Ordered Correlation Forest*

Riccardo Di Francesco†

September 12, 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 these outcomes, they may impose restrictive parametric and distributional assumptions. This paper introduces a novel estimator, the *ordered correlation forest,* that can naturally handle non-linearities in the data and does not assume a specific error term distribution. 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 obtain standard errors for the predicted probabilities and the covariates' marginal effects. Evidence from synthetic data shows that the proposed estimator features a superior prediction performance than alternative forest-based 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 Matteo Iacopini, Michael Lechner, Jana Mareckova, Annalivia Polselli, seminar participants at University of Rome Tor Vergata and SEW-HSG research seminars, and conference participants at the WEEE 2023 for comments and discussions. Gabriel Okasa generously shared the code for implementing part of 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, often referred to as ordered non-numeric outcomes, are commonly observed in empirical studies across the 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, such as ordered logit and ordered probit, are frequently used to analyze the relationship between an ordered 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. However, they 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 introduces a novel machine learning estimator specifically optimized for handling ordered non-numeric outcomes. Employing traditional machine learning estimators "off-the-shelf" can result in biased and inefficient estimation of conditional probabilities. This is because 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.[1] The proposed estimator is designed to mitigate the biases that traditional methods can introduce, ultimately resulting in enhanced predictive performance.

The proposed estimator, named the *ordered correlation forest*, adapts a standard random forest splitting criterion (Breiman, 2001) to the mean squared error relevant to the specific

[1] For comprehensive overviews of traditional classification and regression algorithms, the reader is referred to Hastie et al. (2009) and Efron and Hastie (2016).

estimation problem at hand. The new splitting rule is then used to build a collection of forests, each estimating the conditional probability of a single class. After constructing the individual trees within each forest, the ordered correlation forest employs an unbiased estimator of conditional probabilities within each leaf. Model consistency is ensured, as the predictions always fall within the unit interval by construction. To estimate the covariates' marginal effects, the ordered correlation forest utilizes a nonparametric approximation of derivatives (Lechner & Okasa, 2019).

Under an "honesty" condition (Athey & Imbens, 2016), the ordered correlation forest inherits the asymptotic properties of random forests, namely the consistency and asymptotic normality of their predictions (Wager & Athey, 2018). Honesty is a subsample-splitting technique that requires 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 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 standard errors for the covariates' marginal effects (for a similar approach, see Lechner & Okasa, 2019; Lechner & Mareckova, 2022). We can then use the estimated standard errors to conduct valid inference about the marginal effects as usual, e.g., by constructing conventional confidence intervals.

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, explaining estimation and inference about the statistical targets of interest. Section 4 uses synthetic data 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 provides further comparisons with alternative estimators using real data. Section 6 concludes.

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 W_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):

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

where W_i consists of a set of raw covariates, $g(\cdot)$ is a potentially non-linear regression function, and ϵ_i is independent of W_i and has cumulative distribution $F(\cdot)$. Then, an observational rule links the observed outcome Y_i to the latent outcome Y_i^* using unknown threshold parameters $-\infty = \zeta_0 < \zeta_1 < \cdots < \zeta_{M-1} < \zeta_M = \infty$ that define intervals on the support of Y_i^*, with each interval corresponding to one of the M categories or classes of Y_i:

$$\zeta_{m-1} < Y_i^* \leq \zeta_m \implies Y_i = m, \quad m = 1, \ldots, M \tag{2.2}$$

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:

$$p_m(W_i) := \mathbb{P}(Y_i = m | W_i) \tag{2.3}$$

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:

$$\nabla^j p_m(w) := \begin{cases} \dfrac{\partial p_m(w)}{\partial w_j}, & \text{if } w_j \text{ is continuous} \quad (2.4) \\ p_m\left(\lceil w_j \rceil\right) - p_m\left(\lfloor w_j \rfloor\right), & \text{if } w_j \text{ is discrete} \quad (2.5) \end{cases}$$

where w_j is the j-th element of the vector w and $\lceil w_j \rceil$ and $\lfloor w_j \rfloor$ correspond to w 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 $\nabla^j p_m(\bar{w})$, with $\bar{w}$ denoting a vector of means. Alternatively, we can compute the marginal effect at the median, the mean marginal effect, and the median marginal effect.

