Bayesian Copula Density Deconvolution for Zero-inflated Data with Applications in Nutritional Epidemiology

Abhra Sarkar

Statistics and Data Sciences
The University of Texas at Austin

@ Conference of Texas Statisticians
Lamar University, Beaumont, Texas
Motivating nutritional epidemiology problem: Assessment of long-term dietary habits

Bayesian semiparametric solution: Multivariate copula density deconvolution
Dietary habits are important for our general wellbeing - unhealthy diets are known to be leading causes of many chronic diseases.

Nutritionists are interested in the distribution of long term average daily intake of different dietary components.

What percentage of Americans are vitamin A deficient?

What is the median consumption of sodium?

What percentage of Americans have a healthy diet? (more complicated)
Nutritionists are interested in the distribution of long term average daily intake of different dietary components.

Long term average intakes can NOT be directly measured.

Data are typically collected in the form of 24 hour recalls.

Data generating model: \( Y_{i,j} = X_i + U_{i,j}, \quad E(U_{i,j} \mid X_i) = 0. \)

Data generating density: \( f_Y(Y) = \int f_X(X)f_{U \mid X}(Y - X)dX. \)

Bayesian hierarchical solution: Model \( f_X \) and \( f_{U \mid X}. \)
Nutritionists are interested in joint consumption patterns.

Data generating model: \( Y_{i,j} = X_i + U_{i,j}, \quad E(U_{i,j} | X_i) = 0. \)

Data generating density: \( f_Y(Y) = \int f_X(X)f_{U|X}(Y - X)dX. \)

Bayesian hierarchical solution: model \( f_X \) and \( f_{U|X}. \)
Deconvolution in Nutritional Epidemiology - Zero-Inflated MVT Data

Subject 24-hour recalls

<table>
<thead>
<tr>
<th></th>
<th>Episodic Component</th>
<th>Regular Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_{e,1,1}$</td>
<td>$Y_{e,1,2}$</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>$Y_{e,2,2}$</td>
</tr>
<tr>
<td>3</td>
<td>$Y_{e,3,1}$</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>$Y_{e,4,2}$</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>n</td>
<td>$Y_{e,n,1}$</td>
<td>$Y_{e,n,2}$</td>
</tr>
</tbody>
</table>

- Some of the components may only be consumed episodically.
- Data generating model: \( Y_{i,j} = X_i + U_{i,j}, \) \( E(U_{i,j} \mid X_i) = 0. \)
- Data generating density: \( f_Y(Y) = \int f_X(X)f_{U \mid X}(Y - X)dX. \)
- Bayesian hierarchical solution: model \( f_X \) and \( f_{U \mid X}. \)
Deconvolution in Nutritional Epidemiology - Univariate Marginals

<table>
<thead>
<tr>
<th>Subject</th>
<th>24-hour recalls</th>
<th>Episodic Component</th>
<th>Regular Component</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$Y_{e,1,1}$</td>
<td>$Y_{e,1,2}$</td>
<td>$Y_{e,1,3}$ $Y_{e,1,4}$</td>
</tr>
<tr>
<td>2</td>
<td>$0$</td>
<td>$Y_{e,2,2}$</td>
<td>$Y_{e,2,3}$ $Y_{e,2,4}$</td>
</tr>
<tr>
<td>3</td>
<td>$Y_{e,3,1}$</td>
<td>$0$</td>
<td>$Y_{e,3,3}$ $Y_{e,3,4}$</td>
</tr>
<tr>
<td>4</td>
<td>$0$</td>
<td>$Y_{e,4,2}$</td>
<td>$Y_{e,4,3}$ $0$</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>n</td>
<td>$Y_{e,n,1}$</td>
<td>$Y_{e,n,2}$</td>
<td>$0$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>$Y_{r,n,1}$ $Y_{r,n,2}$ $Y_{r,n,3}$ $Y_{r,n,4}$</td>
</tr>
</tbody>
</table>

- Data generating model: $Y_{i,j} = X_i + U_{i,j}, \quad E(U_{i,j} \mid X_i) = 0.$
- Data generating density: $f_Y(Y) = \int f_X(X)f_{U\mid X}(Y - X)dX.$
- Bayesian hierarchical solution: model $f_X$ and $f_{U\mid X}.$

