The International Journal of Biostatistics

Volume 7, *Issue* 1 2011 *Article* 1

Fitting a Bivariate Measurement Error Model for Episodically Consumed Dietary Components

Saijuan Zhang, *Texas A&M University*
Susan M. Krebs-Smith, *National Cancer Institute*
Douglas Midthune, *National Cancer Institute*
Adriana Perez, *University of Texas School of Public Health*
Dennis W. Buckman, *Information Management Services, Inc.*
Victor Kipnis, *National Cancer Institute*
Laurence S. Freedman, *Gertner Institute for Epidemiology and Public Health Research*
Kevin W. Dodd, *National Cancer Institute*
Raymond J. Carroll, *Texas A&M University*

Recommended Citation:
Zhang, Saijuan; Krebs-Smith, Susan M.; Midthune, Douglas; Perez, Adriana; Buckman, Dennis W.; Kipnis, Victor; Freedman, Laurence S.; Dodd, Kevin W.; and Carroll, Raymond J. (2011) "Fitting a Bivariate Measurement Error Model for Episodically Consumed Dietary Components," *The International Journal of Biostatistics*: Vol. 7: Iss. 1, Article 1.
DOI: 10.2202/1557-4679.1267

Fitting a Bivariate Measurement Error Model for Episodically Consumed Dietary Components

Saijuan Zhang, Susan M. Krebs-Smith, Douglas Midthune, Adriana Perez, Dennis W. Buckman, Victor Kipnis, Laurence S. Freedman, Kevin W. Dodd, and Raymond J. Carroll

Abstract

There has been great public health interest in estimating usual, i.e., long-term average, intake of episodically consumed dietary components that are not consumed daily by everyone, e.g., fish, red meat and whole grains. Short-term measurements of episodically consumed dietary components have zero-inflated skewed distributions. So-called two-part models have been developed for such data in order to correct for measurement error due to within-person variation and to estimate the distribution of usual intake of the dietary component in the univariate case. However, there is arguably much greater public health interest in the usual intake of an episodically consumed dietary component adjusted for energy (caloric) intake, e.g., ounces of whole grains per 1000 kilo-calories, which reflects usual dietary composition and adjusts for different total amounts of caloric intake. Because of this public health interest, it is important to have models to fit such data, and it is important that the model-fitting methods can be applied to all episodically consumed dietary components.

We have recently developed a nonlinear mixed effects model (Kipnis, et al., 2010), and have fit it by maximum likelihood using nonlinear mixed effects programs and methodology (the SAS NLMIXED procedure). Maximum likelihood fitting of such a nonlinear mixed model is generally slow because of 3-dimensional adaptive Gaussian quadrature, and there are times when the programs either fail to converge or converge to models with a singular covariance matrix. For these reasons, we develop a Monte-Carlo (MCMC) computation of fitting this model, which allows for both frequentist and Bayesian inference. There are technical challenges to developing this solution because one of the covariance matrices in the model is patterned. Our main application is to the National Institutes of Health (NIH)-AARP Diet and Health Study, where we illustrate our methods for modeling the energy-adjusted usual intake of fish and whole grains. We demonstrate numerically that our methods lead to increased speed of computation, converge to reasonable solutions, and have the flexibility to be used in either a frequentist or a Bayesian manner.

KEYWORDS: Bayesian approach, latent variables, measurement error, mixed effects models, nutritional epidemiology, zero-inflated data

Author Notes: This paper forms part of the Ph.D. dissertation of the first author at Texas A&M University. The research of Zhang, Perez and Carroll was supported by a grant from the National Cancer Institute (R37-CA057030). This publication is based in part on work supported by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST).

1 INTRODUCTION

This paper is about the important public health problem of understanding the distribution of episodically consumed dietary component intakes in terms of their energy-adjusted amounts, and relating this to diet-disease relationships. Before commenting in more detail, we first discuss the literature for simpler problems that are also of interest.

In nutritional surveillance and nutritional epidemiology, there is considerable interest in understanding the distribution of usual dietary intake, which is defined as long-term daily average intake. In addition, of interest is the regression of this intake on measured covariates, which is needed to correct diet-disease relationships for measurement error in assessing diet. If the dietary component of interest is ubiquitously consumed, as most nutrients are, the data are continuously distributed and methods are well-established for solving both problems. See for example Nusser, et al. (1997) for surveillance and Carroll, et al. (2006) for measurement error modeling.

Another class of dietary components is those which are episodically consumed, as is true of most foods, e.g., fish, red meat, dark green vegetables, whole grains. When consumption is measured by a short-term instrument such as a 24 hour recall, hereafter denoted by 24hr, the episodic nature of these dietary components means that their reported intake may either equal zero on a non-consumption day, or is positive on a day the component is consumed. In many studies, non-consumption days predominate for several episodically consumed foods of interest. For example, in our data example, for fish and whole grains, 65% and 12% reported no consumption on both of two administrations of the 24hr, respectively. Thus, data on episodically consumed dietary components are zero-inflated data with measurement error. Recently, Tooze, et al. (2006) for nutritional surveillance and Kipnis, et al. (2009) for nutritional epidemiology have reported so-called two-part methods, which are actually nonlinear mixed effects models, for analyzing episodically consumed dietary components in the univariate case. These methods are known commonly as the "NCI method" because many of the co-authors of these papers are members of the National Cancer Institute (NCI), and because SAS routines based upon the NLMIXED procedure are available at http://riskfactor.cancer.gov/diet/usualintakes/, an NCI web site. Other

two-part models in different contexts are described for example in Olsen and Schafer (2001), Tooze, et al. (2002) and Li, et al. (2005).

We are interested in the more complex public health problem of understanding the usual intake of an episodically consumed dietary component adjusted for energy intake (caloric intake), along with the distribution of usual intake of energy. This is critical because it addresses the issue of dietary component composition, and makes comparable diets of individuals whose usual intakes of energy are very different. As an example, the U.S. Department of Agriculture's Healthy Eating Index-2005 (www.cnpp.usda.gov/HealthyEatingIndex.htm) is a measure of diet quality that assesses conformance to Federal dietary guidance. One component of that index is the number of ounces of whole grains consumed per 1000 kilocalories: there are other items in the HEI-2005 that deal with episodically consumed dietary components, and all of them are adjusted for energy intake. The data needed to compute such variables are thus the usual intake of the dietary component consumed and the usual amount of calories consumed, and (possibly normalized) ratios of them.

Recently, Kipnis, et al. (2010) have developed a model for an episodically consumed dietary component and energy, see Section 2. They fit this model using nonlinear mixed effects models with likelihoods computed by adaptive Gaussian quadrature using the SAS procedure NLMIXED. However, as described in Section 2 and documented in Section 4, this form of computation can be slow, and can have serious convergence issues. This is extremely problematic, because of the importance of the problem and the fact that solutions will find wide use in the nutrition community, but only if they are numerically stable.

In this paper, we take an alternative Markov Chain Monte Carlo (MCMC) approach to computation, which is faster and numerically more stable. There are many good introductory papers reviewing MCMC, such as Casella, et al. (1992), Chib, et al. (1995) and Kass, et al. (1998). Effectively, we exploit the well-known fact (Lehmann and Casella, 1998, Chapter 6.8) that in fully parametric regular models of the type we study, Bayesian posterior means of parameters are asymptotically equivalent to their corresponding maximum likelihood estimators. To implement an MCMC approach in our problem, there are technical issues that have to be overcome, including the fact that one of the covariance matrices in the model of Kipnis, et al. (2010) is patterned.

DOI: 10.2202/1557-4679.1267

Besides fitting the model, our main focus in this paper is to discuss how to use the parameter estimates to then estimate the distributions of the usual intake of energy and energy-adjusted usual intake of dietary components.

In Section 2, we describe the model of Kipnis, et al. (2010). In Section 2, we also briefly outline some of the details of our implementation, although the technical details are given in the Appendix. In Sections 3 and 4, we take up the analysis of the NIH-AARP Study of Diet and Health (http://dietandhealth.cancer.gov/) as an illustration of our model and method. Section 5 gives concluding remarks.

2 Data and Model

2.1 The Data

In practice, the response data often come from repeated 24hr. Necessarily, due to cost and logistical reasons, the number of recalls is limited, and is rarely greater than 2. In a 24hr, what is observed is whether a dietary component is consumed, and if it is consumed, the reported amount. In addition, the amount of energy reported to be consumed is also available. Thus, for person $i = 1, ..., n$, and for the $k = 1, ..., m_i$ repeats of the 24hr, the data are $\widetilde{\mathbf{Y}}_{\mathbf{ik}} = (Y_{i1k}, Y_{i2k}, Y_{i3k})^{\mathrm{T}}$, where

- Y_{i1k} = Indicator of whether the episodically consumed dietary component is consumed.
- Y_{i2k} = Amount of the dietary component consumed as reported by the 24hr, which equals zero if the dietary component is not consumed.
- Y_{i3k} = Amount of energy consumed as reported by the 24hr.

There are also generally covariates such as age category, ethnic status and in many cases the results of reported intakes from a food frequency questionnaire. We will generically call these covariates $\mathbf{X}$.

2.2 The Model

Here we describe the nonlinear mixed effects latent variable model of Kipnis, et al. (2010). There are $i = 1, ..., n$ individuals and $k = 1, ..., m_i$ repeats

of the 24hr. Also, the observed data have three parts, relating to whether the episodically consumed dietary component is consumed, the amount if it is consumed, and the amount of energy. Also with the observed data, we will have covariates for the individual, generically called $\mathbf{X}$, see below for more precise notation. Finally, Kipnis, et al. (2010) use what are called in nutritional epidemiology "person-specific random effects" which are generically denoted by U, so that individuals actually differ from one another in usual intake when they have the same values of the covariates.

To be more precise, for the i^{th} individual there are covariates $(\mathbf{X_{i1}}, \mathbf{X_{i2}}, \mathbf{X_{i3}})$: $\mathbf{X}_{i1}$ are the covariates for the indicator of consumption, $\mathbf{X}_{i2}$ are the covariates for the consumption amount of the dietary component of interest, and $\mathbf{X}_{i3}$ are the covariates for the consumption of energy. Often, in practice, the covariates for each observed data component are the same, so that $\mathbf{X}_{i1} = \mathbf{X}_{i2} = \mathbf{X}_{i3}$. Along with the covariates, there are corresponding person specific random effects (U_{i1}, U_{i2}, U_{i3}), the role of which is to allow different people who share the same covariates to have different amounts of usual intakes. As we will see shortly, there are also errors accounting for day-to-day variation. Only the covariates, the person-specific random effects, and, because of transformations, the variances of the random errors are relevant to the definitions of usual intake, which are given below at equations (6)-(7).