From (2.1) and (2.2), the conditional choice probabilities write as:

$$\begin{aligned} p_m(W_i) &= \mathbb{P}\left(\zeta_{m-1} < Y_i^* \leq \zeta_m | W_i\right) \\ &= \mathbb{P}\left(\zeta_{m-1} - g(W_i) < \epsilon_i \leq \zeta_m - g(W_i)\right) \\ &= F\left(\zeta_m - g(W_i)\right) - F\left(\zeta_{m-1} - g(W_i)\right) \end{aligned} \tag{2.6}$$

If the regression function $g(\cdot)$ and the distribution $F(\cdot)$ of the error term ϵ_i are known, we can estimate (2.6) directly using standard maximum likelihood methods.

However, in many practical applications, precise knowledge of $g(\cdot)$ is not available. Instead, a common approach is to approximate it using a linear-in-parameter model (see e.g., Belloni & Chernozhukov, 2011):

$$g(W_i) = X_i^T \beta + V_{i,k} \tag{2.7}$$

where $X_i = h(W_i)$ is a k-dimensional vector of constructed covariates (generally the raw 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. Substituting the linear approximation (2.7) into (2.1) gives:

$$Y_i^* = X_i^T \beta + U_i \tag{2.8}$$

where the random error $U_i = \epsilon_i + V_{i,k}$ depends on k through the approximation error $V_{i,k}$ and has cumulative distribution $G(\cdot)$. Then, we can approximate the statistical target $p_m(\cdot)$ as follows:[2]

[2] The ultimate target of estimation is $p_m(\cdot)$. $p_m^*(\cdot)$ serves as an approximation that allows us to tackle the estimation problem as if it were parametric.

$$
\begin{aligned}
p_m^*(W_i) &:= \mathbb{P}(Y_i = m | h(W_i)) \\
&= G\left(\zeta_m - X_i^T\beta\right) - G\left(\zeta_{m-1} - X_i^T\beta\right)
\end{aligned}
\tag{2.9}
$$

We can impose assumptions on the distribution $G(\cdot)$ of the random error U_i to estimate (2.9) using standard maximum likelihood methods. Popular choices are the standard normal and the standard logistic distribution functions, producing the ordered probit and ordered logit models, respectively. In scenarios where $k > n$, regularization techniques such as L1- or L2-type penalization are needed.

Although easy to interpret and computationally simple, this approach features several limitations. First, it imposes strong distributional assumptions generally derived from analytical convenience rather than knowledge about the underlying data generating process. Second, it requires the specification of a linear-in-parameter model such as (2.7) to account for non-linearities in $g(\cdot)$. Third, the estimated marginal effects have the restrictive property of single-crossing, meaning that they can change sign only once when moving from the smallest class to the largest.

Recent developments in statistical learning (see e.g., Hastie et al., 2009; Efron & Hastie, 2016) offer ways to overcome these limitations. For instance, random forest algorithms (Breiman, 2001) offer a nonparametric estimation approach that does not assume a specific error term distribution and can naturally handle non-linearities in $g(\cdot)$ without requiring a linear-in-parameter model. However, classification forests do not leverage the ordering information embedded in the structure of the outcome, and regression forests treat the outcome as if it is measured on a metric scale. Consequently, applying these algorithms "off-the-shelf" can result in 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. 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:

$$p_m(W_i) = \mathbb{E}[\mathbb{1}(Y_i = m) | W_i] \tag{2.10}$$

This allows us to estimate each $p_m(\cdot)$ separately by regressing the binary variable $\mathbb{1}(Y_i = m)$ on W_i using any nonparametric estimator:

$$\hat{p}_m^{MML}(W_i) = \hat{p}_m(W_i) \tag{2.11}$$

Alternatively, we can specify a linear-in-parameter model to estimate the approximate target $p_m^*(\cdot)$ through parametric regression of the binary variable $\mathbb{1}(Y_i = m)$ on X_i.

