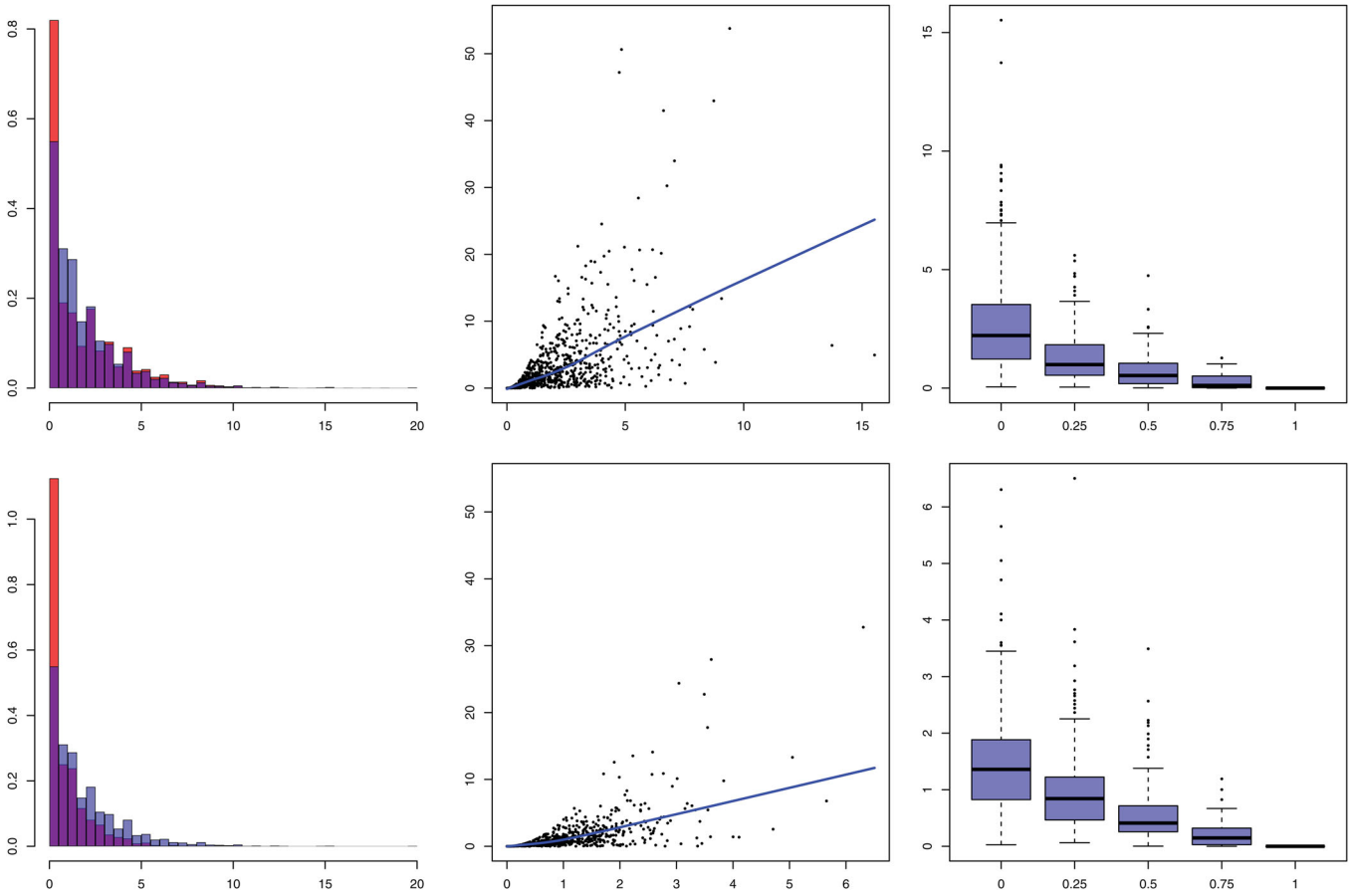

Figure 2. Exploratory plots for milk (top row) and whole grains (bottom row). Left panels: Histogram of recalls $Y_{\ell,i,j}$ (red) and histogram of strictly positive recalls $Y_{\ell,i,j}(>0)$ (blue) superimposed on each other; middle panels: subject-specific means $\overline{Y}_{\ell,i}$ versus subject-specific variances $S^2_{Y,\ell,i}$ when multiple strictly positive recalls are available; right panels: boxplots of proportion of zero recalls versus corresponding subject-specific means $\overline{Y}_{\ell,i}$.

modeled using single component multivariate normal distributions. Estimates of the long-term consumption day intakes were then obtained via individual transformations back to the original scale. Long-term episodic consumptions were finally defined combining these estimates with probabilities of reporting nonconsumptions. As shown in Sarkar et al. (2014), Box–Cox transformations for surrogate observations have severe limitations, including almost never being able to produce transformed surrogates that conform to normality, homoscedasticity, and independence. Transformation–retransformation based methods are thus highly restrictive, even for univariate regularly consumed components.

Despite the limitations, to our knowledge, Zhang, Midthune, et al. (2011) is the only available method that can handle multivariate zero-inflated dietary recall data. It is thus also our main and only competitor.

1.4. Outline of Our Proposed Method

In this article, we develop a Bayesian semiparametric density deconvolution approach specifically designed to address problems with zero-inflated surrogates, carefully accommodating all prominent features of the EATS dataset described above. We build on an augmented latent variable framework which introduces, for each recall of the episodically consumed component, one or two latent continuous proxies, depending on whether the recall was positive or exact zero, effectively translating a deconvolution problem with zero-inflated data to one with all continuous surrogates, albeit some latent ones. This requires modeling an additional pseudo-error distribution for each episodically consumed component, but returns, as potentially useful by-products, estimates of the probabilities of reporting zero recalls for the episodically consumed dietary components. As the right panels of Figure 2 suggest, individuals who consume an episodic component less often (in other words, report more zero recalls) naturally also consume the component in smaller amounts in the long run. The probabilities of reporting zero consumptions are thus informative about the true long-term consumption amounts and conversely. Our proposed latent variable framework appropriately recognizes these features.

Even though the multivariate latent consumptions $\mathbf{X}$ and the associated multivariate errors and pseudo-errors $\mathbf{U}$ become all strictly continuous in our augmented latent variable framework, the approach of Sarkar et al. (2018) to model their distributions using mixtures of multivariate normals is still fraught with serious practical drawbacks as it does not allow much flexibility in modeling the univariate marginals f_{X_ℓ} and $f_{U_\ell|X_\ell}$, especially the marginals of the episodic components which have discontinuities at zero. The issue becomes more critical when inference is based on samples drawn from the posterior using

Markov chain Monte Carlo (MCMC) algorithms. The latent $\mathbf{X}_i$'s are also sampled in the process and the specific parametric form of the assumed multivariate mixture kernel may influence this step in ways that result in density estimates closely resembling its parametric form even when the shape of the true density departs from it.

As opposed to Sarkar et al. (2018) who focused on modeling the joint distributions $f_{\mathbf{X}}$ and $f_{\mathbf{U}|\mathbf{X}}$ first and then deriving the marginals from those estimates, we take the opposite approach of modeling the marginals f_{X_ℓ} and $f_{U_\ell|X_\ell}$ first and then build the joint distributions $f_{\mathbf{X}}$ and $f_{\mathbf{U}|\mathbf{X}}$ by modeling the dependence structures separately using Gaussian copulas. This approach allows us adopt different strategies for modeling the different components of $f_{\mathbf{X}}$ and $f_{\mathbf{U}|\mathbf{X}}$ which proved crucial in accommodating the important features of our motivating datasets. Following Sarkar et al. (2014), we use flexible mixtures of mean restricted normals and mixtures of B-splines to model $f_{U_\ell|X_\ell}$'s and the associated conditional heteroscedasticity functions. Mixtures of normal kernels, as in Sarkar et al. (2014), are, however, not suitable for modeling f_{X_ℓ}'s. We use normalized mixtures of B-splines and mixtures of truncated normal kernels instead which are well suited to model densities with bounded supports and discontinuities at the boundaries.

The literature on copula models in measurement error-free scenarios is vast. See, for example, Nelsen (2007), Joe (2015), Shemyakin and Kniazev (2017), and the references therein. We are, however, unaware of any published work in the context of measurement error problems.

In contrast to Zhang, Midthune, et al. (2011), we model the densities of the latent consumptions and the error and pseudo-errors more directly using flexible models that can accommodate widely varying shapes with discontinuous boundaries as well as conditional heteroscedasticity. In our latent variable framework, the probability of reporting zero recalls depends directly on the latent true consumption day intake, hence informing each other. Applied to our motivating nutritional epidemiology problems, our method thus provides more realistic estimates of the intakes of the episodically consumed dietary components. Additional detailed comparisons of our method with previous approaches for zero-inflated data are presented in Section S.4 in the supplementary materials.

Compared to all previously existing density deconvolution methods, including traditional methods for strictly continuous data as well as methods designed specifically for zero-inflated data, our proposed approach is thus fundamentally novel while also being broadly applicable to both scenarios.

