Ordered Correlation Forest*

Riccardo Di Francesco†

July 30, 2023

Click here for the most recent version.

Abstract

Empirical studies in various social sciences often involve categorical outcomes with inherent ordering, such as self-evaluations of subjective well-being and self-assessments in health domains. While ordered choice models, such as the ordered logit and ordered probit, are popular tools for analyzing such outcomes, they may impose restrictive parametric and distributional assumptions. This paper proposes a novel estimator, the *ordered correlation forest,* which overcomes these limitations. The proposed estimator modifies a standard random forest splitting criterion to build a collection of forests, each estimating the conditional probability of a single class. Under an "honesty" condition, predictions are consistent and asymptotically normal. The weights induced by each forest are used to estimate the variance of the predictions and obtain estimation and inference about the covariates' marginal effects. Evidence from synthetic data sets shows that the proposed estimator features a superior prediction performance than alternative estimators and demonstrates its ability to construct valid confidence intervals for the covariates' marginal effects.

Keywords: Ordered non-numeric outcomes, choice probabilities, machine learning.
JEL Codes: C14, C25, C55

*I especially would like to thank Franco Peracchi for feedback and suggestions. I am also grateful to Michael Lechner, Jana Mareckova, Annalivia Polselli, and seminar participants at University of Rome Tor Vergata and SEW-HSG research seminars for comments and discussions. Gabriel Okasa generously shared the code for implementing the DGPs in the simulation. The R package for implementing the methodology developed in this paper is available at https://github.com/riccardo-df/ocf. The associated vignette is at https://riccardo-df.github.io/ocf/.

†Department of Economics and Finance, University of Rome Tor Vergata, Rome. Electronic correspondence: riccardo.di.francesco@uniroma2.it.

1 Introduction

Categorical outcomes with a natural order are commonly observed in empirical studies across social sciences. For example, happiness research typically employs large surveys to collect self-evaluations of subjective well-being (Frey & Stutzer, 2002), and health economics is heavily based on self-assessments in several health domains (see e.g., Peracchi & Rossetti, 2012, 2013). These outcomes are usually measured on a discrete scale with five or ten classes, where the classes can be arranged in a natural order without any knowledge about their relative magnitude.

Ordered choice models are a popular class of statistical models used to analyze the relationship between this kind of outcome and a set of covariates (see e.g., Greene & Hensher, 2010). These models target the estimation of the conditional choice probabilities, which represent the probability that the outcome belongs to a certain class given the values of the covariates. Common examples of ordered choice models include ordered probit and ordered logit models. However, these models are limited by their dependence on parametric and distributional assumptions that are often based on analytical convenience rather than knowledge about the underlying data generating process. As a result, econometricians may need to consider alternative techniques to produce more accurate and reliable predictions.

This paper proposes a novel estimator that addresses these limitations, the *ordered correlation forest*. The proposed estimator modifies a standard random forest splitting criterion to build a collection of forests, each estimating the conditional probability of a single class. The estimator inherits the asymptotic properties of random forests proven by Wager and Athey (2018), namely the consistency and asymptotic normality of their predictions. This allows valid inference about conditional probabilities to be made using conventional methods, although requires the individual trees to satisfy a fairly strong condition called "honesty" (Athey & Imbens, 2016). Honesty is a subsample-splitting technique that ensures that different observations are used to place the splits and compute leaf predictions and is crucial to achieving consistency of the random forest predictions.

The particular honesty implementation used by the ordered correlation forest estimator allows for a weight-based estimation of the variance of the predicted probabilities. This is achieved by rewriting the random forest predictions as a weighted average of the outcomes (Athey et al., 2019). The weights, which are obtained for the predicted probabilities, can be properly transformed to obtain estimation and inference about the covariates' marginal effects (for a similar approach, see Lechner & Okasa, 2019; Lechner & Mareckova, 2022).

The rest of the paper unfolds as follows. Section 2 provides a brief overview of ordered choice models and discusses some alternative estimation strategies. Section 3 presents the ordered correlation forest estimator, explaining estimation and inference about the statistical targets of interest. Section 4 uses synthetic data sets to compare the ordered correlation forest with alternative estimators and evaluate its performance in estimating and making inference about the covariates' marginal effects. Section 5 concludes. Further comparisons with alternative estimators using real data sets are provided in the Appendix.

2 Ordered Choice Models

Ordered choice models are a class of statistical models used to analyze the relationship between an ordered non-numeric outcome Y_i and a set of covariates X_i (McCullagh, 1980). These models are typically motivated by postulating the existence of a latent and continuous outcome variable of interest Y_i^*, assumed to obey the following regression model (see e.g., Peracchi, 2014):

$$Y_i^* = g\left(W_i\right) + \epsilon_i \tag{2.1}$$

where W_i consists of a set of raw covariates and ϵ_i is independent of W_i. We may approximate the regression function $g\left(W_i\right)$ by a linear-in-parameter model:

$$g\left(W_i\right) = \sum_{j=1}^{k} \beta_j X_{i,j} + V_{i,k} = X_i^T \beta + V_{i,k} \tag{2.2}$$

where $X_i = \left(X_{i,1}, \ldots, X_{i,k}\right)$ contains k constructed covariates (generally the basic covariates W_i plus interactions and polynomials thereof) and $V_{i,k}$ is an approximation error that is assumed to be independent of X_i.[1] Substituting the linear approximation (2.2) into (2.1) gives:

$$Y_i^* = X_i^T \beta + U_i \tag{2.3}$$

where the random error $U_i = \epsilon_i + V_{i,k}$ depends on k through the approximation error $V_{i,k}$.

However, we generally observe only a discretized version Y_i of Y_i^*, which takes on integer values $m = 1, \ldots, M$ corresponding to different categories or classes. Unknown threshold parameters $-\infty = \zeta_0 < \zeta_1 < \cdots < \zeta_{M-1} < \zeta_M = \infty$ define intervals on the support of Y_i^*, each corresponding to one of the M categories of the observed variable Y_i:

$$\zeta_{m-1} < Y_i^* \leq \zeta_m \implies Y_i = m, \quad m = 1, \ldots, M \tag{2.4}$$

Although the M classes have a natural ordering, they are not measured on a cardinal scale. This limits our ability to make precise quantitative comparisons.

Researchers are typically interested in the estimation of the conditional choice probabilities, defined as:

$$p_m\left(X_i\right) := \mathbb{P}\left(Y_i = m | X_i\right) \tag{2.5}$$

However, the marginal effect of the j-th covariate on $p_m(\cdot)$ is a more interpretable measure for ordered choice models. The marginal effect is defined differently depending on whether the j-th covariate is continuous or discrete:

$$\nabla^j p_m\left(x\right) := \frac{\partial p_m\left(x\right)}{\partial x_j} \tag{2.6}$$

$$\nabla^j p_m\left(x\right) := p_m\left(\lceil x_j \rceil\right) - p_m\left(\lfloor x_j \rfloor\right) \tag{2.7}$$

[1] We allow for $k > n$.

where x_j is the j-th element of the vector x and $\lceil x_j \rceil$ and $\lfloor x_j \rfloor$ correspond to x with its j-th element rounded up and down to the closest integer. We can summarize the marginal effects in various ways, such as computing the marginal effect at the mean $\nabla^j p_m(\bar{x})$, with $\bar{x}$ denoting a vector of means. Alternatively, we can compute the marginal effect at the median, the mean marginal effect, and the median marginal effect.