- Univariate submodels: $Y_{\ell,i,j} = X_{\ell,i} + U_{\ell,i,j}, \quad E(U_{\ell,i,j} \mid X_{\ell,i}) = 0.$
- Subject specific mean $\overline{Y}_{\ell,i}$ - crude estimate of $X_{\ell,i}$
- Subject specific variance $S_{Y,\ell,i}^2$ - crude estimate of $\text{var}(U_{\ell,i,j} \mid X_{\ell,i})$
EATS Data Example - Sodium and Energy
EATS Data Example - Milk and Whole Grains
Density Deconvolution with Zero-Inflated Data

Classical Multivariate Deconvolution Model

- Data generating model: \( Y_{i,j} = X_i + U_{i,j}, \ E(U_{i,j} \mid X_i) = 0. \)
- Data generating density: \( f_Y(Y) = \int f_X(X) f_{U \mid X}(Y - X) dX. \)
- Bayesian hierarchical solution: model \( f_X \) and \( f_{U \mid X}. \)

Multivariate Deconvolution for Zero-inflated Data

- \( q \) episodic components: \( Y_{\ell,i,j} \) may equal exact zero, \( \ell = 1, \ldots, q. \)
- \( p \) regular components: \( Y_{\ell,i,j} \) is always positive, \( \ell = q + 1, \ldots, q + p. \)

- Exact zeros!
- Conditional heteroscedasticity!
- Boundary discontinuities!
- Probability of reporting positive consumption!

- Challenging to specify a hierarchical model that accommodates exact zeros and other important data features!!!
Latent Variable Framework

- Record the data slightly differently as

\[ Y_{\ell,i,j} = 0 \] (1) when the recall is 0 (when the recall is positive) for \( \ell = 1, \ldots, q \),
\[ Y_{\ell,i,j} = \text{actual reported recalls for } \ell = q + 1, \ldots, 2q, \text{ zero or positive}, \]
\[ Y_{\ell,i,j} = \text{actual reported recalls for } \ell = 2q + 1, \ldots, 2q + p, \text{ always positive}. \]

- Introduce variables \( W_{\ell,i,j} \) such that

\[ Y_{\ell,i,j} = I(W_{\ell,i,j} > 0), \quad \text{for } \ell = 1, \ldots, q, \]
\[ Y_{\ell,i,j} = Y_{\ell-q,i,j} W_{\ell,i,j}, \quad \text{for } \ell = q + 1, \ldots, 2q, \]
\[ Y_{\ell,i,j} = W_{\ell,i,j}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p. \]

- For \( \ell = 1, \ldots, q \), \( W_{\ell,i,j} \) is always latent - it models whether \( Y_{\ell,i,j} = 0 \) or 1.

- For \( \ell = q + 1, \ldots, q + p \), \( W_{\ell,i,j} = Y_{\ell,i,j} \) when \( Y_{\ell,i,j} \) is positive, \( W_{\ell,i,j} \) is latent otherwise.
Latent Variable Framework

- Introduce variables $W_{\ell,i,j}$ such that
  
  \[
  Y_{\ell,i,j} = I(W_{\ell,i,j} > 0), \quad \text{for } \ell = 1, \ldots, q,
  \]
  
  \[
  Y_{\ell,i,j} = Y_{\ell-q,i,j} W_{\ell,i,j}, \quad \text{for } \ell = q + 1, \ldots, 2q,
  \]
  
  \[
  Y_{\ell,i,j} = W_{\ell,i,j}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p.
  \]

- For $\ell = 1, \ldots, q$, $W_{\ell,i,j}$ is always latent - it models whether $Y_{\ell,i,j} = 0$ or 1.
- For $\ell = q + 1, \ldots, q + p$, $W_{\ell,i,j} = Y_{\ell,i,j}$ when $Y_{\ell,i,j}$ is positive, $W_{\ell,i,j}$ is latent otherwise.
- Introduce variables $\tilde{X}_{\ell,i}$ such that
  
  \[
  \tilde{X}_{\ell,i} = h_\ell(X_{\ell,i}), \quad \text{for } \ell = 1, \ldots, q,
  \]
  
  \[
  \tilde{X}_{\ell,i} = X_{\ell-q,i}^+, \quad \text{for } \ell = q + 1, \ldots, 2q,
  \]
  
  \[
  \tilde{X}_{\ell,i} = X_{\ell-q,i}^+, = X_{\ell-q,i}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p.
  \]