1.5. Outline of the Article

The rest of the article is organized as follows. Section 2 details the proposed Bayesian hierarchical framework. Section 3 presents results of our proposed method applied to the motivating nutritional epidemiology problems. Section 4 concludes with a discussion. A brief review of copula, a detailed comparison of our method with previous approaches to zero-inflated data, an MCMC algorithm to sample from the posterior, simulation studies comparing the proposed method with its main competitor, and some additional results are included in the supplementary materials.

2. Deconvolution Models

2.1. Latent Variable Framework

Our goal is to estimate the marginal and joint consumption patterns of $q + p$ dietary components of which the first q are episodically consumed and the latter p are regularly consumed, including energy. There are a total of n subjects with m_i 24-hr recalls recorded for the ith subject. We let $\mathbf{Y}_{i,j} = (Y_{1,i,j}, \ldots, Y_{2q+p,i,j})^{\mathrm{T}}$ denote the observed data for the jth recall of the ith individual. For $\ell = 1 \ldots, q$, $Y_{\ell,i,j}$ is the indicator of whether the ℓth episodic component is reported to have been consumed. For $\ell = q + 1, \ldots, 2q$, $Y_{\ell,i,j}$ is the reported intake of the ℓth episodically consumed component, and for $\ell = 2q + 1, \ldots, 2q+p$, $Y_{\ell,i,j}$ is the reported intake of the ℓth regularly consumed component. Let $\mathbf{W}_{i,j} = (W_{1,i,j}, \ldots, W_{2q+p,i,j})^{\mathrm{T}}$ denote a vector with all continuous components that are related to the observed data $\mathbf{Y}_{i,j}$ by the relationships

$$\begin{aligned} Y_{\ell,i,j} &= \mathrm{I}(W_{\ell,i,j} > 0), \quad \text{for } \ell = 1, \ldots, q, \\ Y_{\ell,i,j} &= Y_{\ell-q,i,j} W_{\ell,i,j}, \quad \text{for } \ell = q + 1, \ldots, 2q, \\ Y_{\ell,i,j} &= W_{\ell,i,j}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p. \end{aligned} \tag{1}$$

For $\ell = 1, \ldots, q$, $W_{\ell,i,j}$ indicates whether the ℓth episodic component is reported to have been consumed in the jth recall of the ith individual and is always latent except that we know whether it is positive or negative. That is, for $\ell = 1, \ldots, q$, $W_{\ell,i,j}$ is always latent with $W_{\ell,i,j} < 0$ if $Y_{\ell,i,j} = 0$ and $Y_{q+\ell,i,j} = 0$, and $W_{\ell,i,j} \geq 0$ if $Y_{\ell,i,j} = 1$ and $Y_{q+\ell,i,j} > 0$.

For $\ell = q + 1, \ldots, 2q$, $W_{\ell,i,j}$ is latent if the ℓth episodic component is reported to have not been consumed in the jth recall of the ith individual and is observed and equals the reported consumed positive amount $Y_{\ell,i,j}$ otherwise. That is, for $\ell = q + 1, \ldots, 2q$, $W_{\ell,i,j}$ is latent if $Y_{\ell-q,i,j} = 0$ and $Y_{\ell,i,j} = 0$ and is observed with $W_{\ell,i,j} = Y_{\ell,i,j}$ if $Y_{\ell-q,i,j} = 1$ and $Y_{\ell,i,j} > 0$.

For $\ell = 2q+1, \ldots, 2q+p$, $W_{\ell,i,j}$ denotes the reported intake of the ℓth regular component and is always observable. That is, for $\ell = 2q + 1, \ldots, 2q + q$, $W_{\ell,i,j} = Y_{\ell,i,j} > 0$.

We let $\mathbf{X}_i = (X_{1,i}, \ldots, X_{q+p,i})^{\mathrm{T}}$ denote the latent daily average long-term intakes of the ith individual, consumption and nonconsumption days combined. We now let $\mathbf{X}_i^+ = (X_{1,i}^+, \ldots, X_{q+p,i}^+)^{\mathrm{T}}$ denote the latent daily average long-term intakes of the ith individual on consumption days only. We then define $\widetilde{\mathbf{X}}_i = (\widetilde{X}_{1,i}, \ldots, \widetilde{X}_{2q+p,i})^{\mathrm{T}}$ as

$$\begin{aligned} \widetilde{X}_{\ell,i} &= h_\ell(X_{\ell,i}), \quad \text{for } \ell = 1, \ldots, q, \\ \widetilde{X}_{\ell,i} &= X_{\ell-q,i}^+, \quad \text{for } \ell = q + 1, \ldots, 2q, \\ \widetilde{X}_{\ell,i} &= X_{\ell-q,i}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p. \end{aligned}$$

Here, $h_\ell(\cdot)$ is an unknown function to be estimated from data. The reasons behind defining $\widetilde{\mathbf{X}}_i$ in this manner will be clear shortly.

For $i = 1, \ldots, n, j = 1, \ldots, m_i$, we let $\mathbf{U}_{i,j} = (U_{1,i,j}, \ldots, U_{2q+p,i,j})^{\mathrm{T}}$ and consider the model

$$\mathbf{W}_{i,j} = \widetilde{\mathbf{X}}_i + \mathbf{U}_{i,j}, \qquad \mathbb{E}(\mathbf{U}_{i,j} \mid \widetilde{\mathbf{X}}_i) = \mathbf{0}.$$

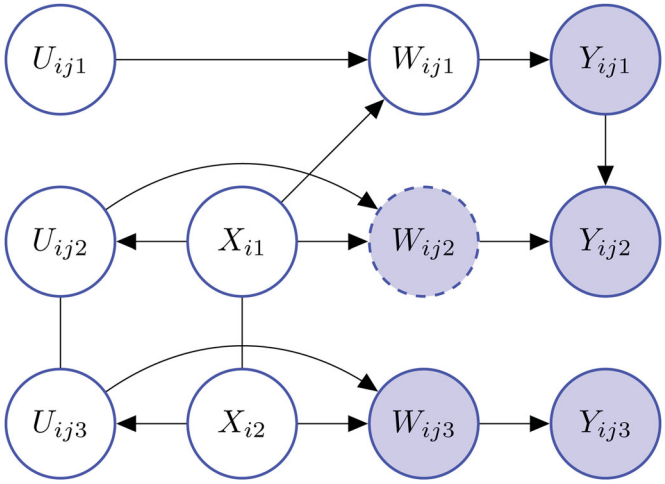

Figure 3. Graphical model depicting the dependency structure of the generative deconvolution model described in Section 2 for one episodically consumed component X_1 and one regularly consumed component X_2. The unfilled and shaded nodes with solid boundaries signify latent and observable variables, respectively. The filled node with dashed boundary may be observed on some occasions and latent on others.

For $\ell = 1, \ldots, q$, $W_{\ell,i,j}$ is always latent and the associated $U_{\ell,i,j}$ represents a pseudo-error that account for their within person daily variations. For $\ell = q+1, \ldots, 2q$, $U_{\ell,i,j}$ denotes the measurement error contaminating $W_{\ell,i,j}$ when it is observed and pseudo-errors when they are latent. Finally, for $\ell = 2q+1, \ldots, 2q+p$, $U_{\ell,i,j}$ denotes the measurement error contaminating $W_{\ell,i,j}$ which are always observed (Figure 3).

According to our model, for $\ell = 1, \ldots, q$, the probability of reporting a positive consumption on the ℓth episodic component, denoted henceforth as $P_\ell(X_{\ell,i})$, is obtained as $\Pr(Y_{\ell,i,j} = 1 \mid X_{\ell,i}) = \Pr(W_{\ell,i,j} > 0 \mid X_{\ell,i}) = \Pr\{U_{\ell,i,j} > -h_\ell(X_{\ell,i}) \mid X_{\ell,i}\}$. For $\ell = 1, \ldots, q$, we also have $\mathbb{E}(Y_{\ell+q,i,j} \mid Y_{\ell,i,j} = 1, \widetilde{X}_{\ell+q,i}) = \mathbb{E}(W_{\ell+q,i,j} \mid Y_{\ell,i,j} = 1, \widetilde{X}_{\ell+q,i}) = \mathbb{E}(W_{\ell+q,i,j} \mid \widetilde{X}_{\ell+q,i}) = \widetilde{X}_{\ell+q,i} = X^+_{\ell,i}$. The positive recalls $Y_{\ell,i,j}$'s and the $W_{\ell+q,i,j}$'s, latent or observed, are thus unbiased for the latent average long-term intakes of the episodic components on consumption days only. For $\ell = 1, \ldots, q$, the expectation $\mathbb{E}(Y_{\ell+q,i,j} \mid X_{\ell,i}, X^+_{\ell,i}) = \Pr(W_{\ell,i,j} > 0 \mid X_{\ell,i})\mathbb{E}(W_{\ell+q,i,j} \mid X^+_{\ell,i}) = P_\ell(X_{\ell,i})X^+_{\ell,i}$ then defines the overall long-term average intake, consumption and nonconsumption days combined. By definition, this is also $X_{\ell,i}$, giving us the relationship $X_{\ell,i} = P_\ell(X_{\ell,i})X^+_{\ell,i}$.