Assumptions on the distribution of U_i are generally imposed to derive closed-form expressions of the conditional probabilities and the marginal effects. Popular choices are the standard normal and the standard logistic distribution functions, producing the ordered probit and ordered logit models, respectively. Estimation is generally performed using standard maximum likelihood methods.

Although easy to interpret and computationally simple, these models feature several limitations. First, they impose strong distributional and functional form assumptions generally derived from analytical convenience rather than knowledge about the underlying data generating process. Second, the definition and estimation of the marginal effects have the restrictive property of single-crossing, meaning that these effects can change sign only once when moving from the smallest class to the largest. Third, if the number of covariates is larger than the number of observations, estimation breaks down.

Several alternatives have been proposed in the literature to overcome these limitations. For example, Boes and Winkelmann (2006) discuss generalizations of the standard ordered choice models. However, they still rely on parametric and distributional assumptions which limit the increase in flexibility they provide.

Recent developments in statistical learning (see e.g., Hastie et al., 2009; Efron & Hastie, 2016) offer ways to relax these assumptions. However, classification algorithms do not leverage the ordering information embedded in the structure of the outcome, and regression algorithms treat the outcome as if it is measured on a metric scale. Thus, applying these algorithms "off-the-shelf" can lead to biased and inefficient estimation of conditional probabilities.

To overcome these limitations, one approach is to transform ordered non-numeric outcomes into a metric scale using scores based on the classes of the observed outcome, thus allowing us to use any regression algorithm on the transformed outcome. For example, Hothorn et al. (2006) propose using the midpoint values of the intervals defined on the support of the latent outcome as score values. In the cases where Y_i^* is not observed, this translates into setting the scores equal to the class labels of Y_i. However, this assumes that the intervals are of equal length, which may not be accurate in practice. To address this issue, Hornung (2020) proposes the ordinal forest estimator, which optimizes the class intervals and uses score values corresponding to these optimized intervals in a standard regression forest. The optimization process involves growing multiple forests using randomly generated candidate score sets, and constructing the final score values by summarizing the score sets with the smallest out-of-bag error. Hornung (2020) shows that the ordinal forest estimator outperforms a standard regression forest that uses class labels as score values using both real and synthetic data sets. However, the optimization process can be computationally expensive, which may limit its practical use for large data sets or real-time applications.

Another approach involves expressing conditional probabilities as conditional expectations of binary variables, which can be estimated by any regression algorithm. One first strategy, which we label *multinomial machine learning*, is to express conditional probabilities as follows:

$$p_m(X_i) = \mathbb{E}\left[\mathbb{1}(Y_i = m) \mid X_i\right] \tag{2.8}$$

This allows us to estimate each $p_m(\cdot)$ separately by regressing the dummy variable $\mathbb{1}(Y_i = m)$ on X_i using any regression algorithm:

$$\hat{p}_m^{MML}(X_i) = \hat{p}_m(X_i) \tag{2.9}$$

However, $\hat{p}_m^{MML}(\cdot)$ does not leverage the information embedded in the ordered structure of the outcome. To overcome this limitation, an alternative strategy that we label *ordered*

machine learning expresses conditional choice probabilities as the difference between the cumulative probabilities of two adjacent classes:

$$\begin{aligned} p_m(X_i) &= \mathbb{P}(Y_i \leq m|X_i) - \mathbb{P}(Y_i \leq m-1|X_i) \\ &= \mathbb{E}[\mathbb{1}(Y_i \leq m)|X_i] - \mathbb{E}[\mathbb{1}(Y_i \leq m-1)|X_i] \\ &= \mu_m(X_i) - \mu_{m-1}(X_i) \end{aligned} \tag{2.10}$$

Then we can separately estimate $\mu_m(\cdot)$ for all $m = 1, \ldots, M-1$ using any regression algorithm and pick the difference between the cumulative probabilities of two adjacent classes to estimate $p_m(\cdot)$:[2]

$$\hat{p}_m^{OML}(X_i) = \hat{\mu}_m(X_i) - \hat{\mu}_{m-1}(X_i) \tag{2.11}$$

However, $\hat{p}_m^{OML}(\cdot)$ can potentially produce negative predictions, thereby contradicting the definition of probabilities.[3]

3 Estimation and Inference

In this section, I discuss the implementation of the ordered correlation forest (OCF) estimator. First, I illustrate the estimation of conditional choice probabilities and marginal effects. Second, I discuss the conditions required for the asymptotic normality and consistency of OCF predictions. Finally, I show how to conduct approximate inference about the statistical targets of interest.

[2]Lechner and Okasa (2019) combine ordered machine learning with random forests (Breiman, 2001) and discuss how to estimate and conduct inference about marginal effects.

[3]Although we might resolve this issue by setting negative predictions to zero, such a solution is suboptimal, and an alternative estimator that does not require truncation may perform better. Therefore, the choice between multinomial and ordered machine learning remains an open question that requires careful empirical investigation.

3.1 Estimation

Similar to the ordered machine learning approach (see equation 2.10), OCF computes the prediction of conditional choice probabilities as the difference between the cumulative probabilities of two adjacent classes. However, instead of estimating $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ separately, OCF internally performs this computation in a single random forest. This allows us to tie the estimation of $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ to correlate the errors made in estimating these two expectations. Additionally, it avoids negative predictions.

To see the importance of correlating the estimation errors, we can decompose the mean squared error of a prediction $\hat{p}_m^{OML}(\cdot)$ at x as follows:[4]

$$\begin{aligned}
\text{MSE}\left(\hat{p}_m^{OML}(x)\right) &= \mathbb{E}\left[\left\{\hat{p}_m^{OML}(x) - p_m(x)\right\}^2\right] \\
&= \mathbb{E}\left[\left\{\hat{\mu}_m(x) - \hat{\mu}_{m-1}(x) - \mu_m(x) + \mu_{m-1}(x)\right\}^2\right] \\
&= \text{MSE}\left(\hat{\mu}_m(x)\right) + \text{MSE}\left(\hat{\mu}_{m-1}(x)\right) - 2\text{EC}\left(\hat{\mu}_m(x), \hat{\mu}_{m-1}(x)\right)
\end{aligned} \tag{3.1}$$

where the last term is the error correlation and captures the degree to which the errors made in estimating $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ are correlated:

$$\text{EC}\left(\hat{\mu}_m(x), \hat{\mu}_{m-1}(x)\right) = \mathbb{E}\left[\left\{\hat{\mu}_m(x) - \mu_m(x)\right\}\left\{\hat{\mu}_{m-1}(x) - \mu_{m-1}(x)\right\}\right] \tag{3.2}$$

Equation (3.1) shows that $\hat{p}_m^{OML}(\cdot)$ is a suboptimal estimator. Besides potentially leading to negative predictions, estimating $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ separately minimizes only the mean squared error terms and ignores the error correlation. Tying the estimation of $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ to correlate the errors could improve estimation performance since errors that move in the same direction cancel out when taking the difference $\hat{\mu}_m(\cdot) - \hat{\mu}_{m-1}(\cdot)$.

