

Random Forest Estimation of the Ordered Choice Model

Michael Lechner & Gabriel Okasa

SEW-HSG

Swiss Institute for Empirical Economic Research
University of St.Gallen, Switzerland

Young Swiss Economist Meeting, ETH Zürich

June 20, 2020



Introduction

Literature

Ordered Forest

Monte Carlo

Empirical Application

Conclusion



Introduction

Ordered Outcomes

- ▶ categorical dependent variable with inherent ordering
- ▶ example: wine quality
 1. very good
 2. good
 3. neutral
 4. bad
 5. very bad
- ▶ many other examples such as education level, income level, opinion surveys, ratings, sport outcomes, ...
- ▶ ordered nature should be taken into account



Introduction

Parametric Models

- ▶ ordered probit & ordered logit
- ▶ assumptions about the distribution of the error term
- ▶ estimation usually via maximum likelihood

Quantities of Interest:

- ▶ choice probabilities:

$$P[Y_i = m \mid X_i = x]$$

- ▶ marginal effects:

$$\frac{\partial P[Y_i = m \mid X_i = x]}{\partial x^k}$$



Introduction

Ordered Forest

- ▶ estimation of ordered choice model based on the random forest algorithm first developed by Breiman (2001)
- ▶ improves on *parametric* models by allowing for:
 - ▶ flexible functional form
- ▶ improves on *nonparametric* models by allowing for:
 - ▶ larger covariate space
- ▶ alternative to standard ordered probit and ordered logit with:
 - ▶ conditional choice probabilities
 - ▶ marginal effects
 - ▶ approximate inference



Literature

Machine Learning for Ordered Outcomes

Tree-based methods:

- ▶ focusing mainly on ordered classification (Kramer et al. (2001), Piccarreta (2008), Pierola et al. (2016))
- ▶ probabilities **X**, marginal effects **X**, inference **X**

Forest-based methods:

- ▶ conditional forest by Hothorn et al. (2006)
- ▶ probabilities ✓, marginal effects **X**, inference **X**
- ▶ ordinal forest by Hornung (2019)
- ▶ probabilities ✓, marginal effects **X**, inference **X**



Random Forests

Recap

- ▶ ensemble of trees based on bootstrap aggregation
- ▶ only a random subset of covariates considered at each split
- ▶ weighting perspective:

$$\hat{R}F^B(x) = \sum_{i=1}^N \hat{w}_i(x) Y_i,$$

where the weights are defined as

$$\hat{w}_{b,i}(x) = \frac{\mathbf{1}(\{X_i \in L_b(x)\})}{|L_b(x)|} \quad \text{with} \quad \hat{w}_i(x) = \frac{1}{B} \sum_{b=1}^B \hat{w}_{b,i}(x).$$



Ordered Forest

Choice Probabilities

- ▶ consider ordered outcome variable $Y_i \in \{1, \dots, M\}$
- ▶ the estimation procedure is then defined as:

1. create $M - 1$ indicator variables such as

$$Y_{m,i} = \mathbf{1}(Y_i \leq m) \quad \text{for} \quad m = 1, \dots, M - 1$$

2. estimate random forest for each of the $M - 1$ indicators
3. obtain predictions, i.e. cumulative probabilities

$$\hat{Y}_{m,i} = \hat{P}[Y_{m,i} = 1 \mid X_i = x] = \sum_{i=1}^N \hat{w}_{m,i}(x) Y_{m,i}$$



Ordered Forest

Choice Probabilities

4. probabilities for each category are computed as follows

$$\hat{P}_{m,i} = \hat{Y}_{m,i} - \hat{Y}_{m-1,i} \quad \text{for} \quad m = 2, \dots, M$$

where

$$\hat{Y}_{M,i} = 1 \quad \text{and} \quad \hat{P}_{1,i} = \hat{Y}_{1,i}$$

and

$$\hat{P}_{m,i} = 0 \quad \text{if} \quad \hat{P}_{m,i} < 0$$

$$\hat{P}_{m,i} = \frac{\hat{P}_{m,i}}{\sum_{m=1}^M \hat{P}_{m,i}} \quad \text{for} \quad m = 1, \dots, M$$



Ordered Forest

Marginal Effects

- ▶ marginal effect for an element x^k of X :
- ▶ for continuous variables as

$$ME_i^{k,m}(x) = \frac{\partial P[Y_i = m \mid X_i^k = x^k, X_i^{-k} = x^{-k}]}{\partial x^k}$$

