Generative models for data with measurement error

Yi Ding supervised by Dr. Wei Huang

Introduction and motivation

- Most statistical methods are designed for error-free data, while lots of real data is measured with errors.
- For example, the U.S. national nutritional surveys traditionally have used the 24-hour recall (24HR) to collect information on food intake.
- This relies heavily on consumers' self-reported behaviours, and often **fails to measure true usual intake of nutrients in the population precisely**.
- Ignoring measurement error can lead to incorrect conclusions, which can negatively affect health policies and treatment guidelines (Figure 2).
- This study aims to develop a **nonparametric approach for modeling diet-outcome relationship that corrects for measurement error**.

Setting

For individual i on day j, $i = 1,..,n$; $j = 1,..,J_i$; let W_{ij} and X_i denote reported 24HR intake and latent individual mean intake, respectively. Let Z_i be a vector of covariates measured without error.

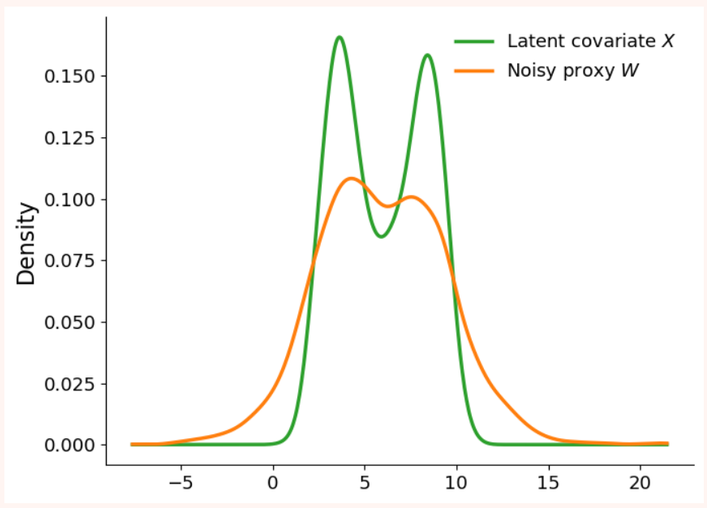


Figure 1. Marginal distribution for W and X

$$W_{ij} = X_i + U_{ij}, \quad \mathbb{E}\left(U_{ij}|X_i\right) = 0. \tag{†}$$

For the classical measurement error model (†), the latent random errors U_{ij} reflect daily intake variations and other sources of errors.

$$\begin{aligned} X_i &= f\left(Z_i, e_{1i}\right) & e_{1i} &\sim N\left(0,1\right) \\ U_{ij} &= g\left(e_{2ij}\right) & e_{2ij} &\sim N\left(0,1\right), \end{aligned}$$

where f and g are unknown functions. **The objective here is to learn the distribution of $X|Z$ and U nonparametrically from the noisy replicated measurements**.

Outcome model

Let Y denote a health outcome possibly related to the **usual intake** $X^* = \mathbb{E}\left(X|Z\right)$ through the regression model

$$Y_i = H\left(X_i^*, Z_i, \epsilon_i\right),$$

where $\epsilon_i \sim N\left(0,1\right)$ is the noise, and we **aim to learn this diet-outcome relationship**.

We observe $(Y_i, W_i, Z_i)_{i=1}^n$ where $W_i = (W_{i1}, ..., W_{iJ_i})^T$.

Naive approach

If we regress Y_i on $\widehat{X_i^*}^{naive} = \overline{W_i} = \frac{1}{J_i}\sum_{j=1}^{J_i} W_{ij}$, then our estimate for H is biased (Figure 2).