To address this limitation, OCF constructs a collection of forests, one for each of the M classes of Y_i. However, rather than the standard criterion, OCF uses equation (3.1) as the splitting rule to build the individual trees in the m-th forest. This allows the estimator to

[4]This decomposition can be applied to any estimation strategy that involves calculating the difference between two surfaces. For example, Lechner and Mareckova (2022) leverage this decomposition to estimate heterogeneous causal effects.

account for the error correlation that $\hat{p}_m^{OML}(\cdot)$ ignores. Intuitively, OCF seeks splits that not only provide good estimates of $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$, but also correlate the errors made in estimating these expectations.

To use (3.1) as the splitting rule, we need to estimate its components. This, in turn, requires an estimator of $\mu_m(\cdot)$ in each node. An unbiased estimator of $\mu_m(\cdot)$ in a child node $C_j \subset \mathcal{X}$ consists of the proportion of observations in C_j whose outcome is lower than or equal to m:

$$\check{\mu}_m(X_i) = \frac{1}{|C_j|} \sum_{i:X_i \in C_j} \mathbb{1}(Y_i \leq m) \tag{3.3}$$

This leads us to estimating $\text{MSE}(\check{\mu}_m(\cdot))$ and $\text{EC}(\check{\mu}_m(\cdot), \check{\mu}_{m-1}(\cdot))$ in each node by their sample analogs:

$$\widehat{\text{MSE}}_j(\check{\mu}_m(X_i)) = \frac{1}{|C_j|} \sum_{i:X_i \in C_j} [\mathbb{1}(Y_i \leq m) - \check{\mu}_m(X_i)]^2 \tag{3.4}$$

$$\widehat{\text{EC}}_j(\check{\mu}_m(X_i), \check{\mu}_{m-1}(X_i)) = \frac{1}{|C_j|} \sum_{i:X_i \in C_j} \mathbb{1}(Y_i \leq m)\,\mathbb{1}(Y_i \leq m-1) - \check{\mu}_m(X_i)\,\check{\mu}_{m-1}(X_i) \tag{3.5}$$

Then, in the m-th forest, OCF constructs individual trees by recursively partitioning each parent node $\mathcal{P} \subseteq \mathcal{X}$ into two child nodes $C_1, C_2 \subset \mathcal{P}$ such that the following minimization problem is solved:

$$\min_{C_1,C_2} \sum_{j=1}^{2} \widehat{\text{MSE}}_j(\check{\mu}_m(X_i)) + \widehat{\text{MSE}}_j(\check{\mu}_{m-1}(X_i)) - 2\widehat{\text{EC}}_j(\check{\mu}_m(X_i), \check{\mu}_{m-1}(X_i)) \tag{3.6}$$

Once the recursive partitioning stops, each tree in the m-th forest unbiasedly estimates $p_m(\cdot)$ at x by computing the proportion of observations in the same leaf as x whose outcome equals m:

$$\begin{aligned} \hat{p}_{m,b}^{OCF}(x) &= \check{\mu}_m(x) - \check{\mu}_{m-1}(x) \\ &= \frac{1}{|L_{m,b}(x)|} \sum_{i \in L_{m,b}(x)} \mathbb{1}(Y_i = m) \end{aligned} \tag{3.7}$$

where $L_{m,b}(x)$ is the set of observations falling in the same leaf of the b-th tree as the prediction point x. The predictions from each tree are then averaged to obtain the forest predictions:[5]

$$\hat{p}_m^{OCF}(x) = \frac{1}{B_m} \sum_{b=1}^{B_m} \hat{p}_{m,b}^{OCF}(x) \tag{3.8}$$

where $b = 1, \ldots, B_m$ indexes the trees in the m-th forest. In contrast to ordered machine learning, OCF ensures model consistency, as the predictions $\hat{p}_m^{OCF}(\cdot)$ always lie in the unit interval by construction.

Estimation of marginal effects proceeds as proposed by Lechner and Okasa (2019). For discrete covariates, we can plug an estimate $\hat{p}_m^{OCF}(\cdot)$ of $p_m(\cdot)$ into equation (2.7) to have a straightforward estimator of $\nabla^j p_m(\cdot)$. For continuous covariates, we use a nonparametric approximation of the infinitesimal change in x_j:

$$\nabla^j \hat{p}_m^{OCF}(x) = \frac{\hat{p}_m^{OCF}(\widehat{\lceil x_j \rceil}) - \hat{p}_m^{OCF}(\widehat{\lfloor x_j \rfloor})}{\bar{x}_j - \underline{x}_j} \tag{3.9}$$

where $\widehat{\lceil x_j \rceil}$ and $\widehat{\lfloor x_j \rfloor}$ correspond to x with its j-th element set to $\bar{x}_j = x_j + \omega \sigma_j$ and $\underline{x}_j = x_j - \omega \sigma_j$, with σ_j the standard deviation of x_j and $\omega > 0$ a tuning parameter.

3.2 Asymptotic Properties

Wager and Athey (2018) establish the consistency and asymptotic normality of random forest predictions. However, besides some regularity and technical assumptions, there are certain conditions regarding the construction of individual trees that must be satisfied. In the following, I define these conditions.

The first condition requires that the trees use different observations to place the splits and compute the leaf predictions. This condition is called *honesty* and is crucial to bounding the bias of forest predictions.

[5] It may be necessary to perform a normalization step to ensure that $\sum_{m=1}^{M} \hat{p}_m^{OCF}(x) = 1$. This is true also for $\hat{p}_m^{MML}(\cdot)$ and $\hat{p}_m^{OML}(\cdot)$.

Definition 1 (*Honesty*). *A tree is honest if it uses the outcome Y_i to either place the splits or compute the leaf predictions, but not both.*

Wager and Athey (2018) implement honesty by first drawing a subsample from the original sample $\mathcal{S}$ for each tree and then splitting the subsample into two halves, using one half to grow the tree and the other half to compute leaf predictions (see also Athey et al., 2019). Alternatively, Lechner and Mareckova (2022) suggest a different implementation (also used by Lechner & Okasa, 2019). First, they split the original sample $\mathcal{S}$ into a training sample $\mathcal{S}^{tr}$ and an honest sample $\mathcal{S}^{hon}$. Then, they use random subsamples from $\mathcal{S}^{tr}$ to construct trees and compute leaf predictions using only $\mathcal{S}^{hon}$. This strategy enables weight-based inference about leaf predictions and their transformations, such as marginal effects. OCF adopts this strategy as well (see Section 3.3).