- ▶ for binary (categorical) variables as

$$\hat{ME}_i^{k,m}(x) = P[Y_i = m \mid X_i^k = 1, X_i^{-k} = x^{-k}] - P[Y_i = m \mid X_i^k = 0, X_i^{-k} = x^{-k}]$$



Ordered Forest

Marginal Effects

- ▶ marginal effect for an element x^k of X :
- ▶ for continuous variables as

$$ME_i^{k,m}(x) = \frac{\partial P[Y_i = m \mid X_i^k = x^k, X_i^{-k} = x^{-k}]}{\partial x^k}$$

- ▶ for binary (categorical) variables as

$$\hat{ME}_i^{k,m}(x) = P[Y_i = m \mid X_i^k = 1, X_i^{-k} = x^{-k}] - P[Y_i = m \mid X_i^k = 0, X_i^{-k} = x^{-k}]$$



Ordered Forest

Marginal Effects

- ▶ marginal effect for an element x^k of X :
- ▶ for continuous variables as

$$ME_i^{k,m}(x) = \frac{\partial P[Y_i = m \mid X_i^k = x^k, X_i^{-k} = x^{-k}]}{\partial x^k}$$

- ▶ for binary (categorical) variables as

$$\hat{ME}_i^{k,m}(x) = P[Y_i = m \mid X_i^k = 1, X_i^{-k} = x^{-k}] - P[Y_i = m \mid X_i^k = 0, X_i^{-k} = x^{-k}]$$



Ordered Forest

Marginal Effects

- ▶ estimate the marginal effect as numeric approximation

$$\hat{ME}_i^{k,m}(x) = \frac{1}{2h} \left\{ \hat{P}[Y_i = m \mid X_i^k = x^k + h, X_i^{-k} = x^{-k}] - \hat{P}[Y_i = m \mid X_i^k = x^k - h, X_i^{-k} = x^{-k}] \right\}$$

where h is 0.1 standard deviation of x^k

- ▶ marginal effect at the mean
 - ▶ evaluate $\hat{ME}_i^{k,m}(x)$ at the population (sample) mean
- ▶ mean marginal effect
 - ▶ $\frac{1}{N} \sum_{i=1}^N \hat{ME}_i^{k,m}(x)$



Ordered Forest

Marginal Effects

- ▶ estimate the marginal effect as numeric approximation

$$\hat{ME}_i^{k,m}(x) = \frac{1}{2h} \left\{ \hat{P}[Y_i = m \mid X_i^k = x^k + h, X_i^{-k} = x^{-k}] - \hat{P}[Y_i = m \mid X_i^k = x^k - h, X_i^{-k} = x^{-k}] \right\}$$

where h is 0.1 standard deviation of x^k

- ▶ marginal effect at the mean
 - ▶ evaluate $\hat{ME}_i^{k,m}(x)$ at the population (sample) mean
- ▶ mean marginal effect
 - ▶ $\frac{1}{N} \sum_{i=1}^N \hat{ME}_i^{k,m}(x)$



Ordered Forest

Approximate Inference

- ▶ Wager and Athey (2018) prove consistency and asymptotic normality of the random forest predictions
 - ▶ subsampling
 - ▶ honesty
- ▶ weight-based inference as proposed by Lechner (2018)
- ▶ use forest weights for deriving the variance of the estimator
- ▶ crucial condition:
 - ▶ weights and outcomes must be independent → sample splitting
 - ▶ requiring honest forest instead of honest trees only



Monte Carlo

Simulation Design

- ▶ General:
 - ▶ DGP simulated as ordered logit
 - ▶ train $N = \{200, 800, 3\ 200\}$ & test $N = 10\ 000$
 - ▶ replications $R = 100$
 - ▶ 72 different DGPs in total
- ▶ Outcomes:
 - ▶ number of classes: $\{3, 6, 9\}$
 - ▶ equally spaced thresholds vs. randomly spaced thresholds
- ▶ Covariates:
 - ▶ low vs. high dimensional
 - ▶ independent vs. correlated
 - ▶ linear vs. nonlinear effects



Monte Carlo

Simulation Design

Considered Methods:

- ▶ Ordered Logit (McCullagh, 1980)
- ▶ Conditional Forest (Hothorn et al., 2006)
- ▶ Ordinal Forest (Hornung, 2019): naive & optimized
- ▶ **Ordered Forest**: standard & honest
- ▶ Multinomial Forest: standard & honest