For regularly consumed components $\ell = q+1, \ldots, q+p$, of course, $\widetilde{X}_{\ell+q,i} = X^+_{\ell,i} = X_{\ell,i}$. The recalls in these cases are all observed and are unbiased for the latent long-term intakes as $\mathbb{E}(Y_{\ell+q,i,j} \mid \widetilde{X}_{\ell+q,i}) = \mathbb{E}(W_{\ell+q,i,j} \mid \widetilde{X}_{\ell+q,i}) = \widetilde{X}_{\ell+q,i}$.

Written in terms of the long-term average intakes $X_{\ell,i}$, the model thus becomes

$$
\begin{aligned}
W_{\ell,i,j} &= h_\ell(X_{\ell,i}) + U_{\ell,i,j}, \quad \text{for } \ell = 1, \ldots, q,\\
W_{\ell,i,j} &= X_{\ell-q,i}/P_{\ell-q}(X_{\ell-q,i}) + U_{\ell,i,j}, \quad \text{for } \ell = q+1, \ldots, 2q,\\
W_{\ell,i,j} &= X_{\ell-q,i} + U_{\ell,i,j}, \quad \text{for } \ell = 2q+1, \ldots, 2q+p. \qquad (2)
\end{aligned}
$$

This formulation now allows the problem to be reduced to that of modeling the components $f_{\mathbf{X}}$, $f_{\mathbf{U}|\widetilde{\mathbf{X}}}$, and $P_\ell(X_\ell)$ in a Bayesian hierarchical framework. It also simplifies the estimation of the distribution energy-adjusted intakes. We address this latter problem in Section 2.5.

The complex nature of our problem warranted the introduction of many different variables representing the many random variables of our model. For easy reference, these variables and a few others to be introduced shortly are listed in Table 2.

2.2. Modeling the Density $f_{\mathbf{X}}$

In this article, $f_{\mathbf{X}}$ is specified using a Gaussian copula density model

$$f_{\mathbf{X}}(\mathbf{X}) = |\mathbf{R}_{\mathbf{X}}|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\mathbf{Y}_{\mathbf{X}}^{\mathrm{T}}(\mathbf{R}_{\mathbf{X}}^{-1} - \mathbf{I}_{q+p})\mathbf{Y}_{\mathbf{X}}\right\} \prod_{\ell=1}^{q+p} f_{X,\ell}(X_\ell),$$

with $F_{X,\ell}(X_\ell) = \Phi(Y_{X,\ell})$ for all ℓ and $\mathbf{R}_{\mathbf{X}}$ is the correlation matrix of $\mathbf{X}$.

In initial attempts, we modeled the marginal densities $f_{X,\ell}$ as flexible mixtures of truncated normal kernels $\mathrm{TN}(\cdot \mid \mu, \sigma^2, [A, B])$ with location μ and scale σ and range restricted to the interval $[A, B]$. In multivariate applications such as ours, where the components represent similar variables and have highly overlapping supports, we can greatly reduce dimension and borrow information across different dietary components, by allowing the component specific parameters of the mixture models to be shared among the variables. We thus modeled the marginal densities as

$$
\begin{aligned}
f_{X,\ell}(X_\ell) &= \textstyle\sum_{k=1}^{K_X} \pi_{X,\ell,k}\, \mathrm{TN}(X_\ell \mid \mu_{X,k}, \sigma^2_{X,k}, [A_\ell, B_\ell]),\\
\boldsymbol{\pi}_{X,\ell} &\sim \mathrm{Dir}(\alpha_{X,\ell}/K_X, \ldots, \alpha_{X,\ell}/K_X),\\
\mu_{X,k} &\sim \mathrm{Normal}(\mu_{X,0}, \sigma^2_{X,0}),\\
\sigma^2_{X,k} &\sim \text{Inv-Ga}(a_{\sigma^2_{X,0}}, b_{\sigma^2_{X,0}}).
\end{aligned}
$$

The models for different components ℓ thus share the same atoms $(\mu_{X,k}, \sigma^2_{X,k})$ but with varying probability weights $\pi_{X,\ell,k}$.

Despite being specifically tailored to capture boundary discontinuities, in numerical experiments, we found the model to often produce steeply decaying and highly peaked estimates with underestimated (local) variance in these regions. After further investigations, we could attribute the issue to smoothness properties of such models, characterized by the variance components $\sigma^2_{X,k}$ which are estimated "locally" utilizing only the data points associated with the corresponding mixture components. For the episodic components, the scarcity of informative observations near the left boundaries often allows the sampled latent $X_{\ell,i}$'s to cluster away from these boundaries, resulting in the associated $\sigma^2_{X,k}$'s to be underestimated and hence the estimated densities to be peaked away from the boundaries. Setting informative lower bounds to the variance parameters solves the problem. Determining such bounds for the latent variables from their contaminated recalls, however, proved to be difficult.

For episodic components, we thus needed models that can accommodate local variations in shape but would also allow the smoothness to be learned from regions where more informative data points are available. To achieve this, we employed flexible penalized normalized mixtures of B-splines with smoothness inducing priors on the coefficients to model the densities of the episodic components. For the ℓth component, we partition the interval $[A_\ell, B_\ell]$ of interest into L_ℓ subintervals using knot points $A_\ell = t_{\ell,1} = \cdots = t_{\ell,d+1} < t_{\ell,d+2} < t_{\ell,d+3} < \cdots <$

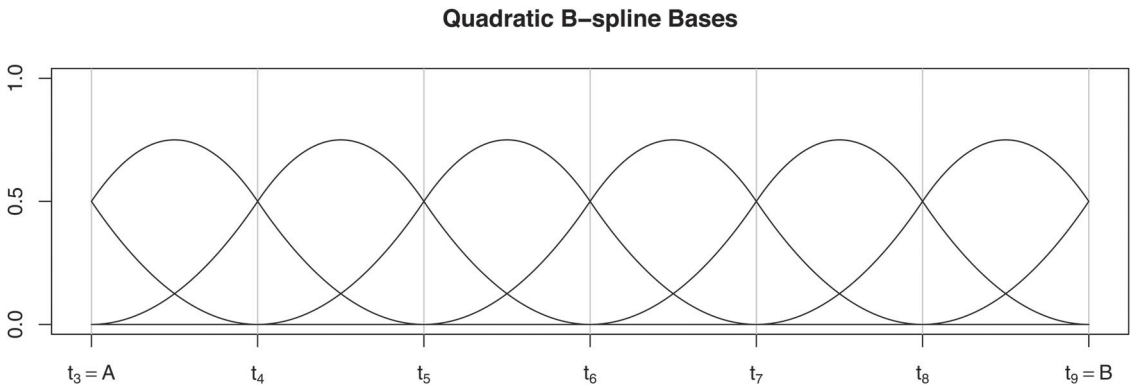

Figure 4. Plot of 9 quadratic ($d = 2$) B-splines on $[A, B]$ defined using 11 knot points that divide $[A, B]$ into $K = 6$ equal subintervals.

$t_{\ell,d+L_k} < t_{\ell,d+L_\ell+1} = \cdots = t_{\ell,2d+L_\ell+1} = B_\ell$. Using these knot points, $J_\ell = (d + L_\ell)$ B-spline bases of degree d, denoted by $\mathbf{B}_{d,\ell,J_\ell} = \{b_{d,\ell,1}, b_{d,\ell,2}, \ldots, b_{d,\ell,J_\ell}\}$, can be defined through a recursion relation (de Boor 2000, p. 90). See Figure 4 and Section S.2 in the supplementary materials. B-splines are nearly orthogonal and locally supported. For equidistant knot points with $\delta_\ell = (t_{\ell,J} - t_{\ell,J-1})$, the areas under these curves can be easily computed as

$$\delta_{\ell,j} = \int_{A_\ell}^{B_\ell} b_{2,\ell,j}(X)dX = \begin{cases} \delta_\ell/6 & \text{for } j = 1, J_\ell, \\ 5\delta_\ell/6 & \text{for } j = 2, J_\ell - 1, \\ \delta_\ell & \text{for } j = 3, \ldots, J_\ell - 2. \end{cases}$$

