

CAUSAL INFERENCE WITH ORDINAL OUTCOMES

DENSITY ESTIMATION BASED APPROACH

CHANHYUK PARK

WASHINGTON UNIVERSITY IN ST. LOUIS

MOTIVATION: THE DISCRETIZATION PROBLEM

- Experiments to identify causal effects of treatments
- Outcomes are usually believed to be continuous in unidimensional space
 - Approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009)
 - Policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005; Wu, 2022)
- The Problem - How we measure it

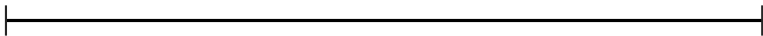
MOTIVATION: THE DISCRETIZATION PROBLEM

- Experiments to identify causal effects of treatments
- Outcomes are usually believed to be continuous in unidimensional space
 - Approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009)
 - Policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005; Wu, 2022)
- The Problem - How we measure it

MOTIVATION: THE DISCRETIZATION PROBLEM

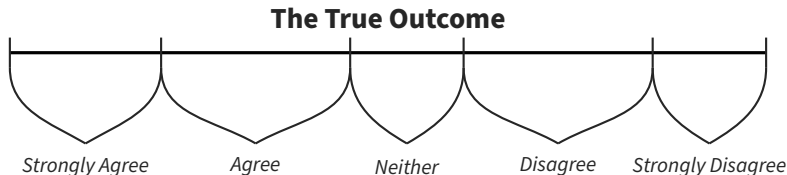
- Experiments to identify causal effects of treatments
- Outcomes are usually believed to be continuous in unidimensional space
 - Approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009)
 - Policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005; Wu, 2022)
- The Problem - How we measure it

The True Outcome



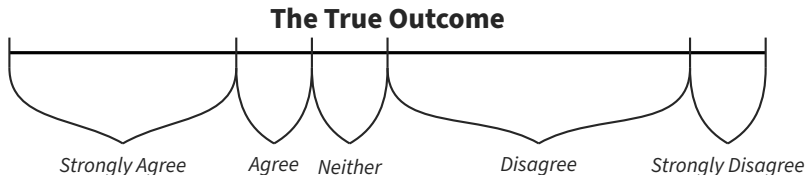
MOTIVATION: THE DISCRETIZATION PROBLEM

- Experiments to identify causal effects of treatments
- Outcomes are usually believed to be continuous in unidimensional space
 - Approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009)
 - Policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005; Wu, 2022)
- The Problem - How we measure it



MOTIVATION: THE DISCRETIZATION PROBLEM

- Experiments to identify causal effects of treatments
- Outcomes are usually believed to be continuous in unidimensional space
 - Approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009)
 - Policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005; Wu, 2022)
- The Problem - How we measure it



THE GOAL

- **Identification:**

Naive causal identification with these "index" may fail

⇒ Normalized Latent Treatment Effect (NLTE)

- **Estimation:**

Standard parametric regression is rigid

⇒ Flexible, density estimation based estimators

THE GOAL

- **Identification:**

Naive causal identification with these "index" may fail

⇒ Normalized Latent Treatment Effect (NLTE)

- **Estimation:**

Standard parametric regression is rigid

⇒ Flexible, density estimation based estimators

POTENTIAL OUTCOME FRAMEWORK

- Denote the binary treatment status of individual $D_i \in \{0, 1\}$
- Denote the potential outcomes in latent preference scale as $Y_i^*(D_i)$
 - $Y_i^*(1) = \tau + f(\beta, X_i) + \epsilon_i$
 - $Y_i^*(0) = f(\beta, X_i) + \epsilon_i$
- Denote the potential outcomes in observed ordinal scale (POO) as $Y_i(d_i) = g(Y_i^*(d_i)) \in \{1, \dots, j, \dots, J\}$
- We are interested in the average treatment effect in the latent space (LTE)

$$\begin{aligned}LTE &= \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] \\ &= \tau\end{aligned}$$

POTENTIAL OUTCOME FRAMEWORK

- Denote the binary treatment status of individual $D_i \in \{0, 1\}$
- Denote the potential outcomes in latent preference scale as $Y_i^*(D_i)$
 - $Y_i^*(1) = \tau + f(\beta, X_i) + \epsilon_i$
 - $Y_i^*(0) = f(\beta, X_i) + \epsilon_i$
- Denote the potential outcomes in observed ordinal scale (POO) as $Y_i(d_i) = g(Y_i^*(d_i)) \in \{1, \dots, j, \dots, J\}$
- We are interested in the average treatment effect in the latent space (LTE)

$$\begin{aligned}LTE &= \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] \\ &= \tau\end{aligned}$$