The second condition is that the leaves of the trees must become small in all dimensions of the covariate space as the sample size increases. This is necessary for achieving consistency of the predictions and is accomplished by introducing randomness in the tree-growing process and enforcing a regularity condition on how quickly the leaves get small.

Definition 2 (*Random-split*). *A tree is random-split if, at every step of the tree-growing procedure, the probability that the next split occurs along the $j-th$ covariate is bounded below by π/k, for some $0 < \pi \leq 1$, for all $j = 1, \ldots, k$.*

Definition 3 (*α-regularity*). *A tree is α-regular if each split leaves at least a fraction α of the observations in the parent node on each side of the split and the trees are fully grown to depth d for some $d \in N$, that is, there are between d and $2d-1$ observations in each terminal node of the tree.*

To achieve α-regularity, OCF ignores splits that do not satisfy this condition. The algorithm always selects the best split from among the candidate splits that would maintain at least a fraction α of the parent node's observations on both sides of the split. This way, we can rule out any influence of the splitting rule on the shape of the final leaves.

Third, trees must be constructed using subsamples drawn without replacement, rather than bootstrap samples, as originally proposed by Breiman (2001).

Lastly, to derive the asymptotic normality, trees must be symmetric.

Definition 4 (*Symmetry*)**.** *A predictor is symmetric if the (possibly randomized) output of the predictor does not depend on the order in which observations are indexed in the training and honest samples.*

Under these conditions, Wager and Athey (2018) establish consistency and asymptotic normality of the random forest predictions. If the M forests constructed by OCF satisfy these conditions, then they inherit these properties, thus producing consistent and asymptotically normally distributed predictions of conditional probabilities.

3.3 Inference

In addition to the consistency and asymptotic normality of the random forest predictions, Wager and Athey (2018) show that the asymptotic variance of such predictions can be consistently estimated by adapting the infinitesimal jackknife estimator proposed by Wager et al. (2014) to the case of subsampling without replacement. This approach can be used to estimate the variance of a prediction $\hat{p}_m^{OCF}(\cdot)$ at x. However, generalizing this method to estimate the variance of marginal effects $\nabla^j \hat{p}_m^{OCF}(\cdot)$ is not straightforward.

To overcome this limitation, OCF employs an alternative approach that leverages the weight-based representation of random forest predictions (Athey et al., 2019) and adapts the weight-based inference proposed by Lechner and Mareckova (2022) (see also Lechner & Okasa, 2019). In particular, OCF implements honesty in a way that guarantees that the weight assigned to the i-th unit is independent of the outcomes of other units. This allows for the derivation of a straightforward formula for the variance of honest predicted probabilities and marginal effects.

First, we express OCF predictions as weighted averages of the outcomes. Let $\mathcal{S}$ denote the observed sample. The following provides an expression for a prediction $\hat{p}_m^{OCF}(\cdot)$ at x

numerically equivalent to that in (3.8):

$$\begin{aligned} \hat{p}_m^{OCF}(x) &= \sum_{i \in \mathcal{S}} \hat{\alpha}_{m,i}(x) \mathbb{1}(Y_i = m) \\ \hat{\alpha}_{m,b,i}(x) &= \frac{\mathbb{1}\left(X_i \in L_{m,b}(x)\right)}{\left|L_{m,b}(x)\right|}, \quad \hat{\alpha}_{m,i}(x) = \frac{1}{B_m} \sum_{b=1}^{B_m} \hat{\alpha}_{m,b,i}(x) \end{aligned} \tag{3.10}$$

where the weights $\hat{\alpha}_{m,1}(x), \ldots, \hat{\alpha}_{m,|\mathcal{S}|}(x)$ determine the forest-based adaptive neighborhood of x. They represent how often the i-th observation in $\mathcal{S}$ shares a leaf with x in the m-th forest. This measures how important the i-th observation is for fitting $p_m(\cdot)$ at x. Notice that $\sum_{i \in \mathcal{S}} \hat{\alpha}_{m,i}(x) = 1$ for all x.

Calculating the variance of a prediction $\hat{p}_m^{OCF}(x)$ in (3.10) is challenging because the weight assigned to the i-th unit $\hat{\alpha}_{m,i}(x)$ is a function of both $\mathcal{S}$ and X_i. Thus, this weight depends on the outcomes of all other units in $\mathcal{S}$, which complicates the formula for the variance.

However, the formula for the variance simplifies under the particular honesty implementation of OCF. Let $\mathcal{S}^{tr}$ and $\mathcal{S}^{hon}$ be a training sample and an honest sample obtained by randomly splitting the observed sample $\mathcal{S}$. Also, let $\hat{\alpha}_{m,i}^{tr}(\cdot)$ be the weights induced by a forest constructed using only $\mathcal{S}^{tr}$. Then, an honest prediction $\tilde{p}_m^{OCF}(\cdot)$ at x is obtained by the following weighted average of observations in $\mathcal{S}^{hon}$:

$$\tilde{p}_m^{OCF}(x) = \sum_{i \in \mathcal{S}^{hon}} \hat{\alpha}_{m,i}^{tr}(x) \mathbb{1}(Y_i = m) \tag{3.11}$$

The new weight assigned to the i-th unit $\hat{\alpha}_{m,i}^{tr}(x)$ is a function of $\mathcal{S}^{tr}$ and of X_i. Thus, under i.i.d. sampling this weight is independent of the outcomes of other units in $\mathcal{S}^{hon}$. This allows us to derive a simple formula for the variance of an honest prediction $\tilde{p}_m^{OCF}(x)$:

$$\mathbb{V}\left(\tilde{p}_m^{OCF}(x)\right) = |\mathcal{S}^{hon}| \, \mathbb{V}\left(\hat{\alpha}_{m,i}^{tr}(x) \mathbb{1}(Y_i = m)\right) \tag{3.12}$$

We can estimate this variance by its sample analog.