Mixtures of B-splines can therefore be easily normalized. A flexible model for the density functions is then obtained as

$$f_{X,\ell}(X_\ell) = \mathbf{B}_{d,\ell,J_\ell}(X_\ell) \exp(\boldsymbol{\xi}_\ell) \left\{ \sum_{m=1}^{J_\ell} \delta_{\ell,m} \exp(\xi_{\ell,m}) \right\}^{-1},$$

$$(\boldsymbol{\xi}_\ell \mid J_\ell, \sigma^2_{\xi,\ell}) \propto (2\pi\sigma^2_{\xi,\ell})^{-J_\ell/2} \exp\{-\boldsymbol{\xi}_\ell^{\mathrm{T}} \mathbf{P}_\ell \boldsymbol{\xi}_\ell/(2\sigma^2_{\xi,\ell})\},$$

$$\sigma^2_{\xi,\ell} \sim \text{Inv-Ga}(a_\xi, b_\xi).$$

Here, $\boldsymbol{\xi}_\ell = \{\xi_{\ell,1}, \xi_{\ell,2}, \ldots, \xi_{\ell,J_\ell}\}^{\mathrm{T}}$; $\exp(\boldsymbol{\xi}_\ell) = \{\exp(\xi_{\ell,1}), \exp(\xi_{\ell,2}), \ldots, \exp(\xi_{\ell,J_\ell})\}^{\mathrm{T}}$; and $\mathbf{P}_\ell = \mathbf{D}_\ell^{\mathrm{T}}\mathbf{D}_\ell$, where $\mathbf{D}_\ell$ is a $(J_\ell + 2) \times J_\ell$ matrix such that $\mathbf{D}_\ell\boldsymbol{\xi}_\ell$ computes the second-order differences in $\boldsymbol{\xi}_\ell$. The prior $p_0(\boldsymbol{\xi}_\ell \mid \sigma^2_{\xi,\ell})$ induces smoothness in the coefficients because it penalizes $\sum_{j=1}^{J_\ell}(\Delta^2\xi_{\ell,j})^2 = \boldsymbol{\xi}_\ell^{\mathrm{T}} P_\ell \boldsymbol{\xi}_\ell$, the sum of squares of the second-order differences in $\boldsymbol{\xi}_\ell$ (Eilers and Marx 1996). The parameters $\sigma^2_{\xi,\ell}$ play the role of smoothing parameters—the smaller the value of $\sigma^2_{\xi,\ell}$, the stronger the penalty and the smoother the associated variance function. The inverse-Gamma hyper-priors on $\sigma^2_{\xi,\ell}$ allow the data to influence the posterior smoothness and make the approach data adaptive. Importantly, the smoothness is now informed by data points across the entire range, resulting in vast improvements in the density estimates near the left boundaries.

For regularly consumed components with strictly positive recalls, we found mixtures of truncated normals to slightly outperform normalized mixtures of B-splines. This is also consistent with findings reported in Sarkar et al. (2014). For regularly consumed components, we thus still use mixtures of truncated normals with shared atoms. With densities smoothed out to zeros at the boundaries, truncations are not strictly needed for regularly consumed dietary components. We still retain the truncations to make our approach broadly applicable to other potential applications where boundary discontinuities may be present even when the recalls are all continuous.

Next, we consider the problem of modeling $\mathbf{R}_\mathbf{X}$. The problem of modeling correlation matrices has garnered some attention in the literature (Barnard, McCulloch, and Meng 2000; Liechty, Liechty, and Müller 2004; Pourahmadi and Wang 2015; Tsay and Pourahmadi 2017). Here, we adapt the model from Zhang, Midthune, et al. (2011) based on spherical coordinate representation of Cholesky factorizations that allows the involved parameters to be treated separately of each other, simplifying posterior computation while guaranteeing the resulting matrix to always be a valid correction matrix. We prove in Section S.3.1 in the supplementary materials that the converse is also true. That is, any correlation matrix can be represented in this form which establishes its nonparametric nature. We drop the subscript $\mathbf{X}$ for the rest of this subsection to keep the notation clean.

Let $\mathbf{V}^{(q+p)\times(q+p)}$ be a lower triangular matrix such that $\mathbf{R} = \mathbf{V}\mathbf{V}^{\mathrm{T}}$. The form of $\mathbf{V}$ is

$$\mathbf{V} = \begin{pmatrix} v_{1,1} & 0 & \ldots & 0 \\ v_{2,1} & v_{2,2} & \ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ v_{q+p,1} & v_{q+p,2} & \ldots & v_{q+p,q+p} \end{pmatrix}.$$

We have $r_{\ell,\ell'} = \sum_{k=1}^{\ell} v_{\ell,k}v_{\ell',k}$ for all $\ell \leq \ell'$. The restriction that $\mathbf{R}$ is a correlation matrix then implies $\sum_{k=1}^{\ell} v^2_{\ell,k} = 1$ for all $\ell = 1, \ldots, (q+p)$. The restrictions are satisfied by the following parameterization

$$\begin{aligned} &v_{1,1} = 1, \\ &v_{2,1} = b_1,\ v_{2,2} = \sqrt{1 - b_1^2}, \\ &v_{3,1} = b_2 \sin\theta_1,\ v_{3,2} = b_2 \cos\theta_1,\ v_{3,3} = \sqrt{1 - b_2^2}, \\ &v_{\ell,1} = b_{\ell-1} \sin\theta_{i_1(\ell)}, \\ &v_{\ell,k} = b_{\ell-1} \cos\theta_{i_1(\ell)} \cos\theta_{i_1(\ell)+1} \ldots \cos\theta_{i_1(\ell)+k-2} \\ &\qquad \sin\theta_{i_1(\ell)+k-1}, \quad \text{for } k = 2, \ldots, (\ell - 2), \\ &v_{\ell,\ell-1} = b_{\ell-1} \cos\theta_{i_1(\ell)} \cos\theta_{i_1(\ell)+1} \ldots \cos\theta_{i_2(\ell)-1} \cos\theta_{i_2(\ell)}, \\ &\qquad v_{\ell,\ell} = \sqrt{1 - b_{\ell-1}^2}, \end{aligned}$$

where $\ell = 4, \ldots, (q + p)$, $i_1(\ell) = 1 + \{1 + \cdots + (\ell - 3)\} = (\ell^2 - 5\ell + 8)/2$ and $i_2(\ell) = i_1(\ell) + (\ell - 3) = (\ell^2 - 3\ell + 2)/2$,

$|b_t| \leq 1, t = 1, \ldots, (q+p-1), |\theta_s| \leq \pi, s = 1, \ldots, i_2(q+p)$. The total number of parameters is $\{1+2+\cdots+(q+p-1)\} = (q+p)(q+p-1)/2$. We have $|\mathbf{R}| = |\mathbf{V}|^2 = \prod_{\ell=2}^{q+p} v_{\ell,\ell}^2 = \prod_{\ell=1}^{q+p-1}(1-b_\ell^2)$. The model for $\mathbf{R}$ is completed by assigning uniform priors on b_t's and θ_s's

$$b_t \sim \text{Unif}(-1,1), \quad \theta_s \sim \text{Unif}(-\pi,\pi).$$

Here, Unif(a,b) denotes a uniform distribution with support (a,b).

2.3. Modeling the Density $f_{\mathbf{U}|\widetilde{\mathbf{X}}}$

The reported intakes of the regularly consumed components exhibit strong conditional heteroscedasticity, so do the reported intakes of the episodic components, when consumed. To accommodate conditional heteroscedasticity, we let

$$\begin{aligned}\mathbf{U}_{i,j} &= \mathbf{S}(\widetilde{\mathbf{X}}_i)\boldsymbol{\epsilon}_{i,j}, \quad \text{with} \quad \mathbb{E}(\boldsymbol{\epsilon}_{i,j}) = \mathbf{0}, \quad \text{and}\\ \mathbf{S}(\widetilde{\mathbf{X}}_i) &= \text{diag}\{1,\ldots,1, s_{q+1}(\widetilde{X}_{q+1,i}), \ldots, s_{2q+p}(\widetilde{X}_{2q+p,i})\}.\end{aligned}$$

The above model implies that $\text{cov}(\mathbf{U}_{i,j} \mid \widetilde{\mathbf{X}}_i) = \mathbf{S}(\widetilde{\mathbf{X}}_i)\,\text{cov}(\boldsymbol{\epsilon}_{i,j})\,\mathbf{S}(\widetilde{\mathbf{X}}_i)$ and marginally $\text{var}(U_{\ell,i,j} \mid \widetilde{X}_{\ell,i}) = s_\ell^2(\widetilde{X}_{\ell,i})\,\text{var}(\epsilon_{\ell,i,j})$. Other features of the distribution of $\mathbf{U}$, including its shape and correlation structure, are derived from $f_{\boldsymbol{\epsilon}}$. The multiplicative structural assumption arises naturally for conditionally heteroscedastic multivariate measurement errors; the model also automatically accommodates multiplicative measurement errors $W_{\ell,i,j} = \widetilde{X}_{\ell,i}\widetilde{U}_{\ell,i,j}$ with $E(\widetilde{U}_{\ell,i,j}) = 1$ and $\widetilde{U}_{\ell,i,j}$ independent of $\widetilde{X}_{\ell,i}$ via a simple reformulation $W_{\ell,i,j} = \widetilde{X}_{\ell,i} + \widetilde{X}_{\ell,i}(\widetilde{U}_{\ell,i,j}-1) = \widetilde{X}_{\ell,i} + s(\widetilde{X}_{\ell,i})\epsilon_{\ell,i,j}$ with $s(\widetilde{X}_{\ell,i}) = \widetilde{X}_{\ell,i}$ and $\epsilon_{\ell,i,j} = (\widetilde{U}_{\ell,i,j}-1)$ (Sarkar et al. 2018).