Monte Carlo

Simulation Results: Complex DGP & Low Dimension

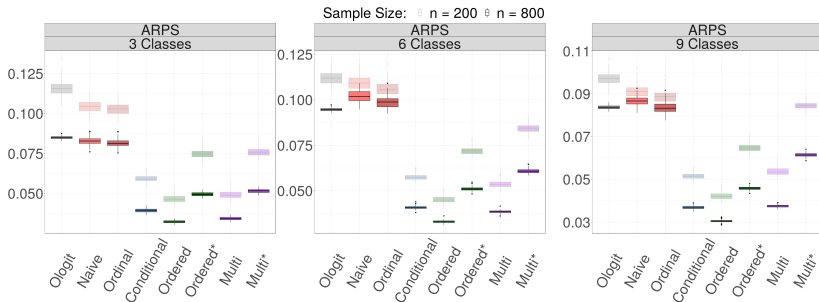


Figure 1: Simulation Results: Average Ranked Probability Score



Monte Carlo

Simulation Results: Complex DGP & Low Dimension

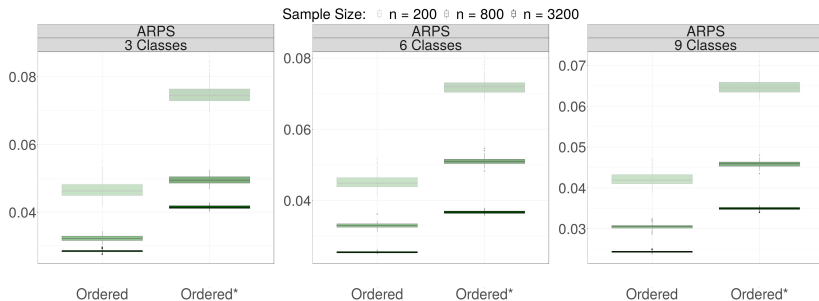


Figure 2: Simulation Results: Average Ranked Probability Score



Empirical Application

Wine Quality Data

Data:

- ▶ quality of wine on an ordered scale from 1 to 6
- ▶ 4893 observations & 11 covariates

Comparisons:

- ▶ ordered forest vs. ordered logit
- ▶ marginal effects estimation
 - ▶ summability of effects
 - ▶ single crossing property



Empirical Application

Mean Marginal Effects

Dataset		Ordered Forest				Ordered Logit					
Variable	Class	Effect	Std.Error	t-Value	p-Value	Effect	Std.Error	t-Value	p-Value		
alcohol	1	-0.0010	0.0011	-0.9028	0.3667	-0.0017	0.0005	-3.5943	0.0003	***	
	2	-0.0036	0.0028	-1.2824	0.1997	-0.0125	0.0023	-5.4096	0.0000	***	
	3	-0.0581	0.0063	-9.2777	0.0000	***	-0.0612	0.0105	-5.8009	0.0000	***
	4	0.0215	0.0067	3.2172	0.0013	***	0.0163	0.0031	5.2566	0.0000	***
	5	0.0350	0.0073	4.8124	0.0000	***	0.0450	0.0077	5.8232	0.0000	***
	6	0.0062	0.0088	0.7074	0.4793		0.0141	0.0026	5.4111	0.0000	***
chlorides	1	0.0033	0.0047	0.7097	0.4779	0.0023	0.0055	0.4147	0.6784		
	2	0.1416	0.0600	2.3600	0.0183	**	0.0166	0.0398	0.4167	0.6769	
	3	1.1575	0.4107	2.8184	0.0048	***	0.0811	0.1945	0.4169	0.6768	
	4	0.1198	0.5269	0.2273	0.8202		-0.0216	0.0518	-0.4175	0.6763	
	5	-1.3265	0.4174	-3.1778	0.0015	***	-0.0596	0.1431	-0.4166	0.6770	
	6	-0.0957	0.4908	-0.1949	0.8454		-0.0187	0.0449	-0.4166	0.6769	
total.sulfur.dioxide	1	0.0000	0.0000	2.0716	0.0383	**	0.0000	0.0000	0.8906	0.3732	
	2	-0.0000	0.0000	-0.9400	0.3472		0.0000	0.0000	0.9069	0.3645	
	3	0.0001	0.0001	1.2302	0.2186		0.0001	0.0001	0.9101	0.3628	
	4	-0.0001	0.0001	-1.0828	0.2789		-0.0000	0.0000	-0.9086	0.3636	
	5	0.0000	0.0001	0.1169	0.9069		-0.0001	0.0001	-0.9095	0.3631	
	6	0.0000	0.0001	0.2671	0.7894		-0.0000	0.0000	-0.9078	0.3640	

Significance levels correspond to: ***. < 0.01, **. < 0.05, *. < 0.1.