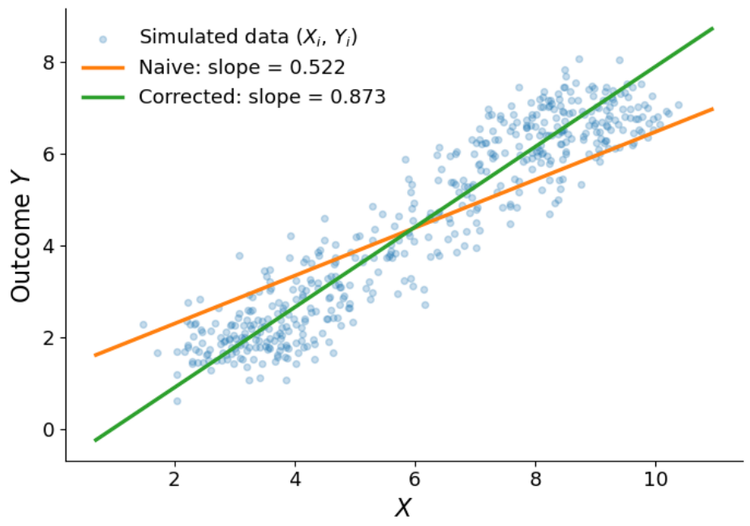


Figure 2. Naive approach v.s. Correcting for measurement error

Here, the underlying true relationship is

$$Y = -0.9 + 0.88\,X^* + \epsilon, \quad \epsilon \sim N(0, 0.4^2).$$

The naive approach grossly underestimates the slope (attenuation bias), whereas our approach that learns the distribution of $X|Z$ produces better estimate.

Identification conditions

The distribution for U is identified through **observed contrasts of replicated measurements**. Contrasts are defined as

$$D_{ijk} = W_{ij} - W_{ik} = U_{ij} - U_{ik} \quad \text{for} \quad j \neq k; j, k \leq J_i.$$

Let the characteristic function for W, $X|Z$, D, and U be denoted as $\varphi_W\left(t\right)$, $\varphi_X\left(t\right)$, $\varphi_D\left(t\right)$, and $\varphi_U\left(t\right)$, respectively.

To ensure **identification for $X|Z$ and U**, we assume

1. U_{ij} are independent of X_i across all i, j
2. U_{ij} are independent identically distributed (i.i.d.) across all i, j; with number of replicates $J_i \geq 2$ for $i = 1, ..., n$
3. The distribution of U is symmetric around 0
4. $\varphi_U\left(t\right) \neq 0$ for all $t \in \mathbb{R}$ (Non-vanishing)

Based on the definition of contrasts, and conditions 2, 3 and 4,

$$\begin{aligned} \varphi_D\left(t\right) &= \varphi_U\left(t\right)\varphi_U\left(-t\right) = \left|\varphi_U\left(t\right)\right|^2 \\ \implies \varphi_U\left(t\right) &= \sqrt{\varphi_D\left(t\right)} \quad \left(\varphi_U\left(0\right) = 1\right). \end{aligned}$$

Based on the error model (†), and conditions 1 and 4,

$$\varphi_W\left(t\right) = \varphi_X\left(t\right)\varphi_U\left(t\right) \implies \varphi_X\left(t\right) = \varphi_W\left(t\right)/\varphi_U\left(t\right).$$

Therefore, the distribution of $X|Z$ and U can be identified through observed D and W.

Methodology

The distribution of U, $X|Z$ and $Y|\left(X^*, Z\right)$ can be recovered through generative models learning using scoring rules.

Energy score loss

The energy score is a **strictly proper** scoring rule that measures the accuracy of probabilistic predictions. Let P denote the true data-generating distribution and Q be a predictive distribution. Let $A \sim P$ and $B, B' \overset{iid}{\sim} Q$. The energy score of Q under P is defined as

$$ES(Q, P) = \underbrace{\mathbb{E}\left\|B - A\right\|}_{\text{prediction error}} - \frac{1}{2}\underbrace{\mathbb{E}\left\|B - B'\right\|}_{\text{prediction spread}}, \tag{‡}$$

where $\|\cdot\|$ denotes the Euclidean norm.