- Then use the classical multivariate deconvolution model
  
  \[
  W_{i,j} = \tilde{X}_i + U_{i,j}, \quad \mathbb{E}(U_{i,j} \mid \tilde{X}_i) = 0.
  \]
\[ W_{i,j} = \tilde{X}_i + U_{i,j}, \quad \mathbb{E}(U_{i,j} \mid \tilde{X}_i) = 0 \]

\[ W_{\ell,i,j} = \tilde{X}_{\ell,i} + U_{\ell,i,j} = h_{\ell}(X_{\ell,i}) + U_{\ell,i,j}, \quad \text{for } \ell = 1, \ldots, q, \]

\[ W_{\ell,i,j} = \tilde{X}_{\ell,i} + U_{\ell,i,j} = X_{\ell-q,i}^+ + U_{\ell,i,j}, \quad \text{for } \ell = q + 1, \ldots, 2q, \]

\[ W_{\ell,i,j} = \tilde{X}_{\ell,i} + U_{\ell,i,j} = X_{\ell-q,i} + U_{\ell,i,j}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p. \]

- We have

\[ \mathbb{E}(Y_{\ell,i,j} \mid X_{\ell-q,i}, X_{\ell-q,i}^+) = \Pr(W_{\ell-q,i,j} > 0 \mid X_{\ell-q,i})\mathbb{E}(W_{\ell,i,j} \mid X_{\ell-q,i}^+) \]

\[ = \Pr\{U_{\ell-q,i,j} > h_{\ell-q}(X_{\ell-q,i}) \mid X_{\ell-q,i}\}\mathbb{E}(W_{\ell,i,j} \mid X_{\ell-q,i}^+) \]

\[ = P_{\ell}(X_{\ell-q,i})X_{\ell-q,i}^+ = X_{\ell-q,i}, \quad \text{for } \ell = q + 1, \ldots, 2q, \]

\[ \mathbb{E}(Y_{\ell,i,j} \mid \tilde{X}_{\ell,i}) = \mathbb{E}(W_{\ell,i,j} \mid \tilde{X}_{\ell,i}) = \tilde{X}_{\ell-q,i}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p. \]

- We thus have

\[ W_{\ell,i,j} = h_{\ell}(X_{\ell,i}) + U_{\ell,i,j}, \quad \text{for } \ell = 1, \ldots, q, \]

\[ W_{\ell,i,j} = X_{\ell-q,i}/P_{\ell-q}(X_{\ell-q,i}) + U_{\ell,i,j}, \quad \text{for } \ell = q + 1, \ldots, 2q, \]

\[ W_{\ell,i,j} = X_{\ell-q,i} + U_{\ell,i,j}, \quad \text{for } \ell = 2q + 1, \ldots, 2q + p. \]
Modeling the Density of Interest $f_X(X)$

$$W_{i,j} = \tilde{X}_i + U_{i,j}, \quad \mathbb{E}(U_{i,j} \mid \tilde{X}_i) = 0,$$

Model $f_X(X)$ and $f_{U \mid \tilde{X}}(U \mid \tilde{X})$ and $P(X)$

- Gaussian copula density for $f_X(X)$:
  $$f_X(X) = |R_X|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} Y_X^T (R_X^{-1} - I_{q+p}) Y_X \right\} \prod_{\ell=1}^{q+p} f_{X,\ell}(X_\ell)$$
  where $F_{X,\ell}(X_\ell) = \Phi(Y_{X,\ell})$ for all $\ell$.

- Mixtures of truncated normals for $f_{X,\ell}(X_\ell), \ell = q+1, \ldots, q+p$:
  $$f_{X,\ell}(X_\ell) = \sum_{k=1}^{K_X} \pi_{X,\ell,k} \text{TN}(X_\ell \mid \mu_{X,k}, \sigma_{X,k}^2, [A_\ell, B_\ell]).$$

- Normalized mixtures of B-splines for $f_{X,\ell}(X_\ell), \ell = 1, \ldots, q$:
  $$f_{X,\ell}(X_\ell) = B_{d,\ell,J}(X_\ell) \exp(\xi_\ell) \left\{ \sum_{m=1}^{J_\ell} \delta_{\ell,m} \exp(\xi_{\ell,m}) \right\}^{-1},$$
  Priors: $(\xi_\ell \mid J_\ell, \sigma_{\xi,\ell}^2) \propto (2\pi \sigma_{\xi,\ell}^2)^{-J_\ell / 2} \exp\{-\xi_\ell^T P_\ell \xi_\ell / (2\sigma_{\xi,\ell}^2)\}$, $\sigma_{\xi,\ell}^2 \sim \text{Inv-Ga}(a_\xi, b_\xi)$. 