Empirical Application

Mean Marginal Effects

Dataset		Ordered Forest				Ordered Logit					
Variable	Class	Effect	Std.Error	t-Value	p-Value	Effect	Std.Error	t-Value	p-Value		
alcohol	1	-0.0010	0.0011	-0.9028	0.3667	-0.0017	0.0005	-3.5943	0.0003	***	
	2	-0.0036	0.0028	-1.2824	0.1997	-0.0125	0.0023	-5.4096	0.0000	***	
	3	-0.0581	0.0063	-9.2777	0.0000	***	-0.0612	0.0105	-5.8009	0.0000	***
	4	0.0215	0.0067	3.2172	0.0013	***	0.0163	0.0031	5.2566	0.0000	***
	5	0.0350	0.0073	4.8124	0.0000	***	0.0450	0.0077	5.8232	0.0000	***
	6	0.0062	0.0088	0.7074	0.4793		0.0141	0.0026	5.4111	0.0000	***
chlorides	1	0.0033	0.0047	0.7097	0.4779	0.0023	0.0055	0.4147	0.6784		
	2	0.1416	0.0600	2.3600	0.0183	**	0.0166	0.0398	0.4167	0.6769	
	3	1.1575	0.4107	2.8184	0.0048	***	0.0811	0.1945	0.4169	0.6768	
	4	0.1198	0.5269	0.2273	0.8202		-0.0216	0.0518	-0.4175	0.6763	
	5	-1.3265	0.4174	-3.1778	0.0015	***	-0.0596	0.1431	-0.4166	0.6770	
	6	-0.0957	0.4908	-0.1949	0.8454		-0.0187	0.0449	-0.4166	0.6769	
total.sulfur.dioxide	1	0.0000	0.0000	2.0716	0.0383	**	0.0000	0.0000	0.8906	0.3732	
	2	-0.0000	0.0000	-0.9400	0.3472		0.0000	0.0000	0.9069	0.3645	
	3	0.0001	0.0001	1.2302	0.2186		0.0001	0.0001	0.9101	0.3628	
	4	-0.0001	0.0001	-1.0828	0.2789		-0.0000	0.0000	-0.9086	0.3636	
	5	0.0000	0.0001	0.1169	0.9069		-0.0001	0.0001	-0.9095	0.3631	
	6	0.0000	0.0001	0.2671	0.7894		-0.0000	0.0000	-0.9078	0.3640	

Significance levels correspond to: ***. < 0.01, **. < 0.05, *. < 0.1.



Empirical Application

Mean Marginal Effects

Dataset		Ordered Forest				Ordered Logit					
Variable	Class	Effect	Std.Error	t-Value	p-Value	Effect	Std.Error	t-Value	p-Value		
alcohol	1	-0.0010	0.0011	-0.9028	0.3667	-0.0017	0.0005	-3.5943	0.0003	***	
	2	-0.0036	0.0028	-1.2824	0.1997	-0.0125	0.0023	-5.4096	0.0000	***	
	3	-0.0581	0.0063	-9.2777	0.0000	***	-0.0612	0.0105	-5.8009	0.0000	***
	4	0.0215	0.0067	3.2172	0.0013	***	0.0163	0.0031	5.2566	0.0000	***
	5	0.0350	0.0073	4.8124	0.0000	***	0.0450	0.0077	5.8232	0.0000	***
	6	0.0062	0.0088	0.7074	0.4793		0.0141	0.0026	5.4111	0.0000	***
chlorides	1	0.0033	0.0047	0.7097	0.4779	0.0023	0.0055	0.4147	0.6784		
	2	0.1416	0.0600	2.3600	0.0183	**	0.0166	0.0398	0.4167	0.6769	
	3	1.1575	0.4107	2.8184	0.0048	***	0.0811	0.1945	0.4169	0.6768	
	4	0.1198	0.5269	0.2273	0.8202		-0.0216	0.0518	-0.4175	0.6763	
	5	-1.3265	0.4174	-3.1778	0.0015	***	-0.0596	0.1431	-0.4166	0.6770	
	6	-0.0957	0.4908	-0.1949	0.8454		-0.0187	0.0449	-0.4166	0.6769	
total.sulfur.dioxide	1	0.0000	0.0000	2.0716	0.0383	**	0.0000	0.0000	0.8906	0.3732	
	2	-0.0000	0.0000	-0.9400	0.3472		0.0000	0.0000	0.9069	0.3645	
	3	0.0001	0.0001	1.2302	0.2186		0.0001	0.0001	0.9101	0.3628	
	4	-0.0001	0.0001	-1.0828	0.2789		-0.0000	0.0000	-0.9086	0.3636	
	5	0.0000	0.0001	0.1169	0.9069		-0.0001	0.0001	-0.9095	0.3631	
	6	0.0000	0.0001	0.2671	0.7894		-0.0000	0.0000	-0.9078	0.3640	

Significance levels correspond to: ***. < 0.01, **. < 0.05, *. < 0.1.