By plugging (3.11) into (3.9), we obtain the following estimator of honest marginal effects:

$$\begin{aligned}\nabla^j \tilde{p}_m^{OCF}(x) &= \frac{1}{\bar{x}_j - \underline{x}_j}\left\{\sum_{i\in\mathcal{S}^{hon}} \hat{\alpha}_{m,i}^{tr}(\widehat{\lceil x_j \rceil})\mathbb{1}(Y_i = m) - \sum_{i\in\mathcal{S}^{hon}} \hat{\alpha}_{m,i}^{tr}(\widehat{\lfloor x_j \rfloor})\mathbb{1}(Y_i = m)\right\} \\ &= \frac{1}{\bar{x}_j - \underline{x}_j}\sum_{i\in\mathcal{S}^{hon}} \check{\alpha}_{m,i}^{tr}(\widehat{\lceil x_j \rceil}, \widehat{\lfloor x_j \rfloor})\mathbb{1}(Y_i = m)\end{aligned} \tag{3.13}$$

with $\check{\alpha}_{m,i}^{tr}(\widehat{\lceil x_j \rceil}, \widehat{\lfloor x_j \rfloor}) = \hat{\alpha}_{m,i}^{tr}(\widehat{\lceil x_j \rceil}) - \hat{\alpha}_{m,i}^{tr}(\widehat{\lfloor x_j \rfloor})$ a transformation of the original weights. Using the same argument as before, under i.i.d. sampling the weight assigned to the i-th unit $\check{\alpha}_{m,i}^{tr}(\widehat{\lceil x_j \rceil}, \widehat{\lfloor x_j \rfloor})$ is independent of the outcomes of other units in $\mathcal{S}^{hon}$. Thus the variance of an honest marginal effect $\nabla^j \tilde{p}_m^{OCF}(x)$ can be expressed as follows:

$$\mathbb{V}\left(\nabla^j \tilde{p}_m^{OCF}(x)\right) = \frac{|\mathcal{S}^{hon}|}{\left(\bar{x}_j - \underline{x}_j\right)^2}\mathbb{V}\left(\check{\alpha}_{m,i}^{tr}(\widehat{\lceil x_j \rceil}, \widehat{\lfloor x_j \rfloor})\mathbb{1}(Y_i = m)\right) \tag{3.14}$$

As before, we can estimate this variance by its sample analog.

Following the discussion of Section 3.2, the honest predicted probabilities in (3.11) are consistent and asymptotically normal, provided that the weights $\hat{\alpha}_{m,i}^{tr}(\cdot)$ are induced by a forest composed of α-regular with $\alpha \leq 0.2$ and symmetric random-split trees grown using subsampling without replacement. With these conditions met, we can use the estimated standard errors of honest predicted probabilities $\tilde{p}_m^{OCF}(\cdot)$ to conduct valid inference as usual, e.g., by constructing conventional confidence intervals.

Furthermore, under the same conditions the honest marginal effects in (3.13) are a linear combination of normally distributed predictions, and thus have a normal distribution as well. Therefore, we can also construct conventional confidence intervals for honest marginal effects $\nabla^j \tilde{p}_m^{OCF}(\cdot)$ using their estimated standard errors.

4 Simulation Results

This section uses synthetic data sets to evaluate the performance of the ordered correlation forest (OCF) estimator. In the next subsection, I present a comparison of OCF with various alternative methods in terms of estimating conditional choice probabilities.[6] Then, I

[6] Appendix A provides a comparison with the same alternative methods using real data sets.

assess the ability of OCF in estimating and making inference about the covariates' marginal effects.

4.1 Conditional Probabilities

I consider three designs that differ in the model for the latent outcome variable:

$$\begin{aligned}
&\textit{Design 1.} \qquad Y_i^* = X_i^T \beta + U_i \\
&\textit{Design 2.} \qquad Y_i^* = \sum_{j=1}^{k} X_{i,j} \mathbb{1}\left(X_{i,j} > 0\right) \beta_j + U_i \\
&\textit{Design 3.} \qquad Y_i^* = \sum_{j=1}^{k} sin\left(2X_{i,j}\right) \beta_j + U_i
\end{aligned}$$

with $U_i \sim logistic\,(0,1)$ in all designs. *Design 1* represents a linear model where all the covariates enter without transformation, serving as a benchmark for assessing the performance of the estimators under a straightforward and interpretable setting. In *Design 2* and *Design 3*, the covariates are transformed while preserving the additive structure of the model, thus allowing us to evaluate the estimators' ability to handle non-linearities arising from covariate transformations. For each design, I consider four sample sizes, $|\mathcal{S}| \in \{500,\ 1000, 2000, 4000\}$. Thus, I consider overall twelve different scenarios. The three designs share all the other settings described below.[7]

I obtain the observed outcomes Y_i by discretizing Y_i^* into nine classes:

$$\zeta_{m-1} < Y_i^* \leq \zeta_m \implies Y_i = m, \quad m = 1, \ldots 9$$

I construct the threshold parameters $\zeta_1, \ldots, \zeta_8$ as follows. First, I draw eight values $\zeta_m^q \sim U\,(0.09, 0.91)$ and sort them in ascending order, so that $\zeta_m^q \leq \zeta_{m+1}^q$.[8] Then, I generate a sample of 1,000,000 Y_i^* and set $\zeta_m = Q\left(\zeta_m^q\right)$, with $Q\,(\cdot)$ the empirical quantile function of Y_i^*.

[7] These settings are inspired by the simulation study of Lechner and Okasa (2019).

[8] If the distance between two adjacent values ζ_m^q and ζ_{m+1}^q is not sufficient, I redraw another set of values.

This way, the threshold parameters are unevenly spaced, and the class widths are randomized and unequal.

I generate $k = 30$ covariates $X_i \sim \mathcal{N}(0, \Sigma)$. The components of the coefficient vector β are $\beta_1, \ldots, \beta_5 = 1$, $\beta_6, \ldots, \beta_{10} = 0.75$, and $\beta_{11}, \ldots, \beta_{15} = 0.5$. The remaining covariates have null coefficients, that is, they are "noise" covariates. The variance-covariance matrix Σ is block diagonal and induces correlation among signal covariates as well as among noise covariates, but there is zero correlation between signal and noise covariates:

$$\Sigma = \begin{pmatrix} \mathbf{A}_{signal} & 0 \\ 0 & \mathbf{A}_{noise} \end{pmatrix} \qquad a_{i,j}^{signal} = a_{i,j}^{noise} = \begin{cases} 1, & i = j \\ 0.8, & i \neq j \cap \{i, j \text{ are odd}\} \\ 0, & otherwise \end{cases}$$

After drawing a sample $\mathcal{S}$, I estimate the conditional choice probabilities using both multinomial and ordered machine learning techniques, combining them with random forests (Breiman, 2001) and penalized logistic regressions with an L1 penalty (Tibshirani, 1996). I refer to the resulting estimators as *multinomial random forest* (MRF), *multinomial L1 regression* ($ML1$), *ordered random forest* (ORF), and *ordered L1 regression* ($OL1$).[9] I also consider two versions of OCF, the "adaptive" version OCF_A and the "honest" version OCF_H. This way, we can quantify the loss in the precision derived from using fewer observations to build the forests, representing the price to pay for valid inference. Finally, I include the ordered logit ($LOGIT$) model as a parametric benchmark for comparison.