The model of Kipnis, et al. (2010) uses a latent variable approach. Let $(W_{i1k}, W_{i2k}, W_{i3k})$ be latent variables that are assumed to follow the linear mixed effects model

$$W_{ijk} = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j + U_{ij} + \epsilon_{ijk} \text{ for } j = 1, 2, 3, \tag{1}$$

where $(U_{i1}, U_{i2}, U_{i3}) = \text{Normal}(0, \mathbf{\Sigma}_u)$ are the person-specific random effects, while the within-person errors that account for day-to-day variation $(\epsilon_{i1k}, \epsilon_{i2k}, \epsilon_{i3k}) = \text{Normal}(0, \mathbf{\Sigma}_\epsilon)$. The (U_{i1}, U_{i2}, U_{i3}) and $(\epsilon_{i1k}, \epsilon_{i2k}, \epsilon_{i3k})$ are mutually independent.

The observed data are related to the latent variables as follows:

$$\begin{aligned} Y_{i1k} &= I(W_{i1k} > 0); & (2)\\ Y_{i2k} &= Y_{i1k} g^{-1}(W_{i2k}, \lambda_F); & (3)\\ Y_{i3k} &= g^{-1}(W_{i3k}, \lambda_E), & (4) \end{aligned}$$

where $I(\cdot)$ is the indicator function and $g^{-1}(x, \lambda)$ is the inverse of the Box-Cox transformation $g(x, \lambda) = (x^\lambda - 1)/\lambda$ for $\lambda \neq 0$ and $g(x, 0) = \log(x)$ if $\lambda = 0$.

DOI: 10.2202/1557-4679.1267

We used the same Box-Cox transformations as Kipnis, et al. (2009, 2010). Under the model defined by (1)-(4), the probability to consume follows the probit model

$$\mathrm{pr}(Y_{i1k} = 1|\mathbf{X}_{i1}, U_{i1}, U_{i2}, U_{i3}) = \Phi(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1}), \tag{5}$$

where $\Phi(\cdot)$ is the standard normal distribution function. The probit model is commonly used to model a relationship between a binary dependent variable and one or more independent variables. The probit link was used in Kipnis, et al. (2010) to allow the day-to-day variation in whether a food is consumed to be correlated with the amount of energy consumed, and in such a way that the day-to-day variation random variables $(\epsilon_{i1k}, \epsilon_{i2k}, \epsilon_{i3k})$ are jointly normal, thus facilitating both nonlinear mixed effects software and the MCMC. The Box-Cox transformations in (3)-(4) allow for skewed distributions typically seen with dietary data. Of course, the notation in (5) means that consumption depends on (U_{i1}, U_{i2}, U_{i3}) only through U_{i1}.

Under the assumption that the 24hr is unbiased for usual (mean) intake, the usual intake of the dietary component and energy are given as $T_{Fi} = E(Y_{i2k}|\mathbf{X_{i1}}, \mathbf{X_{i2}}, U_{i1}, U_{i2})$ and $T_{Ei} = E(Y_{i3k}|\mathbf{X}_{i3}, U_{i3})$. Kipnis, et al. (2009, 2010) use a Taylor series approximation $E\{g^{-1}(v+\epsilon)|v) \approx g^{-1}(v,\lambda) + (1/2)\mathrm{var}(\epsilon)\{\partial^2 g^{-1}(v,\lambda)/\partial v^2\}$. Using this approximation, see equation (19) of Kipnis, et al. (2009), and under the covariance matrix restriction described below in Section 2.3, they show that the usual intake T_{Fi} of the dietary component and the usual intake T_{Ei} of energy for individual i are given as

$$T_{Fi} = \Phi(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1 + U_{i1}) g_*\{\mathbf{X}_{i2}^{\mathrm{T}}\boldsymbol{\beta}_2 + U_{i2}, \lambda_F, \boldsymbol{\Sigma}_\epsilon(2,2)\}, \tag{6}$$

$$T_{Ei} = g_*\{\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3 + U_{i3}, \lambda_E, \boldsymbol{\Sigma}_\epsilon(3,3)\}, \tag{7}$$

where the (j,k) element of $\boldsymbol{\Sigma}_\epsilon$ is denoted as $\boldsymbol{\Sigma}_\epsilon(j,k)$ and $g_*(v,\lambda,\sigma_\epsilon^2) = g^{-1}(v,\lambda) + (1/2)\sigma_\epsilon^2\{\partial^2 g^{-1}(v,\lambda)/\partial v^2\}$. Of course, (6)-(7) are approximations because $g_*(\cdot)$ is an approximate inverse of $g(\cdot)$. We can combine the usual intakes of dietary component and energy in various ways, e.g., the number of ounces of whole grains per 1000 kilo-calories, i.e., $1000 \times T_{Fi}/T_{Ei}$.

Remark 1 The Taylor series approximation to computing expectations of inverses of the Box-Cox transformation is used here because it was used by Kipnis, et al. (2009, 2010). More precise quadrature formulae can be used, and we have done so, finding almost no numerical changes. The computational convenience of the approximation makes it attractive.

2.3 Restriction on the Covariance Matrix

There are two restrictions necessary in the specification of $\mathbf{\Sigma}_\epsilon$. First, following Kipnis, et al. (2009, 2010), we set ϵ_{i1k} and ϵ_{i2k} to be independent. The intuitive way to think about the independence between the first two is that whether the dietary component is consumed or not and the amount consumed are assumed to be independent. This actually makes sense because a dietary component being consumed cannot indicate how much was consumed. Second, for identifiability of $\boldsymbol{\beta}_1$ and the distribution of U_{i1}, we require that $\text{var}(\epsilon_{i1k}) = 1$, because otherwise the marginal probability of consumption is $\Phi\{(\mathbf{X}_{i1}^{\text{T}}\boldsymbol{\beta}_1 + U_{i1})/var^{1/2}(\epsilon_{i1k})\}$. Without this second restriction, $\boldsymbol{\beta}_1$, $\text{var}(U_{i1})$, $\text{cov}(U_{i1}, U_{i2})$ and $\text{cov}(U_{i1}, U_{i3})$ are identified only up to scale factors. Hence we have that

$$\mathbf{\Sigma}_\epsilon = \begin{bmatrix} 1 & 0 & s_{13} \\ 0 & s_{22} & s_{23} \\ s_{13} & s_{23} & s_{33} \end{bmatrix}. \tag{8}$$

The difficulty with parameterizations such as (8) is that $(s_{13}, s_{23}, s_{22}, s_{33})$ cannot be left unconstrained, or else (8) need not be a covariance matrix. Define $s_{13} = \rho_{13}s_{33}^{1/2}$ and $s_{23} = \rho_{23}(s_{22}s_{33})^{1/2}$. Then the determinant $|\mathbf{\Sigma}_\epsilon| = s_{22}s_{33}(1 - \rho_{13}^2 - \rho_{23}^2)$. Since $\mathbf{\Sigma}_\epsilon$ is a covariance matrix, its determinant must be non-negative, and hence we cannot allow the correlations (ρ_{13}, ρ_{23}) to vary freely. There are many ways to parameterize $\mathbf{\Sigma}_\epsilon$ in an unrestricted way that forces it to be positive semi-definite. Here we use a *polar coordinate* representation, $\rho_{13} = \gamma\cos(\theta)$ while $\rho_{23} = \gamma\sin(\theta)$, with $\gamma \in (-1, 1)$ and $\theta \in (-\pi, \pi)$.

The zero entries in (8) are not required, although they are implicit in the two part model used in the original papers involving only the episodically consumed dietary component and not energy (Tooze, et al., 2006; Kipnis, et al., 2009) and they make intuitive sense in our context. We have chosen to use this restriction for these reasons and especially so that the marginal model for the episodically consumed dietary component is the same as that in the literature.

Kipnis, et al. (2010) explore a sample selection model (Heckman, 1976, 1979; Leung and Yu, 1996; Kyriazidou, 1997; Min and Agresti, 2002) that does not have this restriction. They found that such a sample selection model can be very unstable in our context, with the components of $\mathbf{\Sigma}_\mathbf{u}$ and $\mathbf{\Sigma}_\epsilon$ varying wildly.

DOI: 10.2202/1557-4679.1267

Although it is possible to use MCMC computations to fit the sample selection model, given the acceptance of the restriction in nutritional epidemiology and of the NCI method, we focus on the covariance model (8).

Remark 2 It is very important to allow for $\mathbf{\Sigma}_\epsilon$ being non-diagonal. The term $s_{23} \neq 0$ simply reflects the reality that, within a person and hence conditional on (U_{i1}, U_{i2}, U_{i3}), the amount of food reported consumed and the amount of energy consumed are sometimes highly correlated. The reason we allow $s_{13} \neq 0$ is to account for the very real possibility that, again within a person, the very fact that one consumes a food leads to a higher or lower reported energy (caloric) intake.

2.4 Model Fitting and Computation

It is possible in principle to fit model (1)-(8) using nonlinear mixed effects software. Kipnis, et al. (2010) use the SAS procedure PROC NLMIXED. However, we have found that such implementation is slow and not very stable, with many issues of convergence. NLMIXED uses adaptive Gaussian quadrature to integrate the likelihood over the distribution of random effects. NLMIXED can have convergence problems, especially when there are too many, or too few, zeros. What typically happens is that $\text{corr}(U_{i1}, U_{i2})$ tries to go to 1.00 or sometimes even -1.00, or that $\text{var}(U_{i1})$ or $\text{var}(U_{i2})$ tries to go to 0.00. When one of these things happens, the model usually converges, according to the change-in-likelihood criterion, but the Hessian is not positive definite. Occasionally, NLMIXED fails to converge at all. In general, we have found that when NLMIXED does not have such numerical problems, its results and ours are in reasonable agreement. These issues are described in more detail in Section 4.2.