Let $\{A_i\}_{i=1}^n$ be i.i.d. samples from P. For each i, let $\{B_{i1}, \ldots, B_{im}\}$ be i.i.d. samples from Q with $m \geq 2$, independent of A_i. Empirically

$$\begin{aligned} &\widehat{ES}_{n,m}(Q, P) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{m}\sum_{j=1}^{m}\left\|B_{ij} - A_i\right\| - \frac{1}{2m(m-1)}\sum_{j=1}^{m}\sum_{j'=1}^{m}\left\|B_{ij} - B_{ij'}\right\|\right]. \end{aligned}$$

Generative model learning for U

The symmetry property (condition 3) for U is **enforced by construction** to ensure identifiability:

$$g\left(e_2\right) = G\left(e_2\right) - G\left(-e_2\right).$$

Model class: $\mathcal{M}_G = \{G\left(e_2\right)\}$, where $G\left(e_2\right)$ are **parameterized by neural networks**, and optimized by gradient-based methods.

Let D denote the true distribution for observed contrasts, then **the objective loss function** ℓ_{ES}^D is derived from (‡):

$$\begin{aligned} \ell_{ES}^D\left(G, D\right) = \mathbb{E}\Big\|&\left\{g\left(e_2^{(1)}\right) - g\left(e_2^{(2)}\right)\right\} - D\Big\| \\ &- \frac{1}{2}\mathbb{E}\Big\|\left\{g\left(e_2^{(1)}\right) - g\left(e_2^{(2)}\right)\right\} - \left\{g\left(e_2^{(3)}\right) - g\left(e_2^{(4)}\right)\right\}\Big\|, \end{aligned}$$

where $e_2^{(1)}, e_2^{(2)}, e_2^{(3)}, e_2^{(4)} \overset{iid}{\sim} N(0,1)$. Let

$$\tilde{G} \in \arg\min_{G \in \mathcal{M}_G} \ell_{ES}^D(G, D).$$

Under correct model specification, we have

$$\tilde{g}(e_2) = \tilde{G}(e_2) - \tilde{G}(-e_2) \overset{d}{=} U.$$

Generative model learning for $X|Z$ and $Y|\left(X^*, Z\right)$

Similarly, the distribution of $X|Z$ is learned through matching the distribution of W with $f\left(Z, e_1\right)$ from a model class $\mathcal{M}_f$:

$$\begin{aligned} \ell_{ES}^f\left(f, W\right) = \mathbb{E}\Big\|&\left\{f\left(Z, e_1\right) + \tilde{g}(e_2)\right\} - W\Big\| \\ &- \frac{1}{2}\mathbb{E}\Big\|\left\{f\left(Z, e_1\right) + \tilde{g}(e_2)\right\} - \left\{f\left(Z, e_1'\right) + \tilde{g}(e_2')\right\}\Big\|. \end{aligned}$$

After learning a generator $\tilde{f} \in \arg\min_{f \in \mathcal{M}_f} \ell_{ES}^f(f, W)$, estimates of $\widehat{X^*}$ can be obtained through Monte Carlo sampling from $\tilde{f}$.

Finally, a generator $\tilde{H}$ for the full distribution of $Y|\left(X^*, Z\right)$ is learned in the same way, by matching the distribution of Y with $H\left(\widehat{X^*}, Z, \epsilon\right)$.

Results and discussion

We set up our simulation with **nonlinear relationships** designed to reflect realistic dietary intake studies. For the purpose of illustration, Z is simulated as a univariate variable.

Latent intake. The latent individual mean intake X depends on Z nonlinearly with additive noise:

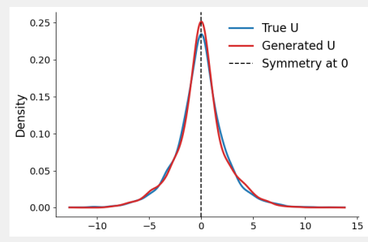
$$X_i = 3\tanh(Z_i/3) + 6 + 0.7e_{1i}, \qquad e_{1i} \sim N\left(0,1\right).$$

Measurement error. Observed intakes are contaminated by Laplace measurement errors:

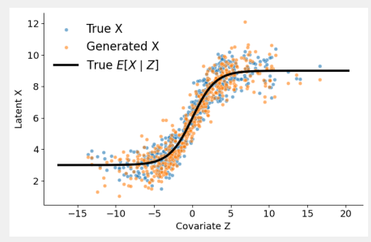
$$W_{ij} = X_i + U_{ij}, \qquad U_{ij} \sim \text{Laplace}\left(0, 2.5/\sqrt{2}\right).$$

Outcome. The health outcome Y depends on usual intake X^* through a nonlinear cubic relationship,

$$Y_i = 20 - 0.2\,X_i^* + \left(X_i^* - 6\right)^3 + 4\epsilon_i, \qquad \epsilon_i \sim N(0,1).$$

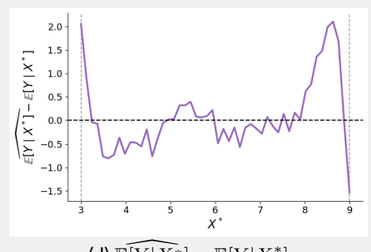


(a) Generated U



(b) Generated $X|Z$

(c) Generated $Y|X^*$



(d) $\widehat{\mathbb{E}[Y|X^*]} - \mathbb{E}[Y|X^*]$

Figure 3. Diagnostics: (a) Measurement error U. (b) Latent intake $X|Z$ and its conditional mean X^*. (c) $Y|X^*$ learned by $\tilde{H}$. (d) Difference between estimated and true conditional mean.

- Based on the diagnostics, our generative models learning approach seems to **recover the full distribution of U, $X|Z$, and $Y|X^*$ well** from the generated samples.
- However, the predictive distribution $\tilde{H}\left(\widehat{X^*}, Z, \epsilon\right)$ slightly **deviates from the true distribution near the boundary of** $\text{supp}(X^*)$ (Figure 3(d)).
- A potential explanation is that recovery of X^* near the boundary is intrinsically more challenging since X^* is inferred indirectly, and the model has no prior knowledge for $\text{supp}(X^*)$. The errors in estimation of X^* then propagate through to $\tilde{H}$.
- We observed **mild oscillation in the training loss** across epochs. This variability is expected in generative models learning where stochastic sampling and noisy gradient estimates can introduce fluctuation, despite stable overall training.

Conclusion and further work

- This study contributes as a stepping stone towards **nonparametric modeling for data with classical measurement error using scoring rules**.
- There are **limitations**, namely deviation of estimates near boundary values and oscillation in training loss across epochs.
- The current approach is also **restrictive** as it requires conditions 1, 2, and 3 to hold for identification, **i.e., the measurement error for reported intake often depends on the mean intake, and may have a skewed distribution, violating the classical measurement error model** (†).
- The most common fix has been to monotonically transform W_{ij} to values $W_{ij}^* = T(W_{ij})$ that more closely follow the classical measurement error model.
- Further work would involve extending the existing nonparametric approach to **non-classical measurement error settings** that generalize better to real-world data.

Acknowledgments and references

All simulations were conducted using a self-modified version of the *engression* package (Shen & Meinshausen, 2024).

[1] V Kipnis, D Midthune, DW Buckman, KW Dodd, PM Guenther, SM Krebs-Smith, AF Subar, JA Tooze, RJ Carroll, and LS Freedman. Modeling data with excess zeros and measurement error: application to evaluating relationships between episodically consumed foods and health outcomes. *Biometrics*, 65(4):1003–1010, 12 2009. doi: 10.1111/j.1541-0420.2009.01223.x.

[2] Xinwei Shen and Nicolai Meinshausen. Engression: extrapolation through the lens of distributional regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, page qkae108, 11 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae108. URL https://doi.org/10.1093/jrsssb/qkae108.

[3] GY Yi, A Delaigle, and P Gustafson. *Handbook of Measurement Error Models*. 1st edition, 2021. doi: 10.1201/9781315101279. URL https://doi.org/10.1201/9781315101279.

yiding3@student.unimelb.edu.au