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:

$$
\begin{aligned}
p_m(W_i) &= \mathbb{P}(Y_i \leq m | W_i) - \mathbb{P}(Y_i \leq m-1 | W_i) \\
&= \mu_m(W_i) - \mu_{m-1}(W_i)
\end{aligned}
\tag{2.12}
$$

with $\mu_m(W_i) := \mathbb{E}[\mathbb{1}(Y_i \leq m) | W_i]$. Then we can estimate each $\mu_m(\cdot)$ separately by regressing the binary variable $\mathbb{1}(Y_i \leq m)$ on W_i using any nonparametric estimator and pick the difference between the cumulative probabilities of two adjacent classes to estimate $p_m(\cdot)$:[3]

$$
\hat{p}_m^{OML}(W_i) = \hat{\mu}_m(W_i) - \hat{\mu}_{m-1}(W_i) \tag{2.13}
$$

As before, we can alternatively specify a linear-in-parameter model to estimate the approximate target $p_m^*(\cdot)$ through parametric regressions of the binary variables $\mathbb{1}(Y_i \leq m)$ and $\mathbb{1}(Y_i \leq m-1)$ on X_i.

However, $\hat{p}_m^{OML}(\cdot)$ can potentially produce negative predictions, thereby contradicting the definition of probabilities. 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. This paper introduces a novel estimator that leverages the ordered structure of the outcome and produces predictions that always fall within the unit interval, thus resulting in enhanced predictive performance compared to existing methods.

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 consistency and asymptotic normality of OCF predictions. Finally, I show how to conduct approximate inference about the statistical targets of interest.

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

3.1 Estimation

Similar to the ordered machine learning approach, OCF computes the prediction of conditional choice probabilities as the difference between the cumulative probabilities of two adjacent classes (see equation 2.12). 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 w as follows:[4].

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

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:

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

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 (Breiman, 2001), OCF uses equation (3.1) as the splitting rule to build the individual trees in the m-th forest. This

[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 under a selection-on-observables assumption

allows the estimator to account for the error correlation that $\hat{p}_m^{OML}(\cdot)$ ignores. Intuitively, during the tree-building process, OCF anticipates that the predictions in the final leaves will involve the difference between two estimated functions. Consequently, it seeks splits that not only yield accurate estimates of $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ but also take into account the correlation between 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 $C_j \subset \mathcal{W}$ consists of the proportion of observations in C_j whose outcome is not greater than m:

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

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:

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

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

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

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

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

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

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

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

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 fall within 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 (2.5) to have a straightforward estimator of $\nabla^j p_m(\cdot)$:

$$\nabla^j \hat{p}_m^{OCF}(w) = \hat{p}_m^{OCF}(\lceil w_j \rceil) - \hat{p}_m^{OCF}(\lfloor w_j \rfloor) \tag{3.9}$$

For continuous covariates, we use a nonparametric approximation of the infinitesimal change in w_j:

$$\nabla^j \hat{p}_m^{OCF}(w) = \frac{\hat{p}_m^{OCF}(\widehat{\lceil w_j \rceil}) - \hat{p}_m^{OCF}(\widehat{\lfloor w_j \rfloor})}{\overline{w}_j - \underline{w}_j} \tag{3.10}$$

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

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

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 discuss 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.

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 drawing a subsample $\mathcal{S}_b$ from the original sample $\mathcal{S}$ and splitting the subsample into two halves $\mathcal{S}_b^{tr}$ and $\mathcal{S}_b^{hon}$, using $\mathcal{S}_b^{tr}$ to grow the b-th tree and $\mathcal{S}_b^{hon}$ to compute its leaf predictions (see also Athey et al., 2019). Alternatively, Lechner and Mareckova (2022) propose a different approach. They divide the original sample $\mathcal{S}$ into a training sample $\mathcal{S}^{tr}$ and an honest sample $\mathcal{S}^{hon}$, constructing trees from random subsamples of $\mathcal{S}^{tr}$ and computing their leaf predictions from $\mathcal{S}^{hon}$. This strategy ensures that, under i.i.d. sampling, the weights assigned to individual units in $\mathcal{S}^{hon}$ are independent of the outcomes of other units, thus allowing for weight-based inference about leaf predictions and their transformations, such as marginal effects. OCF adopts this strategy as well (details in Section 3.3). However, this strategy is somewhat less efficient than the approach proposed by Wager and Athey (2018). This is because, under the latter approach, each data point w will participate in both $\mathcal{S}_b^{tr}$ and $\mathcal{S}_b^{hon}$ of some trees, thus achieving honesty while making more efficient use of the data. However, under this approach, each weight can depend on other units' outcomes, which limits the usage of the weight-based representation of random forest predictions for obtaining standard errors for the leaf predictions and their transformations.

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).