Hence, for stability and speed, we have turned to a Bayesian approach for fitting the model described by equations (1)-(8). We emphasize that the Markov Chain Monte Carlo computation can either be thought of as a strictly Bayesian computation with ordinary Bayesian inference, or as a means of developing *frequentist* estimators of the crucial parameters, based on the well-known fact that in parametric models such as ours, the posterior mean of the parameters is a consistent and asymptotically normally distributed frequentist estimator, see for example Lehmann and Casella (1998, Chapter 6.8).

Our computational algorithm, described in detail in the appendix, uses Gibbs sampling with some Metropolis-Hastings steps. We have implemented this approach in both Matlab and R, and it is fast enough for practical use. In the NIH-AARP Diet and Health Study described in Section 3, with a sample size of 899, for a burn-in of $1,000$ steps followed by $10,000$ MCMC iterations, our Matlab and R programs take approximately 2 minutes and 11.7 minutes on an Intel(R) Xeon(TM) CPU with 3.73GHz and 7.8GB of RAM in a Linux system, respectively. For a burn-in of $5,000$ steps followed by $15,000$ MCMC iterations, our Matlab and R programs take approximately 3 minutes and 17.5 minutes, respectively. Both programs are available from the first author.

We have also developed an implementation in WinBUGS with a BUGS model called from R by using the package R2WinBUGS. Details are available from the third author. As to be expected, the WinBUGS code is much slower than the custom programs, taking approximately 5 hours (Pentium computer with 3.5GHz CPU and 1.99GB of RAM in a Windows system) for a burn-in of $1,000$ steps followed by $10,000$ MCMC samples. We are also currently developing a SAS macro for use by the nutritional community. On various test data sets, the WinBUGS, R, SAS and Matlab code gave very similar answers. In our empirical work, we use the Matlab code.

Remark 3 There are important data conventions that we use. These are described in detail in the Appendix. For example, in Section A.1, we mention that covariates are always standardized to have sample mean zero and sample variance one. The reason is a matter of scaling: energy intake is in terms of calories, which are typically in the 1,000's, so that the corresponding regression parameters, without standardization, with the FFQ energy as a covariate, would necessarily be tiny, making it hard to develop a plausible prior distribution. As described in Section A.1, we also standardize the responses for numerical stability and weaken dependence upon the prior distributions, and in Section A.2 we describe why this standardization makes sense. We have fit our method with various different prior distributions, and there is very little sensitivity to prior specification.

DOI: 10.2202/1557-4679.1267

2.5 The Role of Covariates

Covariates are important for estimating the distribution of usual intakes, for at least three reasons.

- First, as a matter of model specification. Consider abstractly the simple linear regression model $Y = \beta_0 + \beta_1 X + \epsilon$: given X, ϵ might be normally distributed, but if X is not simultaneously normally distributed, then removing it from the model would give a model $Y = \kappa_0 + \xi$, and ξ would not be normally distributed, and our model assumptions would be violated.

- Subar, et al. (2006) studied using food frequency questionnaire (FFQ) data as covariates to estimate the distributions of individual usual intakes of episodically consumed dietary components. They found strong and consistent relationships between FFQ and 24hr. This supports the postulate that FFQ data may provide important covariate information in supplementing 24hr for estimating usual intake of dietary components. Besides FFQ, there are some other clinical covariates such as gender, age, body mass index (BMI), etc. that may be associated with usual intake. Thus, our covariates included an intercept, age, BMI, the FFQ for energy intake and the FFQ for the dietary component of interest. They are used to reduce the error with which the usual intake is estimated, and to make more plausible our distributional assumptions.

- Kipnis, et al. (2009) state in their abstract "*One feature of the proposed method is that additional covariates potentially related to usual intake may be used to increase the precision of estimates of usual intake and of diet-health outcome associations*". In their introduction they state "*In Section 3, using data from the Eating at Americas Table Study (EATS), we quantify the increased precision obtained from including a FFQ report as a covariate*".

A referee has asked whether the β-coefficients for the covariates are interpretable, and whether it would be of interest to make inferences about whether the covariates are associated with usual intake. Because energy adjusted usual intakes involve three β-coefficients for each covariates, interpretation of any one of them is difficult. Whether a particular covariate is associated with usual in-

take is a mildly interesting question, but if far less important than estimating distributions of energy-adjusted usual intakes.

2.6 Simulation Study

We performed a simulation study that was based upon our empirical study given in Section 3, in order to ascertain whether the methodology results in reasonably unbiased estimates of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\Sigma}_{\mathbf{u}}, \boldsymbol{\Sigma}_{\epsilon})$. To test whether our algorithm can produce non-near-zero correlations when the true correlations are actually far from zero, we simulated 200 data sets, each of size $n = 1,000$, roughly the size of the NIH-AARP calibration cohort in Section 3. In this simulation, we used the same covariates for each of the three outcomes, i.e., we set $\mathbf{X}_{i1} = \mathbf{X}_{i2} = \mathbf{X}_{i3}$. The covariate vectors had three components, the first equal to 1.0 for an intercept, and the other two generated as Normal$(0, 1)$. The parameters $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ were generated as Uniform$(0, 1)$ for each simulated data set. We used

$$\boldsymbol{\Sigma}_u = \begin{bmatrix} 0.50 & 0.24 & 0.24 \\ 0.24 & 0.70 & 0.35 \\ 0.24 & 0.35 & 0.70 \end{bmatrix}; \qquad \boldsymbol{\Sigma}_\epsilon = \begin{bmatrix} 1.00 & 0.00 & 0.47 \\ 0.00 & 1.20 & 0.78 \\ 0.47 & 0.78 & 1.40 \end{bmatrix}.$$

The mean of the posterior means of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$ was unbiased overall and are not reported here. The mean of the posterior means of $(\boldsymbol{\Sigma}_{\mathbf{u}}, \boldsymbol{\Sigma}_{\epsilon})$ were

$$\widehat{\boldsymbol{\Sigma}}_u = \begin{bmatrix} 0.51 & 0.27 & 0.27 \\ 0.27 & 0.68 & 0.33 \\ 0.27 & 0.33 & 0.67 \end{bmatrix}; \qquad \widehat{\boldsymbol{\Sigma}}_\epsilon = \begin{bmatrix} 1.00 & 0.00 & 0.39 \\ 0.00 & 1.23 & 0.80 \\ 0.39 & 0.80 & 1.43 \end{bmatrix}.$$

Crucially, for the main purposes of estimating the distribution of usual intakes, the posterior means were essentially unbiased for estimating $\boldsymbol{\Sigma}_u$. As seen in the Appendix, $\boldsymbol{\Sigma}_\epsilon$ also has a role in the definition of usual intake, and it too was essentially unbiased except for a small bias of size 0.08 in estimating $\text{cov}(\epsilon_{i1k}, \epsilon_{i3k})$, a term that does not appear in the definitions of usual intake.

Remark 4 We give here only the results of a single simulation because what we have shown above are representative of other simulations we have done.

DOI: 10.2202/1557-4679.1267

For example, we have simulated cases where the off-diagonal elements of $\mathbf{\Sigma}_u$ were zero and cases where some of them were negative. We have also simulated cases that the diagonal elements of $\mathbf{\Sigma}_u$ were smaller and somewhat larger. In none of the cases did we see any significant bias in the estimates.

Remark 5 We have not displayed the simulation results for the Proc NLMIXED procedure because in those cases that it converges, it is very nearly unbiased, just like our method.

3 Empirical Analysis: Methods

3.1 Introduction to the NIH-AARP Diet and Health Study

The NIH-AARP Diet and Health Study, see http://dietandhealth.cancer.gov/ and Schatzkin, et al. (2001), has two components, the main study with diet assessed by a Food Frequency Questionnaire (FFQ) and a calibration sub-study with additional diet assessment by two 24hr. We considered a part of the main study that consists of $n_p = 142,364$ women, who contributed an FFQ as well as relevant demographic characteristics. The data used were the same as in Sinha, et al. (2010). The covariates $\mathbf{X}$ used included an intercept, age, body mass index, the FFQ for energy intake and the FFQ for the dietary component in question. The 24hr was not available for these subjects. Thus, the primary sample represents data on $\mathbf{X}_i = \mathbf{X}_{i1} = \mathbf{X}_{i2} = \mathbf{X}_{i3}$ for $i = 1, ..., n_p$.

In addition to the primary sample, there was a subsample of $n_v = 899$ women in the calibration sub-study who completed an FFQ and demographic characteristics, so that there are $\mathbf{X_i} = \mathbf{X}_{i1} = \mathbf{X}_{i2} = \mathbf{X}_{i3}$ for $= n_p + 1, ..., n_v + n_p$. In addition, these women completed two 24hr. Hence we observed $(Y_{i1k}, Y_{i2k}, Y_{i3k})$ for $k = 1, 2$ and for $i = n_p + 1, ..., n_v + n_p$.

We illustrate our computational algorithm using data from both the two 24hr and the FFQ for whole grains, fish and energy intake, along with covariates. Following Kipnis, et al. (2009, 2010), the FFQ values for fish, whole grain and energy intake were transformed using $\lambda = 0.25$, $\lambda = 0.33$ and $\lambda = 0.00$, respectively. The 24hr used $\lambda = 0.50$, $\lambda = 0.33$ and $\lambda = 0.33$, respectively.

The MCMC calculations result in samples from the posterior distribution of $\boldsymbol{\mathcal{B}} = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \boldsymbol{\beta}_3^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Sigma}_u$, $\boldsymbol{\Sigma}_\epsilon$ and (U_{i1}, U_{i2}, U_{i3}), the latter only for $i = n_p + 1, ..., n_v + n_p$. The means of the samples for $(\boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_{\mathbf{u}}, \boldsymbol{\Sigma}_\epsilon)$ can be taken as frequentist point estimates of these quantities, and are denoted here as $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \widehat{\boldsymbol{\beta}}_3, \widehat{\boldsymbol{\Sigma}}_u, \widehat{\boldsymbol{\Sigma}}_\epsilon)$. We will use shorthand notation for usual intake:

Usual dietary component intake is T_{Fi}
$$= \mathcal{G}_1\{\mathbf{X}_{i1}, \mathbf{X}_{i2}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, U_{i1}, U_{i2}, \boldsymbol{\Sigma}_\epsilon(2,2)\}, \text{ see (6)};$$
Usual energy intake is $T_{Ei} = \mathcal{G}_2\{\mathbf{X}_{i3}, \boldsymbol{\beta}_3, U_{i3}, \boldsymbol{\Sigma}_\epsilon(3,3)\}$, see (7).