As in Section 2.2, we use a Gaussian copula density model to specify the density $f_{\boldsymbol{\epsilon}}$ but the model now has to satisfy mean zero constraints. Specifically, we let

$$\begin{aligned}f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) = &\prod_{\ell=1}^{q} f_{\epsilon,\ell}(\epsilon_\ell) \times |\mathbf{R}_{\boldsymbol{\epsilon}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\mathbf{Y}_{\boldsymbol{\epsilon}}^{\mathrm{T}}(\mathbf{R}_{\boldsymbol{\epsilon}}^{-1} - \mathbf{I}_p)\mathbf{Y}_{\boldsymbol{\epsilon}}\right\}\\ &\prod_{\ell=q+1}^{2q+p} f_{\epsilon,\ell}(\epsilon_\ell), \quad \text{subject to } \mathbb{E}_{f_{\epsilon,\ell}}(\epsilon_\ell) = 0,\\ &\text{for } \ell = 1,\ldots,2q+p.\end{aligned}$$

Here, $F_{\epsilon,\ell}(\epsilon_\ell) = \Phi(Y_{\epsilon,\ell})$ for all ℓ. The first q components of $\boldsymbol{\epsilon}$ are independent of each other and also independent of the rest of the $q+p$ components. The latter $q+p$ components may be correlated with correlation matrix $\mathbf{R}_{\boldsymbol{\epsilon}}$.

The copula approach again allows us to use different models for the distributions of the pseudo-errors $f_{\epsilon,\ell}(\epsilon), \ell = 1,\ldots,q$, and the distributions of the actual scaled measurement errors $f_{\epsilon,\ell}(\epsilon), \ell = q+1,\ldots,2q+p$.

Here, we only model the correlation between different scaled error components $\epsilon_{\ell,i,j}, \epsilon_{\ell',i,j}$ for $\ell \neq \ell'$ but ignore the correlation between different sampling occasions $\epsilon_{\ell,i,j}, \epsilon_{\ell,i,j'}$ for $j \neq j'$. The correlation between $W_{\ell,i,j}, W_{\ell,i,j'}$ for $j \neq j'$ is thus explained entirely by their shared component $\widetilde{X}_{\ell,i}$. In post model fit correlation analysis with estimated scaled "residuals," presented in Figure S.9 in the supplementary materials, we found no real evidence that the errors $\epsilon_{\ell,i,j}, \epsilon_{\ell,i,j'}$ are significantly correlated for $j \neq j'$.

For $\ell = 1,\ldots,q$, we model the marginal densities $f_{\epsilon,\ell}$ as $f_{\epsilon,\ell}(\epsilon_\ell) = \text{Normal}(\epsilon_\ell \mid 0,1)$. This implies a probit model for the probabilities of consumptions $P_\ell(X_\ell) = \Pr\{U_\ell > -h_\ell(X_\ell)\} = \Phi\{h(X_\ell)\}$. Flexibility of this probability model thus depends on the choice of $h_\ell(X_\ell)$. We discuss this issue in Section 2.4.

For $\ell = q+1,\ldots,2q+p$, we model the marginal densities $f_{\epsilon,\ell}(\epsilon)$ using an adaptation of the moment restricted model in Sarkar et al. (2014) but with shared atoms as

$$\begin{aligned}f_{\epsilon,\ell}(\epsilon_\ell) &= \textstyle\sum_{k=1}^{K_\epsilon} \pi_{\epsilon,\ell,k} f_{c\epsilon}(\epsilon_\ell \mid p_{\epsilon,k}, \widetilde{\mu}_{\epsilon,k}, \sigma^2_{\epsilon,k,1}, \sigma^2_{\epsilon,k,2}),\\ \boldsymbol{\pi}_{\epsilon,\ell} &\sim \text{Dir}(\alpha_{\epsilon,\ell}/K_\epsilon, \ldots, \alpha_{\epsilon,\ell}/K_\epsilon), (p_{\epsilon,k}, \widetilde{\mu}_{\epsilon,k}, \sigma^2_{\epsilon,k,1},\\ \sigma^2_{\epsilon,k,2}) &\sim \text{Unif}(0,1)\ \text{Normal}(0, \sigma^2_{\epsilon,\widetilde{\mu}})\ \text{IG}(a_\epsilon, b_\epsilon)\ \text{IG}(a_\epsilon, b_\epsilon),\end{aligned}$$

where $f_{c\epsilon}(\epsilon \mid p, \widetilde{\mu}, \sigma_1^2, \sigma_2^2) = \{p\ \text{Normal}(\epsilon \mid \mu_1, \sigma_1^2) + (1-p)\ \text{Normal}(\epsilon \mid \mu_2, \sigma_2^2)\}$, with $\mu_1 = c_1\widetilde{\mu}, \mu_2 = c_2\widetilde{\mu}$, $c_1 = (1-p)/\{p^2+(1-p)^2\}^{1/2}$ and $c_2 = -p/\{p^2+(1-p)^2\}^{1/2}$. The zero mean constraint on the errors is satisfied, since $p\mu_1 + (1-p)\mu_2 = \{pc_1 + (1-p)c_2\}\widetilde{\mu} = 0$. Normal densities are included as special cases with $(p,\widetilde{\mu}) = (0.5, 0)$ or $(0,0)$ or $(1,0)$. Symmetric component densities are included as special cases when $p = 0.5$ or $\widetilde{\mu} = 0$. Specification of the prior for f_ϵ is completed assuming non-informative priors for $(p, \widetilde{\mu}, \sigma_1^2, \sigma_2^2)$. Here Unif$(\ell, u)$ denotes a uniform distribution on the interval $[\ell, u]$.

As in the case of $\mathbf{R}_{\mathbf{X}}$, we assume $\mathbf{R}_{\boldsymbol{\epsilon}}^{(q+p)\times(q+p)} = ((r_{\boldsymbol{\epsilon},\ell,\ell'})) = \mathbf{V}_{\boldsymbol{\epsilon}}\mathbf{V}_{\boldsymbol{\epsilon}}^{\mathrm{T}}$ and parameterize the elements of $\mathbf{V}_{\boldsymbol{\epsilon}}$ using spherical coordinates. We assign uniform priors on $b_{\boldsymbol{\epsilon},t}, t = 1,\ldots,(q+p-1)$ and $\theta_{\boldsymbol{\epsilon},s}, s = 1,\ldots,i_2(q+p)$

$$b_{\boldsymbol{\epsilon},t} \sim \text{Unif}(-1,1), \quad \theta_{\boldsymbol{\epsilon},s} \sim \text{Unif}(-\pi,\pi).$$

Finally, for $\ell = q+1,\ldots,2q+p$, we model the variance functions $v_\ell(\widetilde{X}_\ell) = s_\ell^2(\widetilde{X}_\ell)$ by flexible penalized mixtures of B-splines with smoothness inducing priors on the coefficients as in Staudenmayer, Ruppert, and Buonaccorsi (2008) as

$$\begin{aligned}v_\ell(\widetilde{X}_\ell) &= s_\ell^2(\widetilde{X}_\ell) = \textstyle\sum_{j=1}^{J_\ell} b_{d,\ell,j}(\widetilde{X}_\ell)\exp(\vartheta_{\ell,j})\\ &= \mathbf{B}_{d,\ell,J_\ell}(\widetilde{X}_\ell)\exp(\boldsymbol{\vartheta}_\ell),\\ (\boldsymbol{\vartheta}_\ell \mid J_\ell, \sigma^2_{\vartheta,\ell}) &\propto (2\pi\sigma^2_{\vartheta,\ell})^{-J_\ell/2}\exp\{-\boldsymbol{\vartheta}_\ell^{\mathrm{T}}\mathbf{P}_\ell\boldsymbol{\vartheta}_\ell/(2\sigma^2_{\vartheta,\ell})\},\\ \sigma^2_{\vartheta,\ell} &\sim \text{Inv-Ga}(a_\vartheta, b_\vartheta).\end{aligned}$$

As before, the parameters $\sigma^2_{\vartheta,\ell}$ play the role of smoothing parameter, and the inverse-Gamma hyper-priors allow them to be learned from the data themselves.

2.4. Modeling the Consumption Probabilities $P_\ell(X_\ell)$

We recall that, according to our model, the probability of reporting positive consumptions by an individual with long-term average intake X_ℓ is given by

$$P_\ell(X_\ell) = \Pr\{U_\ell > -h_\ell(X_\ell) \mid X_\ell\} = \Phi\{h_\ell(X_\ell)\}.$$

We model $h_\ell(X_\ell)$ using flexible mixtures of B-splines again as

$$\begin{aligned}&h_\ell(X_\ell) = \textstyle\sum_{j=1}^{J_\ell} b_{d,\ell,j}(X_\ell)\beta_{\ell,j} = \mathbf{B}_{d,\ell,J_\ell}(X_\ell)\boldsymbol{\beta}_\ell,\\ &(\boldsymbol{\beta}_\ell \mid J_\ell, \sigma^2_{\beta,\ell}, \boldsymbol{\mu}_{\beta,\ell}, \boldsymbol{\Sigma}_{\beta,\ell}) \propto (2\pi\sigma^2_{\beta,\ell})^{-J_\ell/2}\end{aligned}$$

$$\exp\{-\boldsymbol{\beta}_\ell^{\mathrm{T}}\mathbf{P}_\ell\boldsymbol{\beta}_\ell/(2\sigma^2_{\beta,\ell})\}\,\mathrm{MVN}_{J_\ell}(\boldsymbol{\beta}_\ell \mid \boldsymbol{\mu}_{\beta,\ell,0}, \boldsymbol{\Sigma}_{\beta,\ell,0}),$$
$$\sigma^2_{\beta,\ell} \sim \text{Inv-Ga}(a_\beta, b_\beta).$$

The flexibility of $h_\ell(X_\ell)$ compensates for the parametric nature of the probit link, making the model $P_\ell(X_\ell)$ robust.