Empirical Application

Mean Marginal Effects

Dataset		Ordered Forest				Ordered Logit					
Variable	Class	Effect	Std.Error	t-Value	p-Value	Effect	Std.Error	t-Value	p-Value		
alcohol	1	-0.0010	0.0011	-0.9028	0.3667	-0.0017	0.0005	-3.5943	0.0003	***	
	2	-0.0036	0.0028	-1.2824	0.1997	-0.0125	0.0023	-5.4096	0.0000	***	
	3	-0.0581	0.0063	-9.2777	0.0000	***	-0.0612	0.0105	-5.8009	0.0000	***
	4	0.0215	0.0067	3.2172	0.0013	***	0.0163	0.0031	5.2566	0.0000	***
	5	0.0350	0.0073	4.8124	0.0000	***	0.0450	0.0077	5.8232	0.0000	***
	6	0.0062	0.0088	0.7074	0.4793		0.0141	0.0026	5.4111	0.0000	***
chlorides	1	0.0033	0.0047	0.7097	0.4779	0.0023	0.0055	0.4147	0.6784		
	2	0.1416	0.0600	2.3600	0.0183	**	0.0166	0.0398	0.4167	0.6769	
	3	1.1575	0.4107	2.8184	0.0048	***	0.0811	0.1945	0.4169	0.6768	
	4	0.1198	0.5269	0.2273	0.8202		-0.0216	0.0518	-0.4175	0.6763	
	5	-1.3265	0.4174	-3.1778	0.0015	***	-0.0596	0.1431	-0.4166	0.6770	
	6	-0.0957	0.4908	-0.1949	0.8454		-0.0187	0.0449	-0.4166	0.6769	
total.sulfur.dioxide	1	0.0000	0.0000	2.0716	0.0383	**	0.0000	0.0000	0.8906	0.3732	
	2	-0.0000	0.0000	-0.9400	0.3472		0.0000	0.0000	0.9069	0.3645	
	3	0.0001	0.0001	1.2302	0.2186		0.0001	0.0001	0.9101	0.3628	
	4	-0.0001	0.0001	-1.0828	0.2789		-0.0000	0.0000	-0.9086	0.3636	
	5	0.0000	0.0001	0.1169	0.9069		-0.0001	0.0001	-0.9095	0.3631	
	6	0.0000	0.0001	0.2671	0.7894		-0.0000	0.0000	-0.9078	0.3640	

Significance levels correspond to: ***. < 0.01, **. < 0.05, *. < 0.1.



Conclusion

- ▶ new machine learning estimator for ordered choice models based on the random forest algorithm
- ▶ flexible alternative to parametric models including estimation of choice probabilities and marginal effects with inference
- ▶ simulation evidence shows good predictive performance
- ▶ approach already in use for soccer predictions

- ▶ R-package **orf** available from CRAN repository

```
install.packages("orf") # install orf package from CRAN
```



Thank You for Your Attention!

gabriel.okasa@unisg.ch
okasag.github.io



Monte Carlo

Simulation Design

Evaluation Measures:

- ▶ Average Mean Squared Error

$AMSE =$

$$\frac{1}{R} \sum_{j=1}^R \frac{1}{N} \sum_{i=1}^N \frac{1}{M} \sum_{m=1}^M (P[Y_{i,j} = m | X_{i,j} = x] - \hat{P}[Y_{i,j} = m | X_{i,j} = x])^2$$