Modeling the Conditional Density of the Measurement Errors $f_{U|\tilde{X}}$

$$W_{i,j} = \tilde{X}_i + U_{i,j}, \quad \mathbb{E}(U_{i,j} | \tilde{X}_i) = 0,$$

Model $f_X(X)$ and $f_{U|\tilde{X}}(U | \tilde{X})$ and $P(X)$

$$U_{i,j} = S(\tilde{X}_i)\epsilon_{i,j}, \quad \text{with} \quad \mathbb{E}(\epsilon_{i,j}) = 0,$$

and $S(\tilde{X}_i) = \text{diag}\{1, \ldots, 1, s_{q+1}(\tilde{X}_{q+1,i}), \ldots, s_{2q+p}(\tilde{X}_{2q+p,i})\}$.

- Gaussian copula density for $f_{\epsilon}(\epsilon)$:

$$f_{\epsilon}(\epsilon) = \prod_{\ell=1}^{q} f_{\epsilon,\ell}(\epsilon_{\ell}) \times |R_{\epsilon}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} Y_{\epsilon}^T (R_{\epsilon}^{-1} - I_p) Y_{\epsilon} \right\} \prod_{\ell=q+1}^{2q+p} f_{\epsilon,\ell}(\epsilon_{\ell}),$$

where $F_{\epsilon,\ell}(\epsilon_{\ell}) = \Phi(Y_{\epsilon,\ell})$ for all $\ell$ and $\mathbb{E}_{f_{\epsilon,\ell}}(\epsilon_{\ell}) = 0$, for $\ell = 1, \ldots, 2q + p$.

- Single component unit variance normals for $f_{\epsilon,\ell}(\epsilon_{\ell}), \ell = 1, \ldots, q$:

$$f_{\epsilon,\ell}(\epsilon_{\ell}) = \text{Normal}(0, 1).$$

- Mixtures of mean restricted normals for $f_{\epsilon,\ell}(\epsilon_{\ell}), \ell = q + 1, \ldots, q + p$:

$$f_{\epsilon,\ell}(\epsilon_{\ell}) = \sum_{k=1}^{K_{\epsilon}} \pi_{\epsilon,\ell,k} f_{c\epsilon}(\epsilon_{\ell} | p_{\epsilon,k}, \tilde{\mu}_{\epsilon,k}, \sigma_{\epsilon,k,1}^2, \sigma_{\epsilon,k,2}^2),$$

where $f_{c\epsilon}(\epsilon | p, \tilde{\mu}, \sigma_1^2, \sigma_2^2) = \{p \text{ Normal}(\epsilon | \mu_1, \sigma_1^2) + (1-p) \text{ Normal}(\epsilon | \mu_2, \sigma_2^2)\}$

with $p_{\epsilon}\mu_{\epsilon,1} + (1-p_{\epsilon})\mu_{\epsilon,2} = 0.$
Modeling the Variance Functions \( v_{\ell}(\tilde{X}_\ell), \ell = q + 1, \ldots, q + p \)

\[
W_{i,j} = \tilde{X}_i + U_{i,j}, \quad \mathbb{E}(U_{i,j} | \tilde{X}_i) = 0,
\]

Model \( f_X(X) \) and \( f_{U|\bar{X}}(U | \bar{X}) \) and \( P(X) \)

\[
U_{i,j} = S(\tilde{X}_i)\varepsilon_{i,j}, \quad \text{with} \quad \mathbb{E}(\varepsilon_{i,j}) = 0,
\]

and \( S(\tilde{X}_i) = \text{diag}\{1, \ldots, 1, s_{q+1}(\tilde{X}_{q+1,i}), \ldots, s_{2q+p}(\tilde{X}_{2q+p,i})\} \).

- Penalized mixtures of B-splines for \( s^2_{\ell}(\tilde{X}_\ell), \ell = q + 1, \ldots, q + p \):

\[
v_{\ell}(\tilde{X}_\ell) = s^2_{\ell}(\tilde{X}_\ell) = \sum_{j=1}^{J_{\ell}} b_{d,\ell,j}(\tilde{X}_\ell) \exp(\vartheta_{\ell,j}) = B_{d,\ell,J_{\ell}}(\tilde{X}_\ell) \exp(\vartheta_{\ell}).
\]
Modeling the Consumption Probabilities $P_\ell(X_\ell), \ell = 1, \ldots, q$

$W_{i,j} = \tilde{X}_i + U_{i,j}, \quad \mathbb{E}(U_{i,j} \mid \tilde{X}_i) = 0,$