The right panels of Figure 2 suggest that as X_ℓ increases, the probability of reporting a positive consumption also increases on average. We model this flexibly as $\Phi\{h_\ell(X_\ell)\}$. It is certainly possible that two individuals have (nearly) the same long-term average intakes, even though one of them consumes less often than the other but consumes larger amounts. One could hope that additional subject-specific random effects terms would help capture this heterogeneity. It is, however, not clear that such models would be identifiable in the first place. To see this, consider adding random effects $R_{\ell,i}$ to model (2). Letting $h_\ell(X_{\ell,i}) = X_{\ell,i}$ for simplicity, we then obtain $W_{\ell,i,j} = X_{\ell,i} + R_{\ell,i} + U_{\ell,i,j}$, $\ell = 1,\ldots,q$. With only the standard zero mean assumption on the distribution of the random effects, it is impossible to separately nonparametrically identify the distributions of $X_{\ell,i}$ and $R_{\ell,i}$ in this model.

2.5. Modeling Energy-Adjusted Intakes

We now consider the problem of modeling the distribution of energy-adjusted long-term intakes. We now denote $\mathbf{X} = (X_1,\ldots,X_{q+p})^{\mathrm{T}} = (X_1,\ldots,X_J)^{\mathrm{T}}$ with $J = q+p$ and $X_J = X_{q+p}$ representing the energy intake. We are interested in the distribution of the intakes normalized by energy, that is, the distribution of $\mathbf{Z} = (X_1/X_J,\ldots,X_{J-1}/X_J)$. The joint distribution of $\mathbf{Z}$ is then straightforwardly obtained as

$$f_{\mathbf{Z}}(\mathbf{Z}) = \int X_J^J f_{\mathbf{X}}(Z_1X_J,\ldots,Z_{J-1}X_J,X_J)dX_J.$$

The marginal distribution of any Z_ℓ is likewise obtained as

$$f_{Z,\ell}(Z_\ell) = \int X_J f_{X_\ell,X_J}(Z_\ell X_J, X_J)dX_J.$$

These are integrals of single variables and can thus be easily numerically evaluated.

2.6. Model Flexibility

For most practical purposes, including our motivating applications, our models for the densities of interest $f_{X,\ell}$, the densities of the scaled errors $f_{\ell,\epsilon}$, the variance functions s^2_ℓ, and the probabilities of consumptions $P_\ell(X_\ell)$ are all highly flexible whenever sufficiently large numbers of B-spline bases and mixture components are allowed. Adapting similar results from Sarkar et al. (2018), formal statements and proofs establishing theoretical flexibility of these model components can be easily formulated using known results for B-splines and mixture models. Our model for the correlation matrices $\mathbf{R}$ is also nonparametric. A formal proof is provided in Section S.3.1 in the supplementary materials. The only real parametric component of our model is thus the Gaussian copula. Extending the model to other elliptical classes, like the multivariate t, would be conceptually straightforward. It is, however, often difficult to distinguish between such classes even in much simpler low dimensional measurement error-free scenarios (dos Santos Silva and Lopes 2008). The problem only gets an order of magnitude more difficult when the variables whose densities are being modeled using copulas are all latent. Since the number of parameters in elliptical copulas increases only quadratically with dimension, they also scale well to higher dimensions. It is thus also not clear if other stylized copula classes could be any useful in nutritional epidemiology datasets like ours. Exploration of these issues will be pursued elsewhere.

2.7. Model Identifiability

In the following, we investigate identifiability of our model. For notational simplicity, we drop the subscript i and consider for $j = 1,\ldots,m$, $\mathbf{Y}_j = (Y_{1,j},\ldots,Y_{2q+p,j})^{\mathrm{T}}$, and similarly $\mathbf{W}_j, \mathbf{U}_j, \widetilde{\mathbf{X}}$ and $\mathbf{X}$. Then our proposed hierarchical model can be written as

$$\mathbf{Y}_j = \psi(\mathbf{W}_j), \quad \mathbf{W}_j = \widetilde{\mathbf{X}} + \mathbf{U}_j, \quad \mathbb{E}(\mathbf{U}_j \mid \widetilde{\mathbf{X}}) = \mathbf{0}, \quad \widetilde{\mathbf{X}} = \phi(\mathbf{X}),$$

where the functions $\psi(\cdot) : \mathbb{R}^{2q+p} \to \mathbb{R}^{2q+p}$ and $\phi(\cdot) : \mathbb{R}^{2q+p} \to \mathbb{R}^{2q+p}$ are easily identified from models (1) and (2). Specifically, $\phi(\cdot)$ is given by

$$\begin{aligned} \widetilde{X}_\ell &= h_\ell(X_\ell), && \text{for } \ell = 1,\ldots,q, \\ \widetilde{X}_\ell &= X_{\ell-q}/P_{\ell-q}(X_{\ell-q}), && \text{for } \ell = q+1,\ldots,2q, \\ \widetilde{X}_\ell &= X_{\ell-q}, && \text{for } \ell = 2q+1,\ldots,2q+p, \end{aligned} \tag{3}$$

where, for $\ell = 1,\ldots,q$, $P_\ell(X_\ell) = P(W_{\ell,j} > 0 | X_\ell) = \Phi\{h_\ell(X_\ell)\}$ for some arbitrary functions $h_\ell(\cdot) : \mathbb{R} \to \mathbb{R}$.

We state the basic assumptions needed for identifiability and our main result on identifiability below. The proof is deferred to Section S.3.2 in the supplementary materials.

Assumption 1.

(A1) The number of replicates $m \geq 3$.

(A2) $\mathbf{U}_j \mid \widetilde{\mathbf{X}} \overset{d}{=} \mathbf{S}(\widetilde{\mathbf{X}})\boldsymbol{\epsilon}_j$, $\boldsymbol{\epsilon}_j \sim f_{\boldsymbol{\epsilon}}, j = 1,2,3$, where $f_{\boldsymbol{\epsilon}}$ has a Fourier transform that is non-vanishing everywhere.

Observe that (A2) includes the homoscedastic case, that is, when $s_\ell(X_\ell)$ is a constant function of X_ℓ.

Theorem 1. Under (A1) and (A2), given the observed density $f_{\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Y}_3}$, the equation

$$f_{\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Y}_3}(\mathbf{Y}_1,\mathbf{Y}_2,\mathbf{Y}_3) = \int f_{\mathbf{Y}_1|\widetilde{\mathbf{X}}}(\mathbf{Y}_1 \mid \widetilde{\mathbf{X}}) f_{\mathbf{Y}_2|\widetilde{\mathbf{X}}}(\mathbf{Y}_2 \mid \widetilde{\mathbf{X}}) f_{\mathbf{Y}_3|\widetilde{\mathbf{X}}}(\mathbf{Y}_3 \mid \widetilde{\mathbf{X}}) f_{\widetilde{\mathbf{X}}}(\widetilde{\mathbf{X}}) d\widetilde{\mathbf{X}}$$

admits a unique solution for $f_{\mathbf{Y}_j|\widetilde{\mathbf{X}}}(\mathbf{Y}_j \mid \widetilde{\mathbf{X}})$ for $j = 1,\ldots,3$ and $f_{\widetilde{\mathbf{X}}}(\widetilde{\mathbf{X}})$. Furthermore, if $\mathbf{X}$ and $\widetilde{\mathbf{X}}$ are related by (3), then $f_{\mathbf{X}}(\mathbf{X})$ is uniquely identified from $f_{\mathbf{Y}_j|\widetilde{\mathbf{X}}}(\mathbf{Y}_j \mid \widetilde{\mathbf{X}})$ for $j = 1,\ldots,3$ and $f_{\widetilde{\mathbf{X}}}(\widetilde{\mathbf{X}})$.

In practice, for identifiability, we require $m_i \geq 3$ recalls for at least some values of i. As long as this condition is satisfied, missing values in recall data can be simply ignored. For our motivating EATS dataset, we have $m_i = 4$ for all i with no missing recalls. So the conditions are easily satisfied.