For both usual dietary component intake and usual energy intake, 24hr samples are available for $i = n_p + 1, ..., n_v + n_p$.

3.2 Frequentist Analysis

We are going to write the variable of interest as $\mathcal{H}(T_{Fi}, T_{Ei})$. Thus, (a) the dietary component is $\mathcal{H}(T_{Fi}, T_{Ei}) = T_{Fi}$; (b) energy is $\mathcal{H}(T_{Fi}, T_{Ei}) = T_{Ei}$; and (c) the energy adjusted dietary component is $\mathcal{H}(T_{Fi}, T_{Ei}) = 1000 \times T_{Fi}/T_{Ei}$. In general then, the usual intake variable of interest for person i can be written as

$$\mathcal{Q}_i = \mathcal{H}\left[\mathcal{G}_1\{\mathbf{X}_{i1}, \mathbf{X}_{i2}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, U_{i1}, U_{i2}, \boldsymbol{\Sigma}_\epsilon(2,2)\}, \mathcal{G}_2\{\mathbf{X}_{i3}, \boldsymbol{\beta}_3, U_{i3}, \boldsymbol{\Sigma}_\epsilon(3,3)\}\right],$$

for $i = 1, ..., n_p + n_v$, where we have that $(U_{i1}, U_{i2}, U_{i3}) = \text{Normal}(0, \boldsymbol{\Sigma}_u)$.

Estimation of the distribution of $\mathcal{Q}$ across the population is easily accomplished by a Monte-Carlo computation. This is a different Monte-Carlo computation than the MCMC, and is performed after the MCMC has been done. Specifically, for a large B, where we took $B = 5,000$, and for $b = 1, ..., B$ generate $(U_{bi1}, U_{bi2}, U_{bi3}) = \text{Normal}(0, \widehat{\boldsymbol{\Sigma}}_{\mathbf{u}})$. Here B is not the number of burn-in steps, but simply a large enough number to do numerical integration. Then the distribution of usual intake can be estimated as the empirical distribution of the values

$$\mathcal{Q}_{bi} = \mathcal{H}\left[\mathcal{G}_1\{\mathbf{X}_{i1}, \mathbf{X}_{i2}, \widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, U_{bi1}, U_{bi2}, \widehat{\boldsymbol{\Sigma}}_\epsilon(2,2)\}, \mathcal{G}_2\{\mathbf{X}_{i3}, \widehat{\boldsymbol{\beta}}_3, U_{bi3}, \widehat{\boldsymbol{\Sigma}}_\epsilon(3,3)\}\right],$$

taken across $i = 1, ..., n_v + n_p$ and $b = 1, ..., B$.

Standard errors and confidence intervals for the distribution of usual intake can be formed easily by bootstrapping. We used 400 bootstrap samples in our numerical work.

DOI: 10.2202/1557-4679.1267

Remark 6 For bootstrap confidence intervals, it is often recommended to use at least 399 bootstrap samples, as we have done, see for example Davidson and MacKinnon (1999). We have experimented with using up to $1,000$ bootstrap samples, but this significantly increases computing time without changing the basic results in any material way.

3.3 Bayesian Analysis

As described below, Bayesian inference on the distribution of usual intake depends on estimating the distribution of the covariates. The distribution of usual intake $\mathcal{H}(T_F, T_E)$ in a population can be described as follows. Let $\boldsymbol{\mathcal{X}} = (\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3})$ and let $f_X(\boldsymbol{\mathcal{X}}|\boldsymbol{\zeta}) = f_X(\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \boldsymbol{\zeta})$ be the distribution of $\boldsymbol{\mathcal{X}}$ in the population, based on a parameter $\boldsymbol{\zeta}$. Write $\boldsymbol{\mathcal{U}} = (U_1, U_2, U_3)^{\mathrm{T}}$. Use the shorthand notation

$$\begin{aligned}&\mathcal{K}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{U}}, \boldsymbol{\Sigma}_\epsilon)\\ &\quad = \mathcal{H}\left[\mathcal{G}_1\{\mathbf{X_1}, \mathbf{X_2}, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, U_1, U_2, \boldsymbol{\Sigma}_\epsilon(2,2)\}, \mathcal{G}_2\{\mathbf{X_3}, \boldsymbol{\beta}_3, U_3, \boldsymbol{\Sigma}_\epsilon(3,3)\}\right].\end{aligned}$$

Then the distribution of usual intake is

$$\begin{aligned}F(v|\boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_u, \boldsymbol{\zeta}, \boldsymbol{\Sigma}_\epsilon) &= \mathrm{pr}\left\{\mathcal{K}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{U}}, \boldsymbol{\Sigma}_\epsilon) \le v|\boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_u, \boldsymbol{\Sigma}_\epsilon, \boldsymbol{\zeta}\right\}\\ &= \int I\left\{\mathcal{K}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\mathcal{U}}, \boldsymbol{\Sigma}_\epsilon) \le v\right\} f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{\mathcal{U}}|\boldsymbol{\Sigma}_{\mathbf{u}}) f_X(\boldsymbol{\mathcal{X}}|\boldsymbol{\zeta}) d\boldsymbol{\mathcal{U}} d\boldsymbol{\mathcal{X}}.\end{aligned}$$

We suggest approximating this using Monte-Carlo integration, as follows. Again, let B be large where we took $B = 1,000$, and for $b = 1, ...B$, let $\mathbf{u}_b = \mathrm{Normal}(0, \mathbf{I_3})$. Let $\boldsymbol{\Sigma}_u^{1/2}$ be the symmetric square root of $\boldsymbol{\Sigma}_u$. Then

$$F(v|\boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_u, \boldsymbol{\zeta}, \boldsymbol{\Sigma}_\epsilon) \approx B^{-1}\textstyle\sum_{b=1}^{B} \displaystyle\int I\left\{\mathcal{K}(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_u^{1/2}\mathbf{u}_b, \boldsymbol{\Sigma}_\epsilon) \le v\right\} f_X(\boldsymbol{\mathcal{X}}|\boldsymbol{\zeta}) d\boldsymbol{\mathcal{X}}.$$

The posterior distribution of $F(v|\boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_u, \boldsymbol{\zeta}, \boldsymbol{\Sigma}_\epsilon)$ is then calculated from the MCMC samples: our methods in the Appendix are easily generalized to sample from the posterior distribution of $\boldsymbol{\zeta}$.

In the NIH-AARP Diet and Health Study, with a sample size of $n_p + n_v > 140,000$, we effectively know the distribution of $\boldsymbol{\mathcal{X}}$. Let the values in the data be $\boldsymbol{\mathcal{X}}_i$ for $i = 1, ..., n_v + n_p$. Then we have

$$\begin{aligned}&F(v|\boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_u, \boldsymbol{\zeta}, \boldsymbol{\Sigma}_\epsilon)\\ &\quad \approx \{(n_v + n_p)B\}^{-1}\textstyle\sum_{b=1}^{B}\sum_{i=1}^{n_v+n_p} \displaystyle I\left\{\mathcal{K}(\boldsymbol{\mathcal{X}}_i, \boldsymbol{\mathcal{B}}, \boldsymbol{\Sigma}_u^{1/2}\mathbf{u}_b, \boldsymbol{\Sigma}_\epsilon) \le v\right\}.\end{aligned}$$

The posterior distribution of $F(v|\mathcal{B}, \mathbf{\Sigma}_u, \boldsymbol{\zeta}, \mathbf{\Sigma}_\epsilon)$ can then be calculated from the MCMC samples.

4 Results

Along with illustrating the distributions of usual intakes of the dietary components adjusted for energy, we also compared our results with NLMIXED.

4.1 Analysis

We used a burn-in of 5,000 steps followed by 15,000 MCMC samples. We saved every 10^{th} sample to reduce autocorrelation.

4.1.1 Frequentist Analysis

In Table 1 we present summary statistics (mean, standard deviation and selected percentiles) of the usual intakes as well as the usual intakes adjusted for energy. Figures 1 and 2 give density estimates for usual intake and energy adjusted intake of fish and whole grains, respectively: a similar plot for usual energy intake was also produced but not displayed here. The evident skewness of the usual intakes of fish and whole grains is expected, as are the somewhat less skewed nature of the energy adjusted intakes.

We bootstrapped the validation and primary data sets separately 400 times, see Remark 6, reran the analysis, and formed bootstrap confidence intervals. Since the distribution of the covariates X is essentially known because of the size of the primary study, this bootstrap simply reflects the uncertainty in the parameter estimates as they propagate through to usual intakes. To give a graphical summary including uncertainty, in Figure 3 we plot the actual estimated percentiles of the distribution of adjusted fish intake against the percentile number, as well as the 95% pointwise bootstrap confidence interval for these percentiles.

DOI: 10.2202/1557-4679.1267

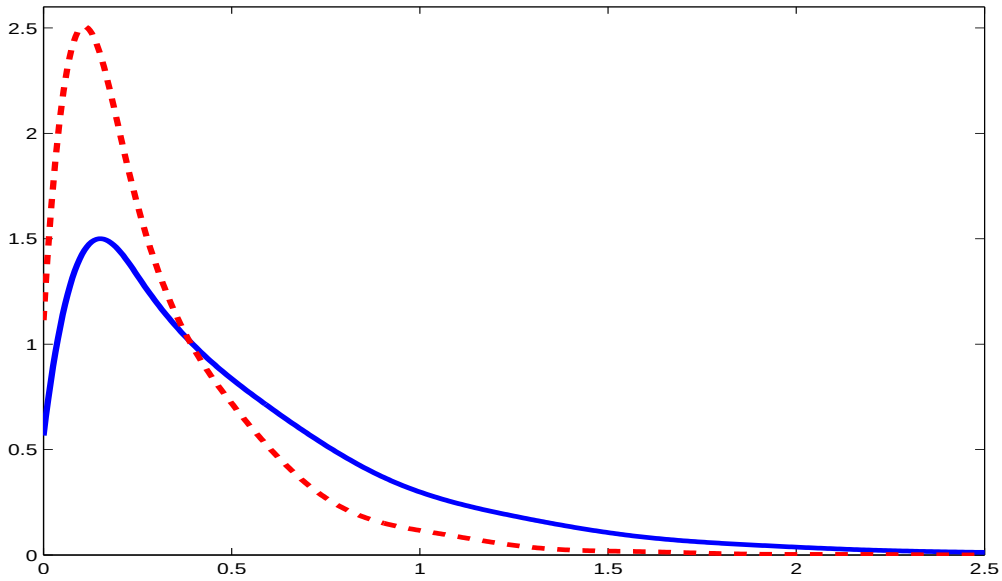

Figure 1: Density estimates for fish. The solid line is the density estimate for usual intake in the unit of oz. The dashed line is the density estimate for usual intake per 1000 kilo-calories.