POTENTIAL OUTCOME FRAMEWORK

- Denote the binary treatment status of individual $D_i \in \{0, 1\}$
- Denote the potential outcomes in latent preference scale as $Y_i^*(D_i)$
 - $Y_i^*(1) = \tau + f(\beta, X_i) + \epsilon_i$
 - $Y_i^*(0) = f(\beta, X_i) + \epsilon_i$
- Denote the potential outcomes in observed ordinal scale (POO) as $Y_i(d_i) = g(Y_i^*(d_i)) \in \{1, \dots, j, \dots, J\}$
- We are interested in the average treatment effect in the latent space (LTE)

$$\begin{aligned}LTE &= \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] \\ &= \tau\end{aligned}$$

POTENTIAL OUTCOME FRAMEWORK

- Denote the binary treatment status of individual $D_i \in \{0, 1\}$
- Denote the potential outcomes in latent preference scale as $Y_i^*(D_i)$
 - $Y_i^*(1) = \tau + f(\beta, X_i) + \epsilon_i$
 - $Y_i^*(0) = f(\beta, X_i) + \epsilon_i$
- Denote the potential outcomes in observed ordinal scale (POO) as $Y_i(d_i) = g(Y_i^*(d_i)) \in \{1, \dots, j, \dots, J\}$
- We are interested in the average treatment effect in the latent space (LTE)

$$\begin{aligned}LTE &= \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] \\ &= \tau\end{aligned}$$

POTENTIAL OUTCOME FRAMEWORK

- Denote the binary treatment status of individual $D_i \in \{0, 1\}$
- Denote the potential outcomes in latent preference scale as $Y_i^*(D_i)$
 - $Y_i^*(1) = \tau + f(\beta, X_i) + \epsilon_i$
 - $Y_i^*(0) = f(\beta, X_i) + \epsilon_i$
- Denote the potential outcomes in observed ordinal scale (POO) as $Y_i(d_i) = g(Y_i^*(d_i)) \in \{1, \dots, j, \dots, J\}$
- We are interested in the average treatment effect in the latent space (LTE)

$$\begin{aligned}LTE &= \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] \\ &= \tau\end{aligned}$$

POTENTIAL OUTCOME FRAMEWORK

- Denote the binary treatment status of individual $D_i \in \{0, 1\}$
- Denote the potential outcomes in latent preference scale as $Y_i^*(D_i)$
 - $Y_i^*(1) = \tau + f(\beta, X_i) + \epsilon_i$
 - $Y_i^*(0) = f(\beta, X_i) + \epsilon_i$
- Denote the potential outcomes in observed ordinal scale (POO) as $Y_i(d_i) = g(Y_i^*(d_i)) \in \{1, \dots, j, \dots, J\}$
- We are interested in the average treatment effect in the latent space (LTE)

$$\begin{aligned}LTE &= \mathbb{E}[Y^*(1)] - \mathbb{E}[Y^*(0)] \\ &= \tau\end{aligned}$$

LTE WITH $Y_i(D_i)$

- If Y_i^* is known, there is not problem
- If g is known and it maps one value of Y_i^* to Y_i , then we may get LTE, but this not true in most settings

$$\mathbb{E}\left[g^{-1}(Y_i(1))\right] - \mathbb{E}\left[g^{-1}(Y_i(0))\right] \neq LTE$$

- Should we give up?

UP TO SCALE IDENTIFICATION

- We can identify LTE up to scale (Theorem 1)
- From the model and data we know:

$$\mathbb{P}(Y(D) \leq j) = \mathbb{P}(Y^* \leq \alpha_j)$$

- We can construct MLE with using this probability.
- However, if we scale Y^* side by a constant, $c > 0$, we still get the same probability

$$\begin{aligned}\mathbb{P}(Y(D) \leq j) &= \mathbb{P}(Y^* \leq \alpha_j) \\ &= \mathbb{P}(cY^* \leq c\alpha_j)\end{aligned}$$

UP TO SCALE IDENTIFICATION

- We can identify LTE up to scale (Theorem 1)
- From the model and data we know:

$$\mathbb{P}(Y(D) \leq j) = \mathbb{P}(Y^* \leq \alpha_j)$$

- We can construct MLE with using this probability.
- However, if we scale Y^* side by a constant, $c > 0$, we still get the same probability

$$\begin{aligned}\mathbb{P}(Y(D) \leq j) &= \mathbb{P}(Y^* \leq \alpha_j) \\ &= \mathbb{P}(cY^* \leq c\alpha_j)\end{aligned}$$