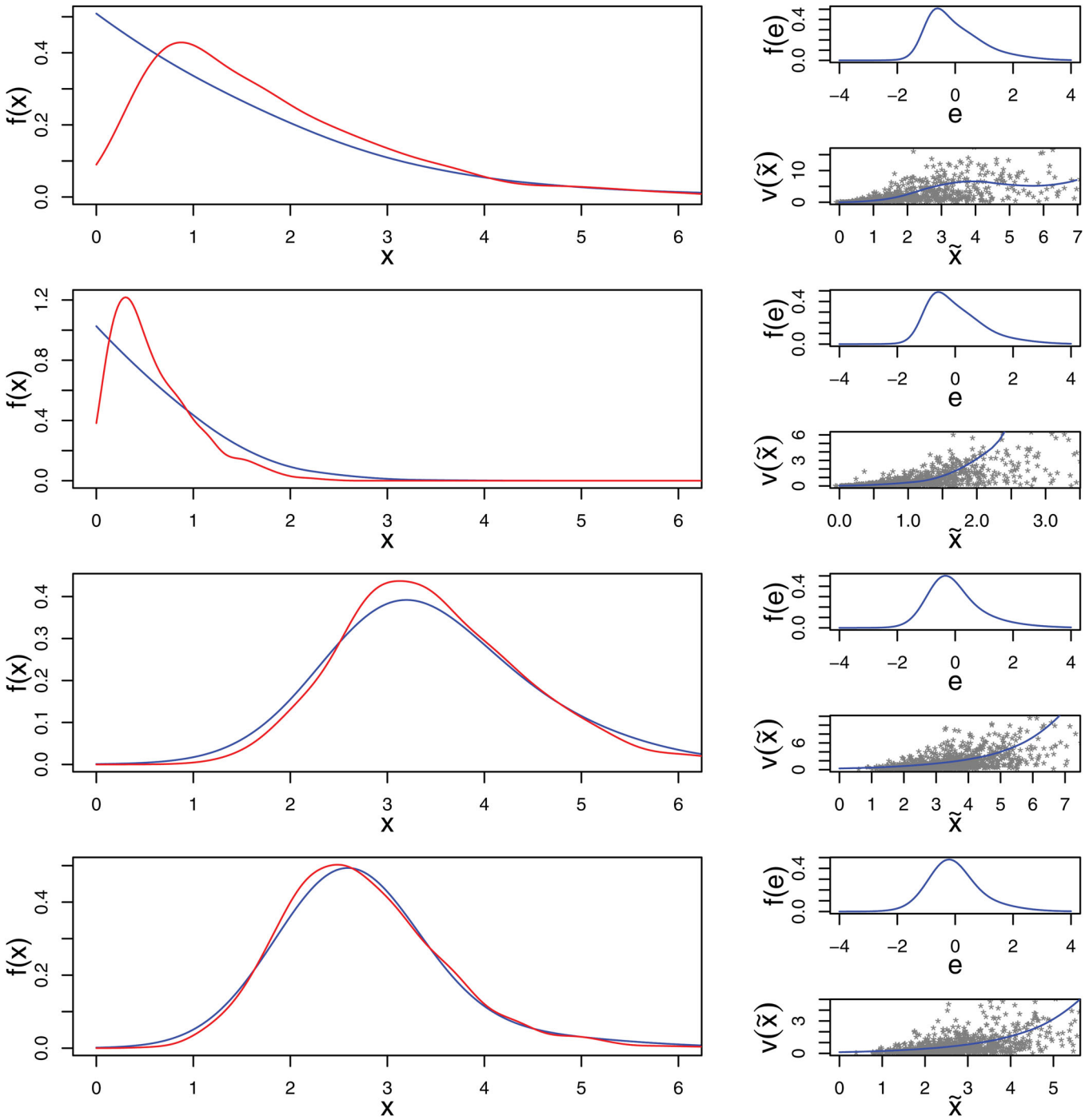

Figure 5. Results for the EATS datasets with sample size $n = 965$, $q = 2$ episodic components, milk and whole grains, and $p = 2$ regular components, sodium and energy, each subject having $m_i = 4$ replicates. From top to bottom, the left panels show the estimated densities $f_{X,\ell}(X_\ell)$ of milk and whole grains, sodium, and energy, respectively, obtained by our method (in blue) and the method of Zhang, Midthune, et al. (2011) (in red). The right panels show the associated distributions of the scaled errors $f_{\epsilon,q+\ell}(\epsilon_{q+\ell})$ and the associated variance functions $v_\ell(\widetilde{X}_\ell) = s^2_\ell(\widetilde{X}_\ell)$, estimated by our method.

3. Applications in Nutritional Epidemiology

In this section, we discuss the results of our method applied to the EATS dataset. Specifically, we consider the problem of estimating the distributions of long-term average daily intakes of two episodic components—milk and whole grains, and two regular components—sodium and energy. The surrogates for milk and whole grains, we recall, had approximately 21% and 37% exact zeros.

Figure 5 shows the estimated marginal densities $f_{X,\ell}$ obtained by our method and the method of Zhang, Midthune, et al. (2011). For sodium and energy, there is general agreement between the estimates obtained by our method and the method of Zhang, Midthune, et al. (2011). For the episodic components milk and whole grains, on the other hand, the estimated densities look very different, especially near the left boundary. Our method shows these densities to continually increase as we approach zero from right, as is expected from Figure 2.

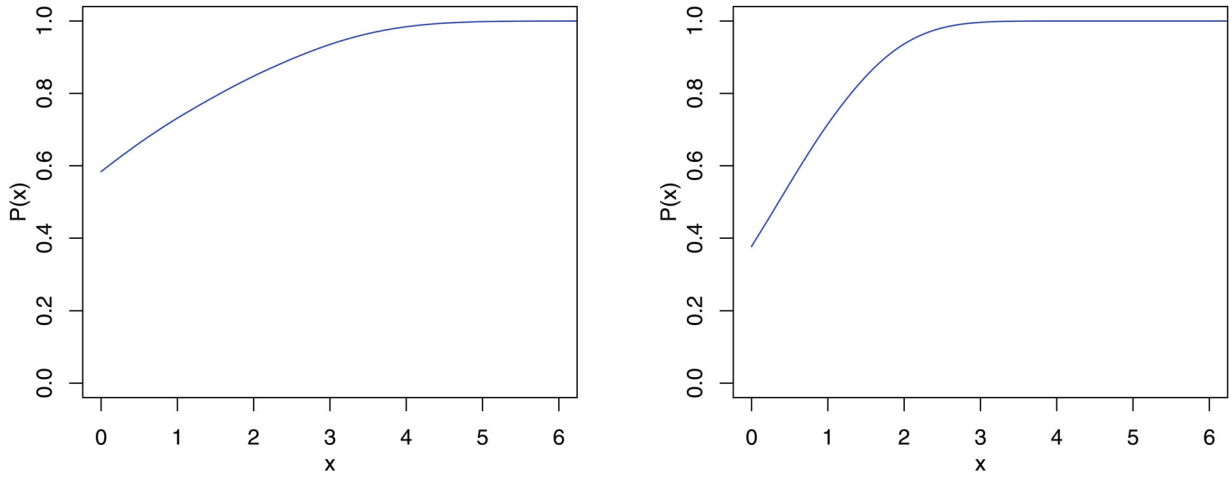

Figure 6. Results for the EATS datasets with sample size $n = 965$, $q = 2$ episodic components, milk and whole grains, and $p = 2$ regular components, sodium and energy, each subject having $m_i = 4$ replicates. The estimated probabilities of reporting positive consumption $P_\ell(X_\ell)$ for the episodic components milk (left panel) and whole grains (right panel), estimated by our method.

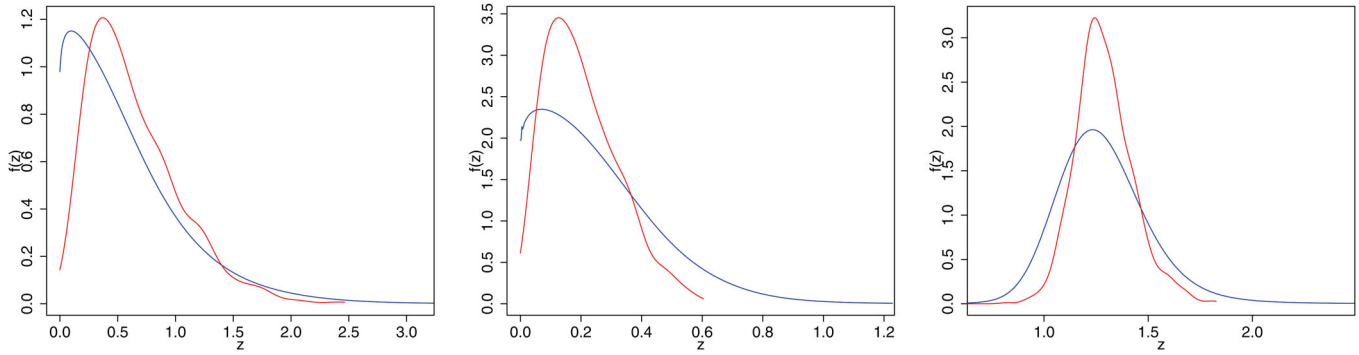

Figure 7. Results for the EATS datasets with sample size $n = 965$, $q = 2$ episodic components, milk and whole grains, and $p = 2$ regular components, sodium and energy, each subject having $m_i = 4$ replicates. From left to right, the estimated distributions of normalized intakes of milk, whole grains and sodium, normalized by total energy, estimated by our method (in blue) and by the method of Zhang, Midthune, et al. (2011) (in red).

Consistent with Figure 2, compared to milk, the distribution of whole grains is also more concentrated near zero. The estimates produced by Zhang, Midthune, et al. (2011), on the other hand, dip near zero, as was also observed in simulation scenarios.

The right panels in Figure 5 show the estimates of the densities of scaled measurement errors $f_{\epsilon,q+\ell}(\epsilon_{q+\ell})$ and the estimates of the variance functions $s^2_\ell(\widetilde{X}_\ell)$. The estimated $f_{\epsilon,q+\ell}$'s are positively skewed for all components. And, as expected from Figures 1 and 2, the estimated s^2_ℓ's show strong patterns of conditional heteroscedasticity for all components. For the episodic components, our method also provides estimates of the probabilities of reporting positive consumptions which are shown in Figure 6. The recalls for whole grains have more zeros than the recalls for milk. Its distribution is also more concentrated near zero. The probability of reporting positive consumptions for whole grains thus increases more rapidly as its true daily average intake increases.