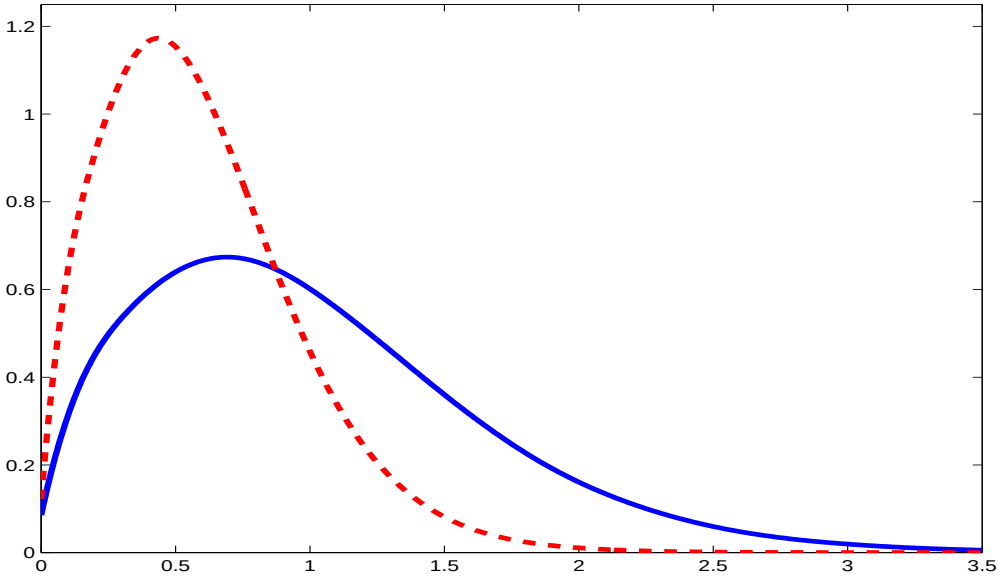

Figure 2: Density estimates for whole grains. The solid line is the density estimate for usual intake in the unit of cups. The dashed line is the density estimate for usual intake per 1000 kilo-calories.

	Whole Grains		Fish			Energy
	Usual Intake (Unit: cup)	per 1000 kcals	Usual Intake (Unit: oz.)	per 1000 kcals	Bayes, per 1000 kcals	Usual Intake (Unit: kcal)
Mean	1.013	0.625	0.539	0.338	0.339	1631.77
s.d.	0.631	0.375	0.486	0.309	0.315	369.16
5^{th}	0.181	0.121	0.053	0.033	0.028	1075.70
10^{th}	0.287	0.189	0.089	0.057	0.057	1180.37
25^{th}	0.536	0.345	0.193	0.122	0.122	1370.29
50^{th}	0.911	0.569	0.399	0.249	0.249	1604.04
75^{th}	1.375	0.841	0.736	0.456	0.456	1863.01
90^{th}	1.867	1.127	1.176	0.731	0.731	2118.74
95^{th}	2.195	1.320	1.508	0.945	0.951	2282.50

Table 1: Estimated distributions of the usual intake for Whole Grains, Fish and Energy and the estimated distributions of energy-adjusted usual intake for Whole Grains and Fish, for women. The 5^{th} percentile of the distribution is labeled as 5^{th}, etc. For energy-adjusted fish intake, we give the results for both the frequentist ("Freq") and the Bayesian ("Bayes") fits. Estimates were very similar for both Freq and Bayes fits and thus we have only displayed results for fish.

4.1.2 Bayesian Analysis

In Table 1 we also give the Bayesian analysis for energy-adjusted fish intake. As seen there, the Bayesian analysis posterior means of the distribution of energy-adjusted fish intake is nearly identical to the frequentist analysis. The same thing was found for all the columns in Table 1.

In addition, posterior credible interval lengths were almost equivalent to those of the frequentist method and are not displayed here.

DOI: 10.2202/1557-4679.1267

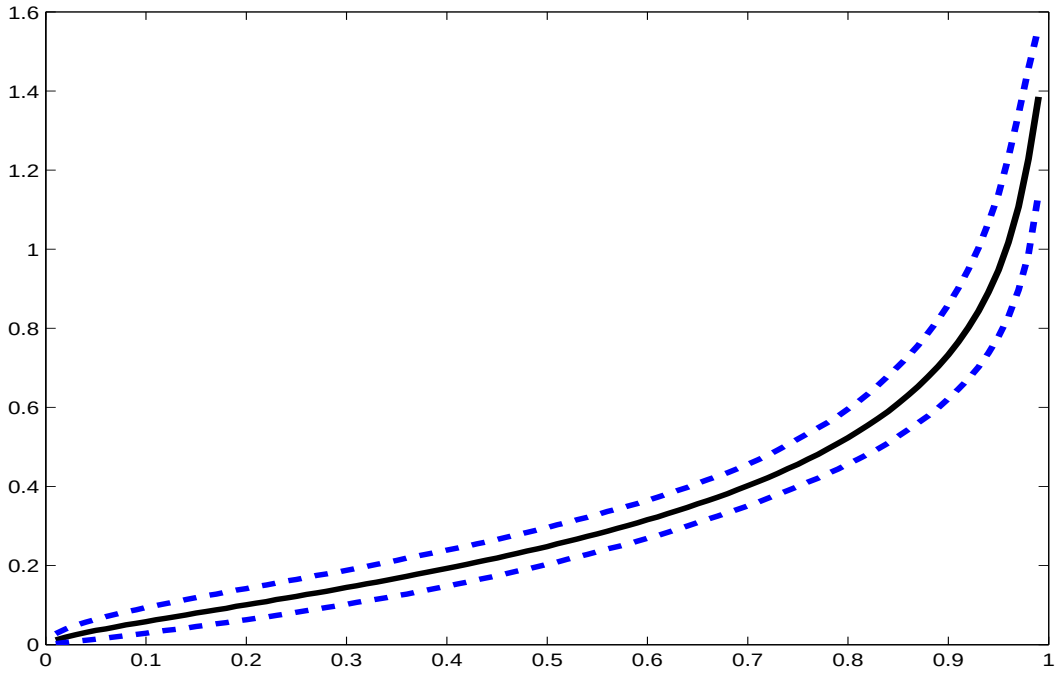

Figure 3: Quantile functions for usual fish intake per 1000 kilo-calories. Horizontal axis is the relative percentile, e.g., the value at 50 is the median. The vertical axis is the estimated percentile (solid line) in the unit of oz./(1000 kcal). Dashed lines are the pointwise 95% bootstrap confidence intervals.

4.2 Comparison With Proc NLMIXED

We described in Section 2.4 some of the motivation for our computational approach. In this section, we show documentation of those claims.

First, in Table 2, we describe aspects of the analysis for women of whole grains, fish and dark-green vegetables, using the AARP data set. The first line in the table is the number of minutes of computation for the nonlinear mixed effects program and our MCMC approach. It can be seen that the MCMC approach is considerably faster. While not displayed here, for Milk for men, which had only 12% reported non-consumption on the 24hr, the nonlinear mixed effects program took 200 minutes, while ours took only 4 minutes. This illustrates our claim concerning speed of computation.

	Whole Grains		Fish		Dark Green	
	NLMIXED	MCMC	NLMIXED	MCMC	NLMIXED	MCMC
Time in Minutes	20	3	12	3	12	4
% zeros on 24hr	32%		77%		73%	
Correlations						
$\text{corr}(U_{i1}, U_{i2})$	0.65	0.48	-0.39	0.08	1.00	0.48
	(0.17)	(0.09)	(0.44)	(0.07)	(N/A)	(0.06)
$\text{corr}(U_{i1}, U_{i3})$	0.20	0.18	0.28	0.26	0.27	0.24
	(0.08)	(0.07)	(0.14)	(0.07)	(N/A)	(0.06)
$\text{corr}(U_{i2}, U_{i3})$	0.37	0.40	0.02	0.02	0.27	0.28
	(0.10)	(0.07)	(0.16)	(0.09)	(N/A)	(0.06)

Table 2: Comparison between two computational methods, "NLMIXED" and "MCMC", to fit the bivariate nonlinear mixed effects model, for whole grains, fish and dark-green vegetables. Displayed are the estimates of correlations among the components of (U_{i1}, U_{i2}, U_{i3}), the estimates for the MCMC approach being posterior means. The numbers displayed in parentheses are the standard errors from the inverse of the Hessian matrix ("NLMIXED") and from MCMC samples ("MCMC"). Here "Dark Green" refers to Dark-Green vegetables, where the nonlinear mixed effects analysis converged but to a singular covariance matrix for Σ_u. The phrase "Time in Minutes" refers to computation time to complete the analysis. The overall % of zeros from the 24hr are also displayed.

A second aspect is that we claimed that sometimes the nonlinear mixed effects analysis of Kipnis, et al. (2010) suffered from convergence to a singular covariance matrix estimate for Σ_u. This occurred for dark-green vegetables, see Table 2, where it was estimated that the correlation between (U_{i1}, U_{i2}), $\text{corr}(U_{i1}, U_{i2})$, was equal to 1.00. This seemingly ridiculous result is in marked contrast to the much more sensible posterior mean of 0.48.

A third aspect of the comparison is that we claimed that when the method of Kipnis, et al. (2010) converged to a reasonable answer, our results were in general agreement with theirs. This is borne out in Table 2, where we have listed the standard errors of the estimates using the Hessian for the nonlinear mixed effects analysis, and using the MCMC samples for our method. The

DOI: 10.2202/1557-4679.1267

estimates are quite similar with the exception of $\mathrm{corr}(U_{i1}, U_{i2})$ for fish, which can be explained as follows. We performed a separate bootstrap calculation for this correlation with our method and the nonlinear mixed effects analysis, which suggested a standard error as large as the difference between the two. The other standard errors are also different, but this may well reflect imprecision in the former caused by using a Hessian in a nonlinear mixed effects model instead of a bootstrap.