- ▶ Average Ranked Probability Score

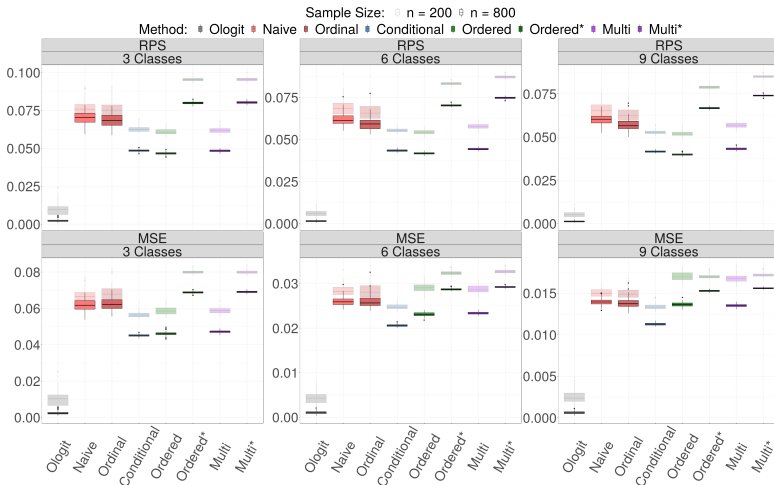
$ARPS =$

$$\frac{1}{R} \sum_{j=1}^R \frac{1}{N} \sum_{i=1}^N \frac{1}{M-1} \sum_{m=1}^M (P[Y_{i,j} \leq m | X_{i,j} = x] - \hat{P}[Y_{i,j} \leq m | X_{i,j} = x])^2$$