I feed the estimators with all covariates without adding any polynomials, interaction terms, or other transformations of the covariates. Thus, $LOGIT$ and $OL1$ are correctly specified in *Design 1* and misspecified in the other designs. To implement OCF_H, I randomly split $\mathcal{S}$ into a training sample $\mathcal{S}^{tr}$ used to construct the trees and an honest sample $\mathcal{S}^{hon}$ used to compute the leaf predictions. I choose $|\mathcal{S}^{tr}| = |\mathcal{S}^{hon}| = |\mathcal{S}|/2$.

I rely on an external validation sample $\mathcal{S}^{val}$ of size $|\mathcal{S}^{val}| = 10,000$ to assess the prediction performance of the estimators. This large number of observations helps minimize the

[9] ORF has been extensively discussed in Lechner and Okasa (2019).

sampling variance. For each replication $r = 1, \ldots, R$, I calculate the mean squared error, mean absolute error, and ranked probability score for each estimator:

$$\text{MSE}_r = \frac{1}{|\mathcal{S}^{val}|} \sum_{i \in \mathcal{S}^{val}} \sum_{m=1}^{M} \left[p_m(X_i) - \hat{p}_{m,r}(X_i) \right]^2 \tag{4.1}$$

$$\text{MAE}_r = \frac{1}{|\mathcal{S}^{val}|} \sum_{i \in \mathcal{S}^{val}} \sum_{m=1}^{M} \left| p_m(X_i) - \hat{p}_{m,r}(X_i) \right| \tag{4.2}$$

$$\text{RPS}_r = \frac{1}{|\mathcal{S}^{val}|} \sum_{i \in \mathcal{S}^{val}} \frac{1}{M-1} \sum_{m=1}^{M} \left[\mu_m(X_i) - \hat{\mu}_{m,r}(X_i) \right]^2 \tag{4.3}$$

with $\hat{p}_{m,r}(\cdot)$ the estimated conditional probabilities in the r-th replication, and $\hat{\mu}_{m,r}(x) = \sum_{j=1}^{m} \hat{p}_{j,r}(x)$ the estimated cumulative distribution function. Notice that, by simulation design, we can compute the true probabilities. I summarize these performance measures by averaging over the replications.[10]

Table 4.1 displays the results obtained with $R = 1,000$ replications. The simulation shows that OCF_A outperforms all other estimators uniformly in all the considered scenarios, except for $LOGIT$ and $OL1$ in *Design 1*. Unsurprisingly, $LOGIT$ and $OL1$ perform best in *Design 1*, since they correctly specify the parametric model and the distributional assumption of the error term. However, their performance deteriorates when the model is misspecified, and in *Design 3* even OCF_H demonstrates superior performance despite using half of the observations to train the model.

Ordered machine learning emerges as the superior choice over multinomial machine learning. Specifically, ORF always features lower MSE, MAE, and RPS than MRF, and $OL1$ outperforms $ML1$ in *Design 1* and *Design 2*. In *Design 3*, the MSE of $OL1$ is between 2–6% larger than that of $ML1$, and minimal disparities are observed in the other performance measures.

However, OCF_A always outperforms $OL1$ when the latter is misspecified, with an advantage that ranges between 19–113% in terms of MSE, 10–60% in terms of MAE, and 15–249%

[10] The objective of this simulation exercise is to evaluate the prediction accuracy of each estimator. Thus, I do not consider the variance or the actual coverage rates of confidence intervals as performance measures, as these aspects are not relevant when the interest lies in prediction accuracy.

in terms of RPS, with larger advantages observed in *Design 3*. Moreover, while OCF_A and ORF achieve the same ranked probability scores, the former consistently exhibits lower mean squared error and mean absolute error across all considered scenarios. In *Design 1*, the advantage of OCF_A over ORF is relatively small, with the MSE of ORF being approximately 6% larger than that of OCF_A, and slight differences in MAE observed. The superiority of OCF_A becomes more pronounced in *Design 2* (around 21% in terms of MSE and 11% in terms of MAE) and *Design 3* (around 12% in terms of MSE and 4% in terms of MAE).

Finally, we compare the prediction performance of OCF_A and OCF_H to quantify the cost of honesty resulting from using fewer observations to construct the forests. The results indicate that in terms of MSE, the cost ranges between 30%–46%, while in terms of MAE,

	Design 1				*Design 2*				*Design 3*			
	500	1,000	2,000	4,000	500	1,000	2,000	4,000	500	1,000	2,000	4,000
Panel 1: $\overline{\text{MSE}}$												
$LOGIT$	0.018	0.009	0.004	0.002	0.075	0.069	0.066	0.064	0.171	0.166	0.163	0.162
MRF	0.150	0.132	0.118	0.107	0.076	0.062	0.051	0.043	0.124	0.111	0.099	0.090
$ML1$	0.174	0.163	0.156	0.150	0.086	0.077	0.071	0.067	0.171	0.167	0.163	0.161
ORF	0.133	0.121	0.109	0.099	0.068	0.058	0.049	0.042	0.120	0.107	0.095	0.085
$OL1$	0.088	0.049	0.026	0.013	0.081	0.061	0.048	0.041	0.181	0.174	0.168	0.164
OCF_A	0.127	0.114	0.103	0.094	0.057	0.048	0.040	0.034	0.107	0.095	0.085	0.077
OCF_H	0.167	0.150	0.136	0.126	0.079	0.067	0.058	0.050	0.138	0.125	0.112	0.102
Panel 2: $\overline{\text{MAE}}$												
$LOGIT$	0.198	0.135	0.094	0.066	0.505	0.484	0.474	0.469	0.907	0.899	0.895	0.894
MRF	0.761	0.705	0.661	0.624	0.562	0.505	0.456	0.415	0.750	0.703	0.662	0.626
$ML1$	0.878	0.847	0.824	0.808	0.620	0.583	0.556	0.537	0.950	0.936	0.925	0.918
ORF	0.682	0.645	0.612	0.581	0.526	0.482	0.441	0.406	0.710	0.666	0.627	0.592
$OL1$	0.460	0.333	0.239	0.170	0.565	0.491	0.437	0.403	0.953	0.933	0.917	0.908
OCF_A	0.683	0.641	0.606	0.574	0.479	0.434	0.395	0.363	0.682	0.639	0.602	0.571
OCF_H	0.853	0.797	0.753	0.714	0.599	0.549	0.507	0.470	0.829	0.780	0.732	0.690
Panel 3: $\overline{\text{RPS}}$												
$LOGIT$	0.003	0.001	0.001	0.001	0.026	0.024	0.022	0.022	0.091	0.088	0.086	0.085
MRF	0.031	0.027	0.024	0.021	0.024	0.019	0.016	0.013	0.043	0.037	0.033	0.029
$ML1$	0.044	0.041	0.038	0.037	0.032	0.029	0.026	0.025	0.091	0.089	0.087	0.086
ORF	0.025	0.023	0.020	0.018	0.018	0.015	0.012	0.011	0.034	0.030	0.026	0.024
$OL1$	0.008	0.004	0.002	0.001	0.021	0.018	0.016	0.015	0.089	0.088	0.086	0.086
OCF_A	0.026	0.023	0.020	0.018	0.018	0.015	0.013	0.011	0.036	0.031	0.027	0.025
OCF_H	0.041	0.035	0.031	0.028	0.031	0.026	0.022	0.019	0.062	0.052	0.044	0.037

Table 4.1: Comparison with alternative estimators. The three panels report the average over the replications of MSE_r ($\overline{\text{MSE}}$), MAE_r ($\overline{\text{MAE}}$), and RPS_r ($\overline{\text{RPS}}$).

the cost ranges between 21%–30%. The cost is larger in terms of RPS, ranging between 51%–77%. Despite this cost, the MSE of OCF_H is lower than that of $LOGIT$ when the latter is misspecified.