Remark 7 While it may seem obvious, it is useful to clarify what we mean by the term "*convergence*". We are not meaning asymptotic rates of convergence, because these are the standard $n^{1/2}$-type one sees in parametric models. We are also not talking about theoretical rates of numerical convergence, e.g., how fast is convergence of the Proc NLMIXED procedure in terms of number of iterations. Instead, for us the term convergence has the meaning that Proc NLMIXED announces that it has converged to a solution with a nonsingular Hessian. Of course, our method, being based on proper priors, converges in the usual MCMC sense.

5 Discussion

Understanding the distribution of energy-adjusted usual intake of episodically consumed dietary components is of considerable public health importance, having implications for basic understanding of both dietary component composition and policy. Being able to correct for measurement error due to within-person variation in short-term assessment of intake, when investigating diet-disease relationships in cohort studies, is equally important. Because of the importance of these problems, models and fitting methods for addressing them will find wide use in the nutrition community. Thus, it is not only important that the models are reasonable, but that the fitting methods be reasonably fast, that they converge, and that the answers from the fitting methods usually make sense. The main point of this paper has been to show that an MCMC approach satisfies these criteria, and has the potential to be used widely in the

nutrition community. The fact that the MCMC approach can be used in a frequentist sense is a new insight for nutritional epidemiology, which is decidedly frequentist in orientation, although the MCMC model fitting can also allow Bayesian inference.

There is an enormous literature on measurement error models, both parametric and nonparametric, for estimating distributions (e.g., Fan, 1991; Wand, 1998; Johnson, et al., 2007; Staudenmeyer, et al., 2008; Delaigle, et al, 2008 among many others) and in regression (Ferrari, et al., 2004, e.g., Liang and Wang, 2005 among many others). Many more references are given in Carroll, et al. (2006). However, none of these papers deal with our topic of episodically consumed and hence zero-inflated dietary components along with continuous components that involve skewness, a structured covariance matrix, correlations of random effects, and usual intakes on the original data scale.

An issue of practically much less importance is that the model of Kipnis, et al. (2010) in equation (6) assumes that each food is consumed by all individuals. Kipnis, et al. (2009) address this issue, by adding a fixed effect regression so as to model never-consumers. They show that even without energy in the model, and with only two 24hr as is standard for such data, their method was numerically very unstable. Our method easily handles such an extension, but its practical implications are not particularly clear when, for example, in other studies, less than 0.5% of subjects claimed on the FFQ never to eat fish or whole grains.

User-friendly SAS macros are being written for distribution to the nutrition community. These programs will also allow sampling weights, so that they can be used in population-based survey samples, and will thus be of interest both nationally and internationally. We are presently working on extending the methods to analyze multiple foods and nutrients simultaneously, with allowance for survey weights, so that analysis of dietary patterns and dietary composite scores can be undertaken.

DOI: 10.2202/1557-4679.1267

Appendix: Details of the MCMC

A.1 Notational Convention

Standardization is important in MCMC applications both for numerical stability and to allow fairly off-the-shelf prior distributions to make sense. Prior to analysis, we standardized the covariates to have mean 0.0 and variance 1.0. The observed, transformed non-zero 24hr were standardized to have mean 0.0 and variance 2.0. More precisely, we first transformed the non-zero dietary component data as $Z_{i2k} = g(Y_{i2k}, \lambda_F)$, and then we standardized these data as $Q_{i2k} = \sqrt{2}(Z_{i2k} - a_F)/s_F$. Similarly, for energy we transformed to $Z_{i3k} = g(Y_{i3k}, \lambda_E)$ and then standardized to $Q_{i3k} = \sqrt{2}(Z_{i3k} - a_E)/s_E$. Of course, whether the dietary component is consumed or not is $Q_{i1k} = Y_{i1k}$. Collected, the data are $\widetilde{\mathbf{Q}}_{ik} = (Q_{i1k}, Q_{i2k}, Q_{i3k})^{\mathrm{T}}$. The terms (a_F, s_F, a_E, s_E) are not random variables but are merely constants used for standardization, and we need not consider inference for them.

We will first describe the algorithm used in terms of the Q_{ijk}, and then in Section A.11, we describe the back-transformation method that we used to obtain estimation and inference for usual intake.

Remark 8 Having the total variability of the non-zero transformed responses equal to 2.0 is extraordinarily convenient. Effectively, this means that $\mathrm{var}(U_{ij}) + \mathrm{var}(\epsilon_{ij}) \approx 2.0$ for $j = 1, 2$. Thus, neither component of this sum is at all likely to be large. Hence, a prior mean for the diagonal elements of Σ_u all equalling 1.0, while too large in our examples, is at least relatively near a reasonable answer. Having priors for $\mathrm{var}(\epsilon_{ij})$ for $j = 1, 2$ that are Uniform$[0, 3]$ is flexible and does not allow ridiculous answers.

A.2 Prior Distributions

Because the data were standardized, following the discussion of Remark 8, we used the following conventions.

- The priors for all $\boldsymbol{\beta}_j$ were normal with mean zero and variance 100.

- The prior for $\mathbf{\Sigma}_u$ was exchangeable with diagonal entries all equal to 1.0 and correlations 0.50. There was 5 degrees of freedom in the inverse Wishart prior, i.e., $m_u = 5$. Thus, the prior is IW$\{(m_u - 3 - 1)\mathbf{\Omega}_u, m_u\}$.

- The priors for s_{22} and s_{33} were Uniform[0,3]. This range is reasonable because of the standardization.

- The priors for (γ, θ) were uniform on their range.

We experimented with different priors for $\mathbf{\Sigma}_u$, e.g., setting the correlations equal to 0.0, setting the diagonal elements equal to 0.5, etc. The results were essentially unchanged when these were done.

A.3 Generating Starting Values for the Latent Variables

While we observe $\widetilde{\mathbf{Q}}_{ik}$, in the MCMC we need to generate the latent variables $\widetilde{\mathbf{W}}_{ik}$ to initiate the MCMC.

- For energy, $Q_{i3k} = W_{i3k}$, no data need to be generated.

- For the amounts, Q_{i2k}, we just simply set $W_{i2k} = Q_{i2k}$.

- For consumption, we generate $\boldsymbol{\mathcal{U}}_i = (U_{i1}, U_{i2}, U_{i3})^{\mathrm{T}}$ as normally distribution with mean zero and covariance matrix given as the prior covariance matrix for $\mathbf{\Sigma}_u$. We then also compute $z_{ik} = |\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_{1,\mathrm{prior}} + U_{i1} + \mathcal{Z}_{ik}|$, where $\mathcal{Z}_{ik} = \mathrm{Normal}(0, 1)$ are generated independently. We then set $W_{i1k} = z_{ik}Q_{i1k} - z_{ik}(1 - Q_{i1k})$.

- We then updated $\widetilde{\mathbf{W}}_{ik}$ by a single application of the updates given in Section A.9.

DOI: 10.2202/1557-4679.1267

A.4 Complete Data Loglikelihood

The loglikelihood of the complete data is

$$\begin{aligned}
&\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\log\{Q_{i1k}I(W_{i1k}>0)+(1-Q_{i1k})I(W_{i1k}<0)\}\\
&+(n/2)\log(|\boldsymbol{\Sigma}_u^{-1}|)-(1/2)\textstyle\sum_{i=1}^{n}\boldsymbol{\mathcal{U}}_i^{\mathrm{T}}\boldsymbol{\Sigma}_u^{-1}\boldsymbol{\mathcal{U}}_i\\
&-(1/2)\textstyle\sum_{j=1}^{3}(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j,\mathrm{prior}})^{\mathrm{T}}\boldsymbol{\Omega}_{\beta,j}^{-1}(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j,\mathrm{prior}})\\
&+\{(m_u+3+1)/2\}\log(|\boldsymbol{\Sigma}_u^{-1}|)-(1/2)\mathrm{trace}(\boldsymbol{\Omega}_u\boldsymbol{\Sigma}_u^{-1})\\
&-(1/2)(2n)\log\{s_{22}s_{33}(1-\gamma^2)\}\\
&-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{ik}-(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1,...,\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}}-\boldsymbol{\mathcal{U}}_i\}^{\mathrm{T}}{\boldsymbol{\Sigma}_\epsilon}^{-1}\\
&\qquad\times\{\widetilde{\mathbf{W}}_{ik}-(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1,...,\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}}-\boldsymbol{\mathcal{U}}_i\}.
\end{aligned}$$

A.5 Complete Conditionals for $(\gamma, \theta, s_{22}, s_{33})$

The complete conditionals for $(\gamma, \theta, s_{22}, s_{33})$ do not have an explicit form, so we use a Metropolis-Hastings within Gibbs sampler to generate them in turn. Since $\boldsymbol{\Sigma}_\epsilon$ is determined by γ, θ, s_{22} and s_{33}, we write it as ${\boldsymbol{\Sigma}_\epsilon}^{-1} \equiv f(\gamma, \theta, s_{22}, s_{33})$. Also, current values are γ_t, θ_t, $s_{22,t}$ and $s_{33,t}$.

Generation of γ. For convenience, we set γ to be discrete with 41 equally-spaced values on its range. Let γ_t be the current value. The candidate value y is selected randomly from γ_t and its two nearest neighbors. The candidate value y is accepted with probability $\alpha(\gamma_t, y)$, $\alpha(\gamma_t, y) = \min\{1, g(y)/g(\gamma_t)\}$, where

$$\begin{aligned}
g(y) \;\propto\; & (1-y^2)^{-n}\\
&\times\exp\Big[-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{ik}-(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1,...,\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}}-\boldsymbol{\mathcal{U}}_i\}^{\mathrm{T}}\\
&\qquad\times f(y,\theta_t,s_{22,t},s_{33,t})\{\bullet\}\Big],
\end{aligned}$$

where $\{\bullet\}$ means that the term before $f(\cdot)$ is transposed and substituted. If the candidate y is accepted, then $\gamma_{t+1} = y$. Otherwise, $\gamma_{t+1} = \gamma_t$.