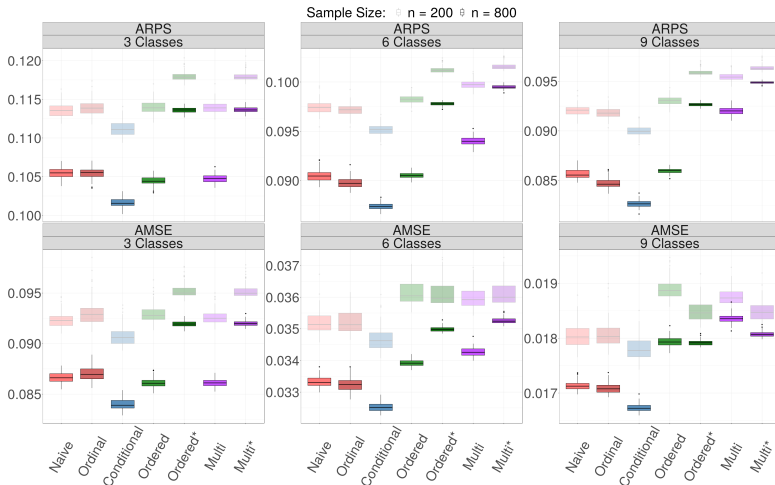
Monte Carlo

Simulation Results: Simple DGP & Low Dimension



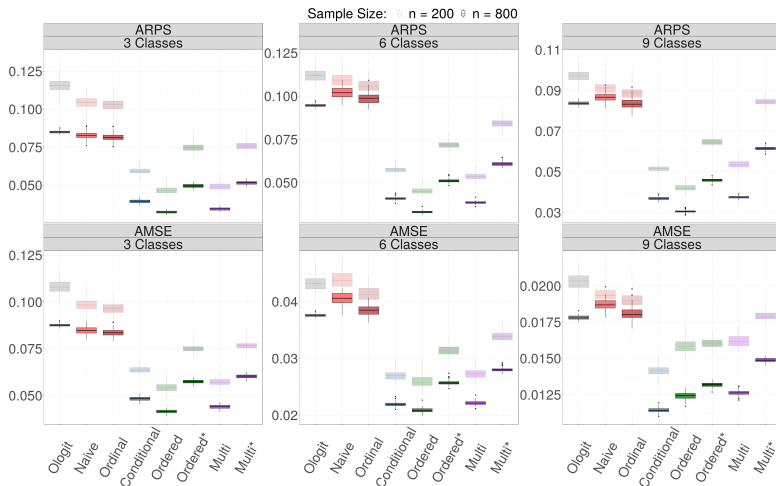
Monte Carlo

Simulation Results: Simple DGP & High Dimension



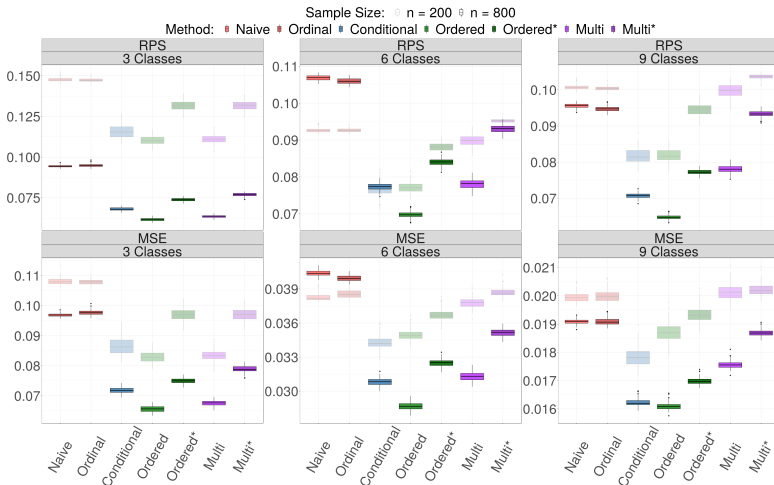
Monte Carlo

Simulation Results: Complex DGP & Low Dimension



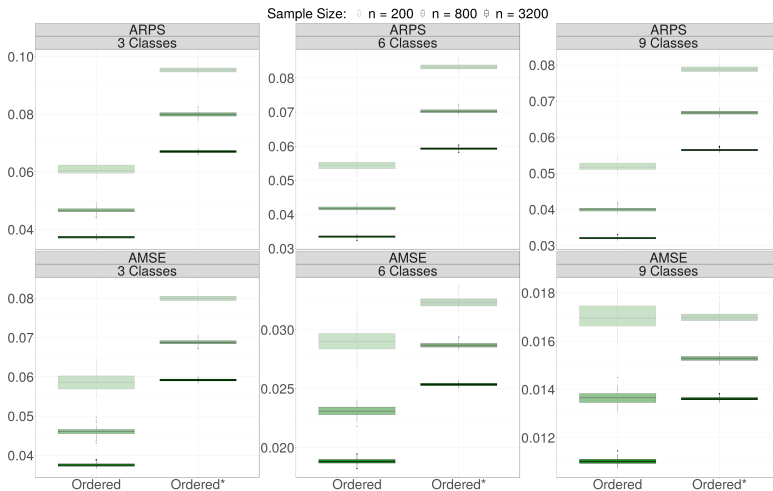
Monte Carlo

Simulation Results: Complex DGP & High Dimension



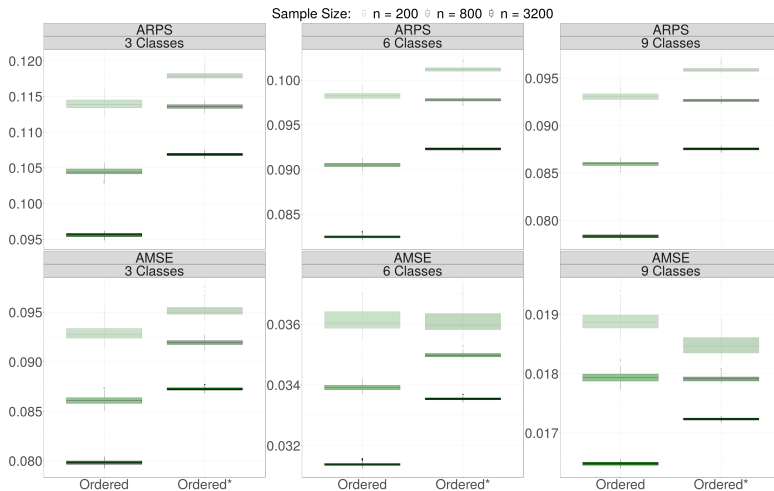
Monte Carlo

Big Simulation Results: Simple DGP & Low Dimension



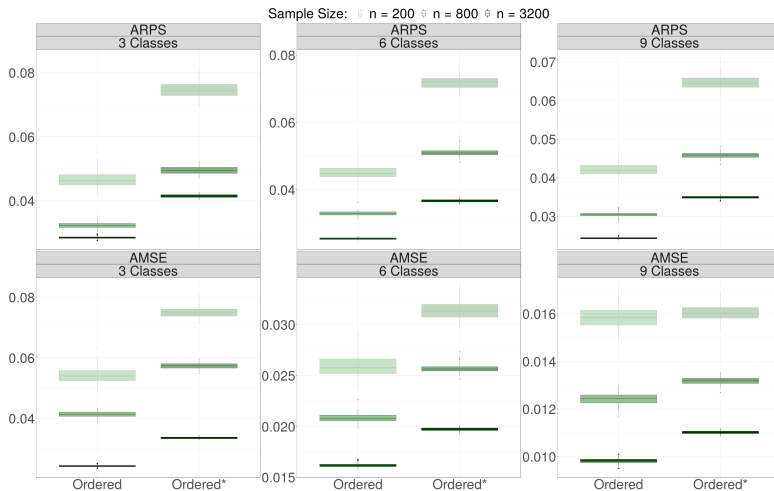
Monte Carlo

Big Simulation Results: Simple DGP & High Dimension



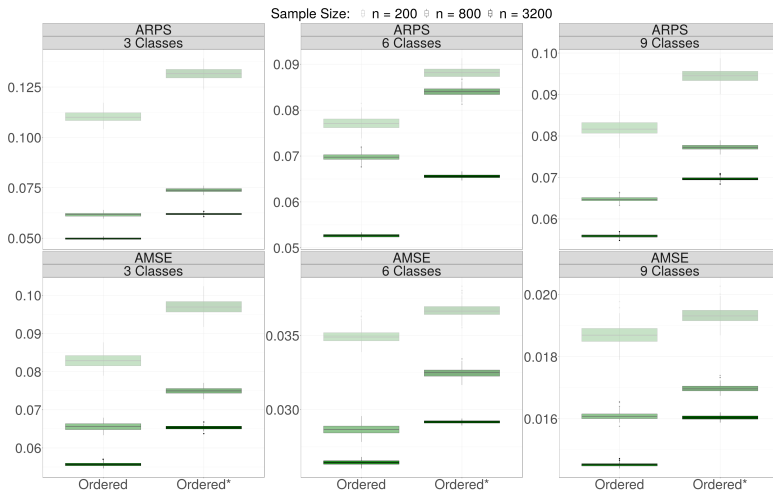
Monte Carlo

Big Simulation Results: Complex DGP & Low Dimension



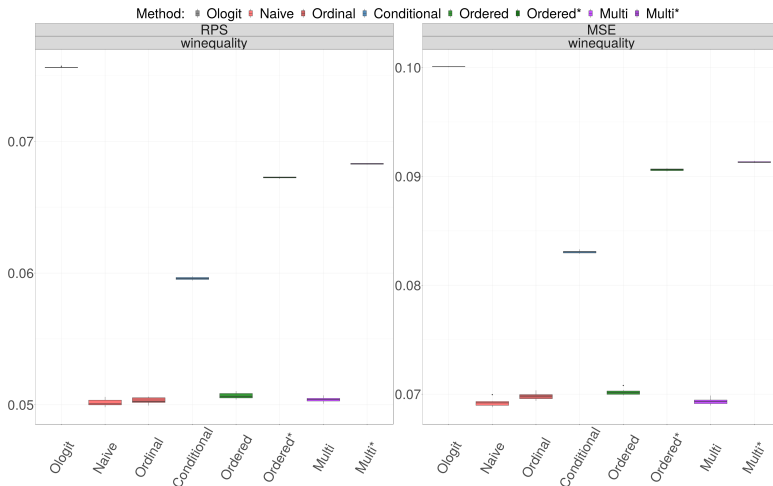
Monte Carlo

Big Simulation Results: Complex DGP & High Dimension







Empirical Application

Choice Probabilities







References I

-  Breiman, L. “Random Forests”. In: *Machine Learning* 45.1 (2001), pp. 5–32.
-  Hornung, Roman. “Ordinal Forests”. In: *Journal of Classification* (2019), pp. 1–14.