4.2 Marginal Effects

In this section, I assess the ability of OCF in estimating and making inference about the covariates' marginal effects using the same DGPs discussed in the previous subsection. However, estimating honest marginal effects (3.13) requires more time compared to estimating conditional choice probabilities due to the computation of the weights $\breve{\alpha}^{tr}_{m,i}(\cdot,\cdot)$. This additional computational time can accumulate in a Monte Carlo exercise, making it infeasible to perform. To mitigate this issue, I simplify the DGPs in two main ways.

First, I generate a reduced set of covariates by considering only $k = 4$ variables, thus significantly reducing the number of marginal effects to be estimated. To account for the differences in the estimation of marginal effects for continuous and discrete covariates (see equation 3.9 and relative discussion), I include both types in the analysis. Specifically, I set $X_{i,1}, X_{i,3} \sim \mathcal{N}(0,1)$ and $X_{i,2}, X_{i,4} \sim Bernoulli(0.5)$. The components of the coefficient vector β are $\beta_1 = \beta_2 = 1$ and $\beta_3 = \beta_4 = 0$.

Second, I obtain the observed outcomes Y_i by discretizing Y_i^* into three classes rather than nine. This further reduces the number of marginal effects to be estimated.

After drawing a sample $\mathcal{S}$, I split it into a training sample $\mathcal{S}^{tr}$ and an honest sample $\mathcal{S}^{hon}$ of equal size. Then, I use $\mathcal{S}^{tr}$ to construct the forests, and $\mathcal{S}^{hon}$ to estimate honest marginal effects at the mean and median of the covariates as in equation (3.13).[11] Additionally, I use the sample analog of equation (3.14) to get standard errors for the honest marginal effects. These standard errors are then used to construct conventional 95% confidence intervals.

To assess the performance of the estimator, I calculate the squared bias and variance for each marginal effect, as well as the actual coverage rates of their corresponding confidence

[11] Estimating the mean marginal effect would involve computing the weights $\breve{\alpha}^{tr}_{m,i}(\cdot,\cdot)$ for each prediction point x, which would result in an impractically long computational time for a Monte Carlo exercise.

	Design 1				*Design 2*				*Design 3*			
	500	1,000	2,000	4,000	500	1,000	2,000	4,000	500	1,000	2,000	4,000
Panel 1: Marginal effects at mean												
$Bias^2$	0.006	0.006	0.006	0.007	0.006	0.006	0.007	0.007	0.006	0.007	0.007	0.007
Var	0.016	0.018	0.019	0.02	0.015	0.018	0.02	0.021	0.015	0.019	0.02	0.02
Coverage 95%	0.92	0.94	0.95	0.95	0.92	0.94	0.95	0.95	0.92	0.94	0.95	0.95
Panel 1: Marginal effects at median												
$Bias^2$	0.006	0.006	0.006	0.007	0.007	0.006	0.007	0.006	0.007	0.007	0.007	0.007
Var	0.016	0.018	0.019	0.02	0.016	0.019	0.02	0.021	0.015	0.019	0.02	0.02
Coverage 95%	0.92	0.93	0.95	0.95	0.92	0.94	0.95	0.95	0.92	0.94	0.94	0.95

Table 4.2: Estimation and inference about the covariates' marginal effects. The first panel reports results for the marginal effects at the mean, and the second panel reports results for the marginal effects at the median.

intervals. I summarize these performance measures by averaging across all marginal effects.

Table 4.2 displays the results obtained with $1,000$ replications. The estimated squared bias is close to zero, indicating that the estimator is approximately unbiased. The actual coverage rates of the confidence intervals are in line with the nominal rate. In smaller samples, there is a slight deviation from the nominal rate, but as the sample size increases, the coverage rates converge to the nominal rate.

5 Conclusion

This paper proposes a novel estimator for ordered non-numeric outcomes, the *ordered correlation forest*. The estimator modifies a standard random forest splitting criterion to build a collection of forests, each estimating the conditional probability of a single class.

Under an honesty condition, the proposed estimator inherits the asymptotic properties of random forests proven by Wager and Athey (2018), namely the consistency and asymptotic normality of its predictions. This allows valid inference about conditional probabilities to be made using conventional methods. Moreover, transforming the weights induced by each forest provides a methodology to obtain estimation and inference about the covariates' marginal effects.

Evidence from synthetic data sets shows that the proposed estimator features a superior prediction performance than alternative estimators and demonstrates its ability to construct valid confidence intervals for the covariates' marginal effects.

References

Athey, S., & Imbens, G. W. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360.

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *Annals of Statistics*, *47*(2), 1148–1178.

Boes, S., & Winkelmann, R. (2006). Ordered response models. *Allgemeines Statistisches Archiv*, *90*(1), 167–181.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.

Epstein, E. S. (1969). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, *8*(6), 985–987.

Frey, B. S., & Stutzer, A. (2002). What can economists learn from happiness research? *Journal of Economic Literature*, *40*(2), 402–435.

Greene, W. H., & Hensher, D. A. (2010). *Modeling ordered choices: A primer*. Cambridge University Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Hornung, R. (2020). Ordinal forests. *Journal of Classification*, *37*(1), 4–17.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Janitza, S., Tutz, G., & Boulesteix, A.-L. (2016). Random forest for ordinal responses: Prediction and variable selection. *Computational Statistics & Data Analysis*, *96*, 57–73.

Lechner, M., & Mareckova, J. (2022). Modified causal forest. *arXiv preprint arXiv:2209.03744*.

Lechner, M., & Okasa, G. (2019). Random forest estimation of the ordered choice model. *arXiv preprint arXiv:1907.02436*.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society: Series B (Methodological)*, *42*(2), 109–127.

McCullagh, P., & Nelder, J. A. (2019). *Generalized linear models*. Routledge.

Murphy, A. H. (1970). The ranked probability score and the probability score: A comparison. *Monthly Weather Review*, *98*(12), 917–924.

Peracchi, F. (2014). Econometric methods for ordered responses: Some recent developments. In *Econometric methods and their applications in finance, macro and related fields* (pp. 133–165). World Scientific.