UP TO SCALE IDENTIFICATION

- We can identify LTE up to scale (Theorem 1)
- From the model and data we know:

$$\mathbb{P}(Y(D) \leq j) = \mathbb{P}(Y^* \leq \alpha_j)$$

- We can construct MLE with using this probability.
- However, if we scale Y^* side by a constant, $c > 0$, we still get the same probability

$$\begin{aligned}\mathbb{P}(Y(D) \leq j) &= \mathbb{P}(Y^* \leq \alpha_j) \\ &= \mathbb{P}(cY^* \leq c\alpha_j)\end{aligned}$$

UP TO SCALE IDENTIFICATION

- We can identify LTE up to scale (Theorem 1)
- From the model and data we know:

$$\mathbb{P}(Y(D) \leq j) = \mathbb{P}(Y^* \leq \alpha_j)$$

- We can construct MLE with using this probability.
- However, if we scale Y^* side by a constant, $c > 0$, we still get the same probability

$$\begin{aligned}\mathbb{P}(Y(D) \leq j) &= \mathbb{P}(Y^* \leq \alpha_j) \\ &= \mathbb{P}(cY^* \leq c\alpha_j)\end{aligned}$$

NORMALIZATION

- Up to scale identification is disappointing
- Proper normalization helps interpretation and comparison across models and across samples
- Fixing $Var(\varepsilon_i) = 1$ is one way
- This works as mapping Y_i^* to the space where the variance of the error is a unit / probit space

NORMALIZATION

- Up to scale identification is disappointing
- Proper normalization helps interpretation and comparison across models and across samples
- Fixing $Var(\varepsilon_i) = 1$ is one way
- This works as mapping Y_i^* to the space where the variance of the error is a unit / probit space

NORMALIZATION

- Up to scale identification is disappointing
- Proper normalization helps interpretation and comparison across models and across samples
- Fixing $Var(\varepsilon_i) = 1$ is one way
- This works as mapping Y_i^* to the space where the variance of the error is a unit / probit space

NORMALIZATION

- Up to scale identification is disappointing
- Proper normalization helps interpretation and comparison across models and across samples
- Fixing $Var(\varepsilon_i) = 1$ is one way
- This works as mapping Y_i^* to the space where the variance of the error is a unit / probit space

COMMON TRANSFORMATION: CARDINALIZATION AND BINARIZATION

- Cardinalization
 - Relies on luck
- Binarization
 - Lose significant efficiency
 - Estimate may be sensitive to how you group the ordinal outcomes

SMALL EXAMPLE

	$y_i^*(0)$	$y_i^*(1)$
A	1.37	0.09
B	-0.56	1.71
C	0.36	0.11
D	0.63	2.22
E	0.40	0.14

- The LTE in the latent preference space is: 0.41

SMALL EXAMPLE – CARDINALIZATION

- Suppose two transformation functions: g_a and g_b
- Both are monotone, but with different thresholds (α_j)

	g_a		g_b	
	$Y_{i,g_a}(0)$	$Y_{i,g_a}(1)$	$Y_{i,g_b}(0)$	$Y_{i,g_b}(1)$
A	Agree	Agree	Agree	Disagree
B	Neither	Agree	Disagree	Agree
C	Agree	Agree	Neither	Disagree
D	Agree	Strongly Agree	Neither	Agree
E	Agree	Agree	Neither	Disagree

SMALL EXAMPLE – CARDINALIZATION

- A researcher impose common cardinalization of 1 to 5,
- In case the true transformation is g_a , the estimate is -0.4
- In case the true transformation is g_b , the estimate is 0.2
- Without compelling reason to impose such numeric labels, cardinalization relies on pure luck

ESTIMATION

- There has been tools for ordinal outcomes
- Standard parametric approaches such as Ordered Probit and Ordered Logit
- Two estimators based on density estimation techniques

ORDERED LOGIT AND PROBIT

- Assume that the error term follows a specific distribution (Logistic and Standard Normal)
- And use maximum likelihood to estimate the coefficients
- Identify coefficients up to scale
- Once distributional assumption is violated, inconsistent

ORDERED LOGIT AND PROBIT

- The true log-likelihood is:

$$\ell(\alpha, \beta, \tau) = \sum_{i=1}^n \sum_{j=0}^J 1_{\{Y_i=j\}} \left\{ \log F(\alpha_j - f(X_i, \beta) + D_i^\top \tau) - \log F(\alpha_{j-1} - f(X_i, \beta) + D_i^\top \tau) \right\}$$

- F is the CDF of the true error.
- If F is not standard normal or standard logistic, MLE is inconsistent.
- The error is never known, and omitted or unobserved confounder may also distort the error