Model $f_X(X)$ and $f_{U \mid \tilde{X}}(U \mid \tilde{X})$ and $P(X)$

$U_{i,j} = S(\tilde{X}_i) \epsilon_{i,j}, \quad \text{with} \quad \mathbb{E}(\epsilon_{i,j}) = 0,$

and $S(\tilde{X}_i) = \text{diag}\{1, \ldots, 1, s_{q+1}(\tilde{X}_{q+1,i}), \ldots, s_{2q+p}(\tilde{X}_{2q+p,i})\}.$

- Probit models for the mixture probabilities $P_\ell(X_\ell), \ell = 1, \ldots, q:$$P_\ell(X_\ell) = \Pr(W_\ell > 0 \mid X_\ell) = \Pr\{U_\ell > -h_\ell(X_\ell) \mid X_\ell\} = \Phi\{h(X_\ell)\}.$

- Penalized mixtures of B-splines for $h_\ell(X_\ell), \ell = 1, \ldots, q:$$h_\ell(X_\ell) = \sum_{j=1}^{J_\ell} b_{d,\ell,j}(X_\ell) \beta_{\ell,j} = B_{d,\ell,J_\ell}(X_\ell) \beta_\ell.$
Modeling the Copula Correlation Matrices $R_X$ and $R_\epsilon$

$$W_{i,j} = \tilde{X}_i + U_{i,j}, \quad \mathbb{E}(U_{i,j} \mid \tilde{X}_i) = 0,$$

Model $f_X(X)$ and $f_{U \mid \tilde{X}}(U \mid \tilde{X})$ and $P(X)$

$$U_{i,j} = S(\tilde{X}_i)\epsilon_{i,j}, \quad \text{with} \quad \mathbb{E}(\epsilon_{i,j}) = 0,$$

and $S(\tilde{X}_i) = \text{diag}\{1, \ldots, 1, s_{q+1}(\tilde{X}_{q+1,i}), \ldots, s_{2q+p}(\tilde{X}_{2q+p,i})\}$.

• Cholesky factorization based models for $R$: $R = VV^T$

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

$v_{1,1} = 1$, $v_{2,1} = b_1$, $v_{2,2} = \sqrt{1 - b_1^2}$,

$v_{3,1} = b_2 \sin \theta_1$, $v_{3,2} = b_2 \cos \theta_1$, $v_{3,3} = \sqrt{1 - b_2^2}$, \ldots

We have $|R| = |V|^2 = \prod_{\ell=2}^{q+p} v_{\ell,\ell}^2 = \prod_{\ell=1}^{q+p-1} (1 - b_\ell^2)$. 
Upper triangle and black lines represent the truth. Lower triangle and red lines show the estimates.
Upper triangle and black lines represent the truth.
Lower triangle and red lines show the estimates.
Simulation Results - Marginal Densities of Interest

Black lines represent the truth. Blue lines are for the copula based model. Red lines are for the model of Zhang, et al. (2011).
Simulation Results - Probabilities of Reporting Positive Intakes

Black lines represent the truth. Blue lines are for the copula based model.
Simulation Results - Adjusted Densities of Interest

Black lines represent the truth. Blue lines are for the copula based model. Red lines are for the model of Zhang, et al. (2011).
Upper triangle and blue lines are for the copula based model. 
Lower triangle and red lines are for the model of Zhang, et al. (2011).
Blue lines are for the copula based model.
Red lines are for the model of Zhang, et al. (2011).
Blue lines are for the copula based model.
Blue lines are for the copula based model.
Red lines are for the model of Zhang, et al. (2011).
Bayesian hierarchical frameworks provide powerful tools for solving complex measurement error problems.

- A model for $X$.
- A model $M$ for everything else given $X$.

Bayesian treatment of measurement error problems is still NOT trivial.

So much more to be done.
Bayesian hierarchical frameworks provide powerful tools for solving complex measurement error problems.

Bayesian treatment of measurement error problems is still NOT trivial.

- Measurement errors can NOT be treated in the same ways as regression errors.
- Methods that are extremely successful in measurement error free scenarios may NOT be useful or appropriate for measurement error problems (e.g. mixtures of truncated normals).
- Conversely, methods that are NOT relevant for measurement error free scenarios may be useful for measurement error problems (e.g. normalized mixtures of splines).
- Methods need to be specifically designed for measurement error problems that accommodate their unique features.

Much more to be done.
Bayesian hierarchical frameworks provide powerful tools for solving complex measurement error problems.

Bayesian treatment of measurement error problems is still NOT trivial.

Much more to be done.

- Deconvolution for data with hard zeros (never consumers)
- Covariate informed density deconvolution
- Multivariate regression in present of measurement errors