Peracchi, F., & Rossetti, C. (2012). Heterogeneity in health responses and anchoring vignettes. *Empirical Economics*, *42*(2), 513–538.

Peracchi, F., & Rossetti, C. (2013). The heterogeneous thresholds ordered response model: Identification and inference. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(3), 703–722.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242.

Wager, S., Hastie, T., & Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *The Journal of Machine Learning Research*, *15*(1), 1625–1651.

Wright, M. N., & Ziegler, A. (2017). Ranger: A fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, *77*(1), 1–17.

Appendix A Empirical Results

This section uses real data sets to compare the performance of the ordered correlation forest estimator with the same estimators of Section 4.1.

I utilize the same data sets considered by Janitza et al. (2016), Hornung (2020), and Lechner and Okasa (2019). These data sets differ in terms of the number of covariates, observations, and classes of the observed outcome. Table A.1 provides a summary of the data sets. For further details on the background of each data set, the reader is referred to Janitza et al. (2016).

To assess the prediction accuracy of each estimator, I employ a ten-fold cross-validation procedure. Specifically, I randomly divide each data set into ten folds $\mathcal{S}^1, \ldots, \mathcal{S}^{10}$ with roughly equal sizes. For each fold $f = 1, \ldots, 10$, I fit all the estimators using the observations from all the other folds except for $\mathcal{S}^f$. Then, I calculate the same performance measures of Section 4.1 using the held-out $\mathcal{S}^f$:

$$\text{MSE}_f = \frac{1}{|\mathcal{S}^f|} \sum_{i \in \mathcal{S}^f} \sum_{m=1}^{M} \left[\mathbb{1}\left(Y_i = m\right) - \hat{p}_{m,f}\left(X_i\right) \right]^2 \tag{A.1}$$

$$\text{MAE}_f = \frac{1}{|\mathcal{S}^f|} \sum_{i \in \mathcal{S}^f} \sum_{m=1}^{M} \left| \mathbb{1}\left(Y_i = m\right) - \hat{p}_{m,f}\left(X_i\right) \right| \tag{A.2}$$

$$\text{RPS}_f = \frac{1}{|\mathcal{S}^f|} \sum_{i \in \mathcal{S}^f} \frac{1}{M-1} \sum_{m=1}^{M} \left[\mathbb{1}\left(Y_i \leq m\right) - \hat{\mu}_{m,f}\left(X_i\right) \right]^2 \tag{A.3}$$

with $\hat{p}_{m,f}(\cdot)$ the estimated conditional probabilities using all the other folds except for $\mathcal{S}^f$, and $\hat{\mu}_{m,f}(x) = \sum_{j=1}^{m} \hat{p}_{j,f}(x)$ the estimated cumulative distribution function. Finally, I repeat

Data Sets						
Data set	Sample Size	Outcome	Class range			N. Covariates
vlbw	218	Apgar score	1 (Life-threatening)	–	9 (Optimal)	10
mammography	412	Last mammography	1 (Never)	–	3 (Over a year)	5
support	798	Functional disability	1 (None)	–	5 (Fatal)	15
nhanes	1,914	Health status	1 (Excellent)	–	5 (Poor)	26
wines	4,893	Quality	1 (Moderate)	–	6 (High)	11

Table A.1: Summary of data sets, sorted in increasing order of sample size.

this process ten times. This approach eliminates the dependence of the results on a particular training-validation sample split.

Figure A.1 reports the results. It displays boxplots showing the median and interquartile range of the estimated mean squared error (upper row), mean absolute error (mid row), and ranked probability score (lower row), together with their minima and maxima.[12]

Overall, the results indicate that OCF_A performs competitively compared to the other estimators. Contrary to the simulation results, there is no clear superiority between ordered and multinomial machine learning. The performance of the estimators does not exhibit significant differences, except in the case of the *wines* data set, where OCF_A emerges as one

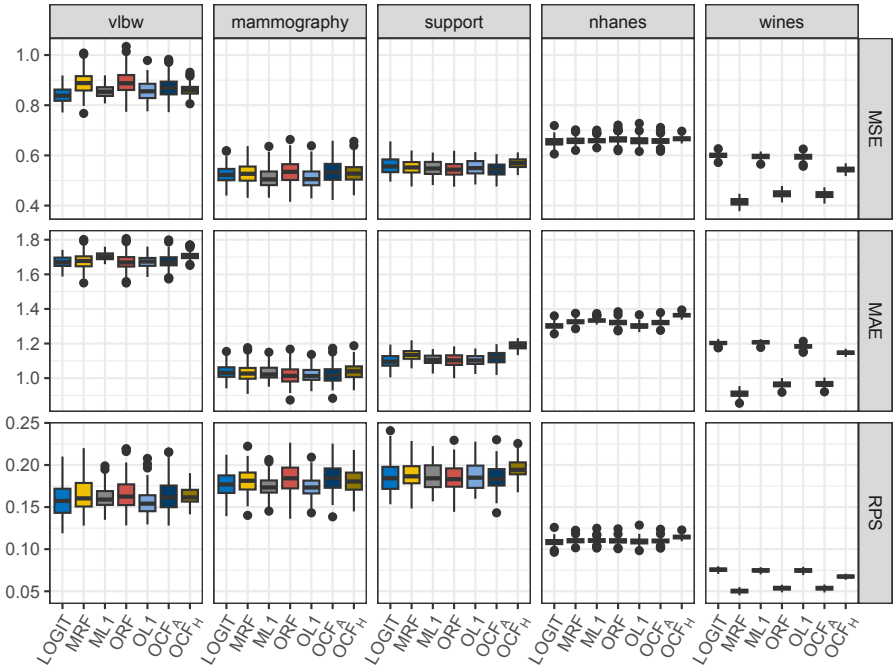

Figure A.1: Prediction performance on real data sets. Each row contains boxplots showing the median and interquartile range of the estimated mean squared error (upper row), mean absolute error (mid row), and ranked probability score (lower row). Each column refers to a different data set, with the data set name displayed at the top of each column. Data sets are sorted according to their sample size.

[12]The cross-validation exercise yields a smaller sample size compared to the simulation results presented in Section 4.1. Consequently, estimates of expected MSE, MAE, and RPS can be more imprecise and influenced by outliers. I report the distribution of the estimated MSE, MAE, and RPS using boxplots to provide a more robust assessment of the prediction performance of each estimator.

of the best estimators together with MRF and ORF. This result highlights the advantage of forest-based methods over those based on penalized regressions.

The cost of honesty resulting from constructing the forests with fewer observations is relatively small and varies with sample size. Across the three largest data sets, the cost ranges between 1%–22% in terms of median MSE, between 3%–18% in terms of median MAE, and between 4%–25% in terms of median RPS. In the two smallest data sets, minimal differences in performance between OCF_A and OCF_H are observed.