ALTERNATIVE: ESTIMATE THE F

- Instead of assuming a specific F , we can estimate as \hat{F}
- Then the log-likelihood becomes:

$$\hat{\ell}(\alpha, \beta, \tau) = \sum_{i=1}^n \sum_{j=0}^J \mathbf{1}_{\{Y_i=j\}} \left\{ \log \hat{F}(\alpha_j - f(X_i, \beta) + D_i^\top \tau) \right. \\ \left. - \log \hat{F}(\alpha_{j-1} - f(X_i, \beta) + D_i^\top \tau) \right\}$$

- I propose to use two density estimation methods: KDE and Normalizing Flows

KERNEL DENSITY ESTIMATION

- Nonparametric method
- Smooth each observation using a kernel (usually Gaussian)

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right)$$

- Generally, the moderate bandwidth ensures enough convergence rate for \sqrt{n} consistency for the main estimator (semiparametric efficiency)

NORMALIZING FLOWS

- A flexible class of models that transform complex continuous density to simple base density (usually Gaussian)
- Use the change-of-variable formula

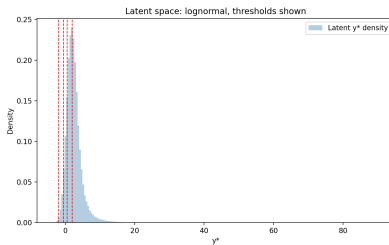
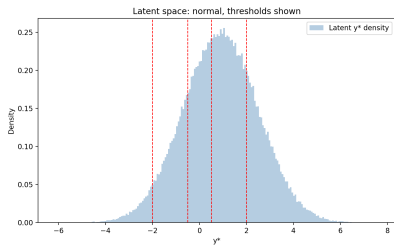
$$f_{\theta}(h) = f_Z(T_{\theta}^{-1}(h)) \left| \det dT_{\theta}^{-1}(h) \right|,$$

- T_{θ} is a set of *invertible* transformation
- Rational Quadratic Spline Flow
- Since this is fully parametric, for finite θ , MLE ensures \sqrt{n} consistency

SIMULATION

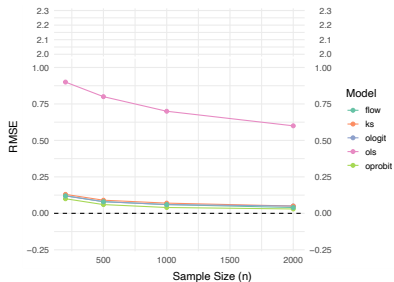
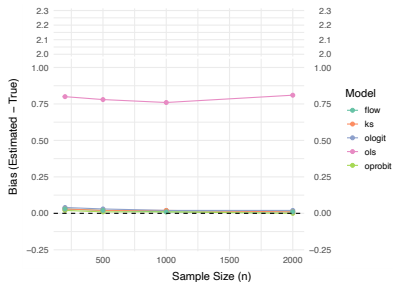
- Ordered Probit Ordered Logit and NF based only
- Errors: Standard Normal and Log Normal distribution
- A binary treatment randomized
- Covariates: One binary and three continuous (following normal distribution)
- Thresholds: $-2, -1, 1, 2.5 \rightarrow 5$ categories
- Sample sizes: 200, 500, 1000, 2000
- 200 Replications

SIMULATION



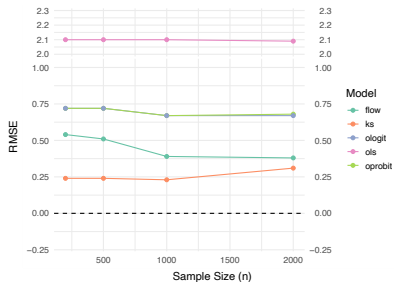
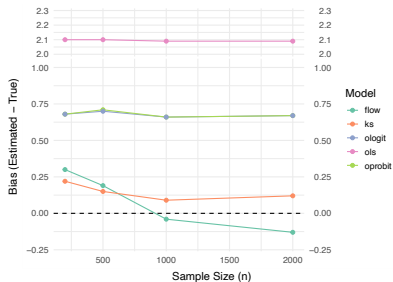
SIMULATION – NORMAL CASE

- True treatment size: 1.0



SIMULATION – LOGNORMAL CASE

- True treatment size: 0.46



CONCLUSION

- With ordinal outcomes, treatment effects can only be identified up to scale
- Cardinalization and Binarization have its pitfalls
- Standard parametric approaches relies on strong distributional assumption
- KDE and NF based approaches may provide flexible ways to estimate the true effects
- For KDE, the normalization of coefficients is an issue...
- For NF, Hyperparameter selection is an issue, trying to automatically adjust it while traing