Generation of θ. This is done exactly as for γ, except now

$$g(y) \propto \exp\Bigg[-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \boldsymbol{\mathcal{U}}_i\}^{\mathrm{T}} \times f(\gamma_{t+1}, y, s_{22,t}, s_{33,t})\{\bullet\}\Bigg].$$

If the candidate y is accepted, then $\theta_{t+1} = y$. Otherwise, $\theta_{t+1} = \theta_t$.
Generation of s_{22}. Suppose the current value of s_{22} is $s_{22,t}$. A candidate value y is generated from the Uniform distribution of length 0.4 with mean $s_{22,t}$: y $\sim$ Uniform[$s_{22,t}$ - 0.2, $s_{22,t}$ + 0.2]. The candidate value y is accepted with probability $\alpha(s_{22,t}, y)$, where

$$\begin{aligned}\alpha(s_{22,t}, y) &= \min\left\{(1, g(y)I_{[0,3]}(y)/g(s_{22,t})\right\};\\ g(y) &\propto y^{-n}\exp\Bigg[-(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \boldsymbol{\mathcal{U}}_i\}^{\mathrm{T}} \times f(\gamma_{t+1}, \theta_{t+1}, y, s_{33,t})\{\bullet\}\Bigg]\end{aligned}$$

If the candidate is accepted, then $s_{22,t+1} = y$. Otherwise, $s_{22,t+1} = s_{22,t}$.
Generation of s_{33}. This is the same as that for s_{22}, except now

$$\begin{aligned}\alpha(s_{33,t}, y) &= \min\left\{1, g(y)I_{[0,3]}(y)/g(s_{33,t})\right\};\\ g(y) &\propto y^{-n}\exp\Bigg[-\frac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{ik} - (\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1, ..., \mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}} - \boldsymbol{\mathcal{U}}_i\}^{\mathrm{T}} \times f(\gamma_{t+1}, \theta_{t+1}, s_{22,t+1}, y)\{\bullet\}\Bigg].\end{aligned}$$

If the candidate is accepted, then $s_{33,t+1} = y$. Otherwise, $s_{33,t+1} = s_{33,t}$.

A.6 Complete Conditional for $\boldsymbol{\Sigma}_u$

By "rest", we mean all the observable data, latent variables and parameters other than the one in question. By inspection, the complete conditional for $\boldsymbol{\Sigma}_u$ is

$$[\boldsymbol{\Sigma}_u|\text{rest}] = \text{IW}\{(m_u - K - 1)\boldsymbol{\Omega}_u + \textstyle\sum_{i=1}^{n}\boldsymbol{\mathcal{U}}_i\boldsymbol{\mathcal{U}}_i^{\mathrm{T}}, n + m_u\}.$$

DOI: 10.2202/1557-4679.1267

A.7 Complete Conditionals for β

Let the elements of $\mathbf{\Sigma}_\epsilon^{-\mathbf{1}}$ be $\sigma_\epsilon^{j\ell}$. For any j, except for irrelevant constants,

$$\begin{aligned}
\log\left[\boldsymbol{\beta}_j|\text{rest}\right] &= -(1/2)(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j,\text{prior}})^{\mathrm{T}}\mathbf{\Omega}_{\beta,j}^{-1}(\boldsymbol{\beta}_j-\boldsymbol{\beta}_{j,\text{prior}}) \\
&\quad -(1/2)\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}(W_{ijk}-\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j-U_{ij})^2\sigma_\epsilon^{jj} \\
&\quad -\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\sum_{\ell\neq j}\sigma_\epsilon^{j\ell}(W_{ijk}-\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j-U_{ij})(W_{i\ell k}-\mathbf{X}_{i\ell}^{\mathrm{T}}\boldsymbol{\beta}_\ell-U_{i\ell}) \\
&= \mathfrak{C}_1^{\mathrm{T}}\boldsymbol{\beta}_j-(1/2)\boldsymbol{\beta}_j^{\mathrm{T}}\mathfrak{C}_2^{-1}\boldsymbol{\beta}_j
\end{aligned}$$

which implies $\left[\boldsymbol{\beta}_j|\text{rest}\right]\sim\text{Normal}(\mathfrak{C}_2\mathfrak{C}_1,\mathfrak{C}_2)$, where

$$\begin{aligned}
\mathfrak{C}_2 &= (\mathbf{\Omega}_{\beta,j}^{-1}+2\textstyle\sum_{i=1}^{n}\sigma_\epsilon^{jj}\mathbf{X}_{ij}\mathbf{X}_{ij}^{\mathrm{T}})^{-1}; \\
\mathfrak{C}_1 &= \mathbf{\Omega}_{\beta,j}^{-1}\boldsymbol{\beta}_{j,\text{prior}}+\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\sigma_\epsilon^{jj}\mathbf{X}_{ij}(W_{ijk}-U_{ij}) \\
&\quad +\textstyle\sum_{i=1}^{n}\sum_{k=1}^{2}\sum_{\ell\neq j}\sigma_\epsilon^{j\ell}(W_{i\ell k}-\mathbf{X}_{i\ell}^{\mathrm{T}}\boldsymbol{\beta}_\ell-U_{i\ell})\mathbf{X}_{ij}.
\end{aligned}$$

A.8 Complete Conditionals for $\mathcal{U}_i$

Except for irrelevant constants, and remembering that $j=1,...,3$,

$$\begin{aligned}
\log\left[\widetilde{\mathbf{U}}_{\mathbf{i}}|\text{rest}\right] &= -(1/2)\boldsymbol{\mathcal{U}}_i^{\mathrm{T}}\mathbf{\Sigma}_{\mathbf{u}}^{-\mathbf{1}}\boldsymbol{\mathcal{U}}_i \\
&\quad -(1/2)\textstyle\sum_{k=1}^{2}\{\widetilde{\mathbf{W}}_{ik}-(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1,...,\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}}-\boldsymbol{\mathcal{U}}_i\}^{\mathrm{T}}\mathbf{\Sigma}_\epsilon{}^{-1} \\
&\qquad\qquad \times\{\widetilde{\mathbf{W}}_{ik}-(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1,...,\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}}-\boldsymbol{\mathcal{U}}_i\} \\
&= \mathfrak{C}_1^{\mathrm{T}}\boldsymbol{\mathcal{U}}_i-(1/2)\boldsymbol{\mathcal{U}}_i^{\mathrm{T}}\mathfrak{C}_2^{-1}\boldsymbol{\mathcal{U}}_i
\end{aligned}$$

which implies $[\boldsymbol{\mathcal{U}}_i|\text{rest}]\sim\text{Normal}(\mathfrak{C}_2\mathfrak{C}_1,\mathfrak{C}_2)$, where

$$\begin{aligned}
\mathfrak{C}_2 &= (\mathbf{\Sigma}_u^{-1}+2\mathbf{\Sigma}_\epsilon{}^{-1})^{-1}; \\
\mathfrak{C}_1 &= \textstyle\sum_{k=1}^{2}\mathbf{\Sigma}_\epsilon{}^{-1}\{\widetilde{\mathbf{W}}_{ik}-(\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1,...,\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3)^{\mathrm{T}}\}.
\end{aligned}$$

A.9 Complete Conditionals for W_{i1k}

Here we do the complete conditional for $W_{i\ell k}$ with $\ell = 1$. Except for irrelevant constants,

$$\begin{aligned}
\log[W_{i\ell k}|\text{rest}] &= \log\{Q_{i\ell k}I(W_{i\ell k}>0)+(1-Q_{i\ell k})I(W_{i\ell k}<0)\} \\
&\quad -(1/2)(W_{i1k}-\mathbf{X}_{i1}^{\mathrm{T}}\boldsymbol{\beta}_1-U_{i1},...,W_{i3k}-\mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3-U_{i3})\boldsymbol{\Sigma}_\epsilon{}^{-1}(\bullet) \\
&= \log\{Q_{i\ell k}I(W_{i\ell k}>0)+(1-Q_{i\ell k})I(W_{i\ell k}<0)\} \\
&\quad -(1/2)\sigma_\epsilon^{\ell\ell}(W_{i\ell k}-\mathbf{X}_{i\ell}^{\mathrm{T}}\boldsymbol{\beta}_\ell-U_{i\ell})^2 \\
&\quad -\textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{i\ell k}-\mathbf{X}_{i\ell}^{\mathrm{T}}\boldsymbol{\beta}_\ell-U_{i\ell})(W_{ijk}-\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j-U_{ij}) \\
&= \log\{Q_{i\ell k}I(W_{i\ell k}>0)+(1-Q_{i\ell k})I(W_{i\ell k}<0)\} \\
&\quad +\mathcal{C}_1W_{i\ell k}-(1/2)W_{i\ell k}^2\mathcal{C}_2^{-1},
\end{aligned}$$

where

$$\begin{aligned}
\mathcal{C}_2 &= 1/(\sigma_\epsilon^{\ell\ell}) \\
\mathcal{C}_1 &= \sigma_\epsilon^{\ell\ell}(\mathbf{X}_{i\ell}^{\mathrm{T}}\boldsymbol{\beta}_\ell+U_{i\ell})-\textstyle\sum_{j\neq\ell}\sigma_\epsilon^{\ell j}(W_{ijk}-\mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j-U_{ij}).
\end{aligned}$$

If we use the notation $\mathrm{TN}_+(\mu,\sigma,c)$ for a normal random variable with mean μ, standard deviation σ is truncated from the left at c, and $\mathrm{TN}_-(\mu,\sigma,c)$ is truncated from the right at c, then it follows that with $\mu=\mathcal{C}_2\mathcal{C}_1$ and $\sigma=\mathcal{C}_2^{1/2}$,

$$\begin{aligned}
[W_{i\ell k}|\text{rest}] &= Q_{i\ell k}\mathrm{TN}_+(\mu,\sigma,0)+(1-Q_{i\ell k})\mathrm{TN}_-(\mu,\sigma,0) \\
&= \mu+Q_{i\ell k}\mathrm{TN}_+(0,\sigma,-\mu)+(1-Q_{i\ell k})\mathrm{TN}_-(0,\sigma,-\mu) \\
&= \mu+Q_{i\ell k}\mathrm{TN}_+(0,\sigma,-\mu)-(1-Q_{i\ell k})\mathrm{TN}_+(0,\sigma,\mu) \\
&= \mu+\sigma\{Q_{i\ell k}\mathrm{TN}_+(0,1,-\mu/\sigma)-(1-Q_{i\ell k})\mathrm{TN}_+(0,1,\mu/\sigma)\}.
\end{aligned}$$

Generating $\mathrm{TN}_+(0,1,c)$ is easy: if $c<0$, simply do rejection sampling of a Normal$(0,1)$ until you get one that is $>c$. If $c>0$, there is an adaptive rejection scheme (Robert, 1995). The "truncated normal" was used because the latent variable W_{i1k} is associated with Y_{i1k} which indicates whether the dietary component is consumed or not. If the dietary component is indeed consumed, then based on our model (2), W_{i1k} should have a positive value. Similarly, if the dietary component is actually not consumed, then W_{i1k} should

DOI: 10.2202/1557-4679.1267

have a negative value. In order to achieve these, we need a truncated distribution. Besides, the conditional distribution of W_{i1k} proportional to a normal distribution, thus we chose truncated normal.