-  Hothorn, Torsten, Kurt Hornik, and Achim Zeileis. “Unbiased recursive partitioning: A conditional inference framework”. In: *Journal of Computational and Graphical Statistics* 15.3 (2006), pp. 651–674.
-  Kramer, Stefan, Gerhard Widmer, Bernhard Pfahringer, and Michael De Groot. “Prediction of Ordinal Classes Using Regression Trees”. In: *Fundamenta Informaticae* 47 (2001), pp. 1–13.



References II

-  Lechner, Michael. “Modified Causal Forests for Estimating Heterogeneous Causal Effects”. In: *arXiv preprint arXiv: 1812.09487v2* (2018).
-  McCullagh, Peter. “Regression Models for Ordinal Data”. In: *Journal of the Royal Statistical Society. Series B.* 42.2 (1980), pp. 109–142.
-  Piccarreta, Raffaella. “Classification trees for ordinal variables”. In: *Computational Statistics* 23.3 (2008), pp. 407–427.
-  Pierola, A., I. Epifanio, and S. Alemany. “An ensemble of ordered logistic regression and random forest for child garment size matching”. In: *Computers and Industrial Engineering* 101 (2016), pp. 455–465.



References III



Wager, Stefan and Susan Athey. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. In: *Journal of the American Statistical Association* (2018), pp. 1228–1242.