Figure 7 shows the distributions of normalized intakes obtained by our method and the method of Zhang, Midthune, et al. (2011). The estimates look very different, including the one for the regular component sodium. Our method provides more realistic estimates of the distribution of normalized intakes that are more concentrated near zero but are more widely spread.

Figure S.7 in the supplementary materials shows the estimated bivariate marginals for produced by our method and the method of Zhang, Midthune, et al. (2011). Figure S.8 in the supplementary materials additionally illustrates how the redundant mixture components become empty after reaching steady states in our MCMC based implementation. Figure S.8 also shows how in practice the mixture component specific parameters get shared across different dimensions in our models with shared parameters for the marginal densities.

More formal model comparisons for real data, using component wise log pseudo marginal likelihoods (Geisser and Eddy 1979) and widely applicable information criteria (WAIC) (Watanabe 2010), are discussed in Section S.9 in the supplementary materials.

4. Discussion

4.1. Summary

In this article, we considered the problem of multivariate density deconvolution when replicated proxies are available but, complicating the challenges, the proxies also include exact zeros for some of the components. The problem is important in nutritional epidemiology for estimating long-term intakes of

episodically consumed dietary components. We developed a novel copula based deconvolution approach that focuses on the marginals first and then models the dependence among the components to build the joint densities, allowing us to adopt different modeling strategies for different marginal distributions which proved crucial in accommodating important features of our motivating datasets. In contrast to previous approaches of modeling episodically consumed dietary components, our novel Bayesian hierarchical modeling framework allows us to model the distributions of interest more directly, resulting in vast improvements in empirical performances while also providing estimates of quantities of secondary interest, including probabilities of reporting nonconsumptions, measurement errors' conditional variability, etc.

4.2. Other Potential Applications

Applications of the multivariate deconvolution approach developed here are not limited to zero-inflated data only but also naturally include data with strictly continuous recalls, as was shown in the simulations. Advanced multivariate deconvolution methods are also needed to correct for measurement errors in regression settings when multiple error contaminated predictors are needed to be included in the model.

4.3. Methodological Extensions

Other methodological extensions and subjects of ongoing research include inclusion of associated exactly measured covariates like age, sex, etc. that can potentially influence the consumption patterns, establishing theoretical convergence guarantees for the posterior, accommodation of dietary components which, unlike regular or episodic components, are never consumed by a percentage of the population, accommodation of subject specific survey weights, exploration of non-Gaussian copula classes, inclusion of additional information provided by food frequency questionnaires, etc.

4.4. HEI Index

Aside being of independent interest, episodic dietary components also contribute to the Healthy Eating Index (HEI, *https://www.cnpp.usda.gov/healthyeatingindex*), a performance measure developed by the US Department of Agriculture (USDA) to assess and promote healthy diets (Guenther, Reedy, and Krebs-Smith 2008; Krebs-Smith et al. 2018). The index is based on 13 energy adjusted dietary components, as many as 8 of which are episodic, and is currently calculated using the NCI method discussed in Section S.4 in the supplementary materials. The methodology developed in this article provides a much more sophisticated framework for modeling the HEI index and makes up an important component of our ongoing research.

Supplementary Materials

The supplementary materials present a brief review of copula and explicit formula of quadratic B-splines for easy reference. The supplementary materials also provide a detailed comparison of our method with previous approaches to zero-inflated data and numerical experiments comparing it with its main competitor. The supplementary materials additionally detail the choice of hyper-parameters and the MCMC algorithm used to sample from the posterior, presents some additional figures, and the results of some additional numerical experiments. R programs implementing the deconvolution methods developed in this article are included in the supplementary materials. The EATS data analyzed in Section 3 can be accessed from National Cancer Institute by arranging a Material Transfer Agreement. A simulated dataset, simulated according to one of the designs described in Section S.6 in the supplementary material, and a "readme" file providing additional details are also included in the supplementary material.

Acknowledgments

We thank the University of Texas Advanced Computing Center (TACC) for providing computing resources that contributed to the research reported here.

Funding

Pati's research was supported in part by NSF grant DMS1613156. Mallick's research was supported by grant R01CA194391 from the National Cancer Institute and grant CCF-1934904 from the NSF. Carroll's research was supported in part by grant U01-CA057030 from the National Cancer Institute.

References

Barnard, J., McCulloch, R., and Meng, X.-L. (2000), "Modeling Covariance Matrices in Terms of Standard Deviations and Correlations, With Application to Shrinkage," *Statistica Sinica*, 10, 1281–1311. [7]

Box, G. E., and Cox, D. R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B, 26, 211–252. [3]

Buonaccorsi, J. P. (2010), *Measurement Error: Models, Methods, and Applications*, Chapman & Hall/CRC Interdisciplinary Statistics Series, Boca Raton, FL: CRC Press. [2]

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

de Boor, C. (2000), *A Practical Guide to Splines*, New York: Springer. [7]

dos Santos Silva, R., and Lopes, H. F. (2008), "Copula, Marginal Distributions and Model Selection: A Bayesian Note," *Statistics and Computing*, 18, 313–320. [9]

Eilers, P. H. C., and Marx, B. D. (1996), "Flexible Smoothing With *B*-Splines and Penalties," *Statistical Science*, 11, 89–102. [7]

Escobar, M. D., and West, M. (1995), "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588. [3]

Frühwirth-Schnatter, S. (2006), *Finite Mixture and Markov Switching Models*, New York: Springer. [3]

Geisser, S., and Eddy, W. F. (1979), "A Predictive Approach to Model Selection," *Journal of the American Statistical Association*, 74, 153–160. [11]

Guenther, P. M., Reedy, J., and Krebs-Smith, S. M. (2008), "Development of the Healthy Eating Index-2005," *Journal of the American Dietetic Association*, 108, 1896–1901. [12]

Joe, H. (2015), *Dependence Modeling With Copulas*, Boca Raton, FL: CRC Press. [5]

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. [3]

Krebs-Smith, S. M., Pannucci, T. E., Subar, A. F., Kirkpatrick, S. I., Lerman, J. L., Tooze, J. A., Wilson, M. M., and Reedy, J. (2018), "Update of the Healthy Eating Index: HEI-2015," *Journal of the Academy of Nutrition and Dietetics*, 118, 1591–1602. [12]

Liechty, J. C., Liechty, M. W., and Müller, P. (2004), "Bayesian Correlation Estimation," *Biometrika*, 91, 1–14. [7]

Nelsen, R. B. (2007), *An Introduction to Copulas*, New York: Springer. [5]

Pourahmadi, M., and Wang, X. (2015), "Distribution of Random Correlation Matrices: Hyperspherical Parameterization of the Cholesky Factor," *Statistics & Probability Letters*, 106, 5–12. [7]

Sarkar, A., Mallick, B. K., Staudenmayer, J., Pati, D., and Carroll, R. J. (2014), "Bayesian Semiparametric Density Deconvolution in the Presence of Conditionally Heteroscedastic Measurement Errors," *Journal of Computational and Graphical Statistics*, 24, 1101–1125. [3,4,5,7,8]

Sarkar, A., Pati, D., Chakraborty, A., Mallick, B. K., and Carroll, R. J. (2018), "Bayesian Semiparametric Multivariate Density Deconvolution," *Journal of the American Statistical Association*, 113, 401–416. [3,4,5,8,9]

Shemyakin, A., and Kniazev, A. (2017), *Introduction to Bayesian Estimation and Copula Models of Dependence*, Hoboken, NJ: Wiley. [5]

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. R. (2008), "Density Estimation in the Presence of Heteroscedastic Measurement Error," *Journal of the American Statistical Association*, 103, 726–736. [2,3,8]

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," *American Journal of Epidemiology*, 154, 1089–1099. [2]

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002), "Analysis of Repeated Measures Data With Clumping at Zero," *Statistical Methods in Medical Research*, 11, 341–355. [3]

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," *Journal of the American Dietetic Association*, 106, 1575–1587. [3]

Tsay, R. S., and Pourahmadi, M. (2017), "Modelling Structured Correlation Matrices," *Biometrika*, 104, 237–242. [7]

Watanabe, S. (2010), "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory," *Journal of Machine Learning Research*, 11, 3571–3594. [11]

Zhang, S., Krebs-Smith, S. M., Midthune, D., Pérez, A., Buckman, D. W., Kipnis, V., Freedman, L. S., Dodd, K. W., and Carroll, R. J. (2011), "Fitting a Bivariate Measurement Error Model for Episodically Consumed Dietary Components," *International Journal of Biostatistics*, 7, 1–17. [3]

Zhang, S., Midthune, D., Guenther, P. M., Krebs-Smith, S. M., Kipnis, V., Dodd, K. W., Buckman, D. W., Tooze, J. A., Freedman, L., and Carroll, R. J. (2011), "A New Multivariate Measurement Error Model With Zero-Inflated Dietary Data, and Its Application to Dietary Assessment," *Annals of Applied Statistics*, 5, 1456–1487. [3,4,5,7,10,11]