A.10 Complete Conditionals for W_{i2k} When it is Not Observed

For $p = 2$, the variable W_{ipk} is not observed when $Q_{i,p-1,k} = 0$, or, equivalently, when $W_{i,p-1,k} < 0$. Except for irrelevant constants,

$$\begin{aligned}\log\left[W_{ipk}|\text{rest}\right] &= -(1/2)\sum_j\sum_\ell \sigma_\epsilon^{j\ell}(W_{ijk} - \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_j - U_{ij})(W_{i\ell k} - \mathbf{X}_{i\ell}^{\mathrm{T}}\boldsymbol{\beta}_\ell - U_{i\ell}) \\ &= -(1/2)W_{ipk}^2\mathcal{C}_2^{-1} + \mathcal{C}_1 W_{ipk}\end{aligned}$$

where

$$\begin{aligned}\mathcal{C}_2 &= 1/(\sigma_\epsilon^{pp}); \\ \mathcal{C}_1 &= \sigma_\epsilon^{pp}(\mathbf{X}_{ip}^{\mathrm{T}}\boldsymbol{\beta}_p + U_{ip}) - \sum_{\ell\neq p}\sigma_\epsilon^{p\ell}(W_{i\ell k} - \mathbf{X}_{i\ell}^{\mathrm{T}}\boldsymbol{\beta}_\ell - U_{i\ell}).\end{aligned}$$

Therefore,

$$[W_{ipk}|\text{rest}] = Q_{ipk}Q_{i,p-1,k} + (1 - Q_{i,p-1,k})\text{Normal}(\mathcal{C}_2\mathcal{C}_1, \mathcal{C}_2).$$

A.11 Usual Intake, Standardization and Transformation

Here we show how to go from the transformed and standardized data to usual intakes. We first consider energy, where we used the transformation

$$Q_{i3k} = \sqrt{2}\{g(Y_{i3k}, \lambda_E) - a_E\}/s_E = g_{\text{tr}}(Y_{i3k}, \lambda_E, a_E, s_E) = \mathbf{X}_{i3}^{\mathrm{T}}\boldsymbol{\beta}_3 + U_{i3} + \epsilon_{i3k}.$$

When $\lambda_E = 0$, the back-transformation is

$$\begin{aligned}g_{\text{tr}}^{-1}(z, 0, a_E, s_E) &= \exp\left\{a_E + s_E z/\sqrt{2}\right\}; \\ \partial^2 g_{\text{tr}}^{-1}(z, 0, a_E, s_E)/\partial z^2 &= \frac{s_E^2}{2}g_{\text{tr}}^{-1}(z, 0).\end{aligned}$$

When $\lambda_E \neq 0$, the back-transformation is

$$g_{\rm tr}^{-1}(z, \lambda_E, a_E, s_E) = \left[1 + \lambda_E \left\{a_E + s_E z/\sqrt{2}\right\}\right]^{1/\lambda_E}; \quad \text{(A.1)}$$

$$\partial^2 g_{\rm tr}^{-1}(z, \lambda_E, a_E, s_E)/\partial z^2 = \frac{s_E^2}{2}(1-\lambda_E)\left[1 + \lambda_E \left\{a_E + s_E z/\sqrt{2}\right\}\right]^{-2+1/\lambda_E} \quad \text{(A.2)}$$

Define

$$g_{\rm tr}^*\{v, \lambda_E, a_E, s_E, \mathbf{\Sigma}_\epsilon(3,3)\} = g_{\rm tr}^{-1}(v, \lambda_E, a_E, s_E) + (1/2)\mathbf{\Sigma}_\epsilon(3,3)\frac{\partial^2 g_{\rm tr}^{-1}(v, \lambda_E, a_E, s_E)}{\partial v^2}.$$

As in Kipnis, et al. (2009), the usual intake of energy for person i is

$$\begin{aligned} T_{Ei} &= E\left\{g_{\rm tr}^{-1}(\mathbf{X}_{i3}^{\rm T}\boldsymbol{\beta}_3 + U_{i3} + \epsilon_{i3}, \lambda_E, a_E, s_E)|\mathbf{X}_{i3}, U_{i3}\right\} \\ &\approx g_{\rm tr}^*\left\{\mathbf{X}_{i3}^{\rm T}\boldsymbol{\beta}_3 + U_{i3}, \lambda_E, a_E, s_E, \mathbf{\Sigma}_\epsilon(3,3)\right\}. \end{aligned}$$

Similarly, a person's usual intake of the dietary component on the original scale is defined as

$$T_{Fi} = \Phi(\mathbf{X}_{i1}^{\rm T}\boldsymbol{\beta}_1 + U_{i1}) g_{\rm tr}^*\left\{\mathbf{X}_{i2}^{\rm T}\boldsymbol{\beta}_2 + U_{i2}, \lambda_F, a_F, s_F, \mathbf{\Sigma}_\epsilon(2,2)\right\}.$$

References

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

Casella, G. and George, E. (1992). Explaining the Gibbs Sampler. *The American Statistician*. 46, 167-174

Chib, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327-335

Davidson, R. and MacKinnon, J. G. (1999). Bootstrap testing in nonlinear models. *International Economic Review*, 40, 487-508.

Delaigle, A., Hall, P. and Meister, A. (2008). On Deconvolution with repeated measurements. *Annals of Statistics*, 36, 665-685.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Annals of Statistics*, 19, 1257-1272.

DOI: 10.2202/1557-4679.1267

Ferrari, P., Kaaks, R., Fahey, M. T., Slimani, N., Day, N. E., Pera, G., Boshuizen, H. C., Roddam, A., Boeing, H., Nagel, G., Thiebaut, A., Orfanos, P., Krogh, P., Braaten, T., and Riboli, E. (2004). Within- and between-cohort variation in measured macronutrient intakes, taking account of measurement errors, in the European Prospective Investigation Into Cancer and Nutrition Study. *American Journal of Epidemiology*, 160, 814-822.

Heckman, J. (1976). The common structure of statistical models of truncation, sample selection, and limited dependent variables and a simple estimator for such models. *Annals of Economics and Social Management*, 5, 475-592.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, 153-161.

Johnson, B. A., Herring, A. H., Ibrahim, J. G., and Siega-Riz, A. M. (2007). Structured measurement error in nutritional epidemiology: Applications in the pregnancy, infection, and nutrition (PIN) study. *Journal of the American Statistical Association*, 102, 856-866.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *Statistical Science*, 52(2), 93-100.

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.

Kipnis, V., Freedman, L. S., Carroll, R. J. and Midthune, D. (2010). A bivariate measurement error model for an episodically consumed dietary component and energy: application to epidemiology. Preprint.

Kyriazidou, E. (1997). Estimation of a panel data sample selection model. *Econometrica*, 65, 1335-1364.

Lehmann, E. L. and Casella, G. (1998). *Theory of Point Estimation*. Springer, New York.

Leung, S.,F. and Yu S. (1996). On the choice between sample selection and two-part models. *Journal of Econometrics*, 72, 197-229.

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

Liang, H. and Wang, N. S. (2005). Partially linear single-index measurement error models. *Statistica Sinica*, 15, 99-116.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: a survey. *Journal of the Iranian Statistical Society*, 1-2, 7-33.

Nusser, S. M., Fuller, W. A., and Guenther, P. M. (1997). Estimating usual dietary intake distributions: Adjusting for measurement error and non-normality in 24-hour food intake data. In Lyberg, L, Biemer, P, Collins, M, Deleeuw, E, Dippo, C, Schwartz, N, and Trewin, D (editors). *Survey Measurement and Process Quality*, pp.670-689, New York: Wiley, 1997.

Olsen, M. K. and Schafer, J. L. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96, 730-745.

Robert, C. P. (1995). Simulation of truncated normal variables. *Statistics and Computing*, 5, 121-125.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., Hurwitz, P. E., Coyle, L., Schussler, N., Michaud, D. S., Freedman, L. S., Brown, C. C., Midthune, D. and Kipnis, V. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: the National Institutes of Health-AARP Diet and Health Study. *American Journal of Epidemiology*, 154, 1119 25.

Sinha, S., Mallick, B. K., Kipnis, V. and Carroll, R. J. (2010). Semiparametric Bayesian analysis of nutritional epidemiology data in the presence of measurement error. *Biometrics*, to appear.

Staudenmayer, J., Ruppert, D., and Buonaccorsi, J. (2008). Density estimation in the presence of heteroskedastic measurement error. *Journal of the American Statistical Association*, 103, 726-736.

Subar, A.F., Dodd, K.W., Guenther, P.M., Kipnis, V., Midthune, D., McDowell, M., Tooze, J.A., Freedman, L.S. and Krebs-Smith, S.M. (2006) The food propensity questionnaire: concept, development, and validation for use as a covariate in a model to estimate usual food intake. *Journal of the American Dietetic Association*,106(10):1556-63.

Tooze, J. A., Grunwald, G. K., and Jones, R. H. (2002). Analysis of repeated measures data clumping at zero. *Statistical Methods in Medical Research* ,11, 341-355.

DOI: 10.2202/1557-4679.1267

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.

Wand, M. P. (1998). Finite sample performance of deconvolving kernel density estimators. *Statistics and Probability Letters*, 37, 131-139.