Fourth, to establish 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 w. 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 w numerically equivalent to that in (3.8):

$$\begin{aligned} \hat{p}_m^{OCF}(w) &= \sum_{i \in \mathcal{S}} \hat{\alpha}_{m,i}(w) \, \mathbb{1}(Y_i = m) \\ \hat{\alpha}_{m,b,i}(w) = \frac{\mathbb{1}\left(W_i \in L_{m,b}(w)\right)}{\left|L_{m,b}(w)\right|}, &\quad \hat{\alpha}_{m,i}(w) = \frac{1}{B_m} \sum_{b=1}^{B_m} \hat{\alpha}_{m,b,i}(w) \end{aligned} \tag{3.11}$$

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

Calculating the variance of a prediction $\hat{p}_m^{OCF}(w)$ in (3.11) is challenging because the weight assigned to the i-th unit $\hat{\alpha}_{m,i}(w)$ is a function of both $\mathcal{S}$ and W_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 $\hat{\alpha}^{tr}_{m,i}(\cdot)$ be the weights induced by a forest constructed using only $\mathcal{S}^{tr}$. Then, an honest prediction $\tilde{p}^{OCF}_{m}(\cdot)$ at w is obtained by the following weighted average of observations in $\mathcal{S}^{hon}$:

$$\tilde{p}^{OCF}_{m}(w) = \sum_{i \in \mathcal{S}^{hon}} \hat{\alpha}^{tr}_{m,i}(w) \mathbb{1}(Y_i = m) \tag{3.12}$$

The new weight assigned to the i-th unit $\hat{\alpha}^{tr}_{m,i}(w)$ is a function of $\mathcal{S}^{tr}$ and of W_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 $\tilde{p}^{OCF}_{m}(w)$:

$$\mathbb{V}\left(\tilde{p}^{OCF}_{m}(w)\right) = |\mathcal{S}^{hon}| \, \mathbb{V}\left(\hat{\alpha}^{tr}_{m,i}(w) \mathbb{1}(Y_i = m)\right) \tag{3.13}$$

We can estimate this variance by its sample analog.

By plugging (3.12) into (3.10), we obtain the following estimator of honest marginal effects:[6]

$$\begin{aligned} \nabla^j \tilde{p}^{OCF}_{m}(w) &= \frac{1}{\overline{w}_j - \underline{w}_j} \left\{ \sum_{i \in \mathcal{S}^{hon}} \hat{\alpha}^{tr}_{m,i}(\widehat{\lceil w_j \rceil}) \mathbb{1}(Y_i = m) - \sum_{i \in \mathcal{S}^{hon}} \hat{\alpha}^{tr}_{m,i}(\widehat{\lfloor w_j \rfloor}) \mathbb{1}(Y_i = m) \right\} \\ &= \frac{1}{\overline{w}_j - \underline{w}_j} \sum_{i \in \mathcal{S}^{hon}} \check{\alpha}^{tr}_{m,i}(\widehat{\lceil w_j \rceil}, \widehat{\lfloor w_j \rfloor}) \mathbb{1}(Y_i = m) \end{aligned} \tag{3.14}$$

with $\check{\alpha}^{tr}_{m,i}(\widehat{\lceil w_j \rceil}, \widehat{\lfloor w_j \rfloor}) = \hat{\alpha}^{tr}_{m,i}(\widehat{\lceil w_j \rceil}) - \hat{\alpha}^{tr}_{m,i}(\widehat{\lfloor w_j \rfloor})$ 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 $\check{\alpha}^{tr}_{m,i}(\widehat{\lceil w_j \rceil}, \widehat{\lfloor w_j \rfloor})$ is independent of the outcomes of other units in $\mathcal{S}^{hon}$. Thus the variance of an honest marginal effect $\nabla^j \tilde{p}^{OCF}_{m}(w)$ can be expressed as follows:

$$\mathbb{V}\left(\nabla^j \tilde{p}^{OCF}_{m}(w)\right) = \frac{|\mathcal{S}^{hon}|}{\left(\overline{w}_j - \underline{w}_j\right)^2} \mathbb{V}\left(\check{\alpha}^{tr}_{m,i}(\widehat{\lceil w_j \rceil}, \widehat{\lfloor w_j \rfloor}) \mathbb{1}(Y_i = m)\right) \tag{3.15}$$

[6] Similar results are obtained for discrete covariates by plugging (3.12) into (3.9).

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

Following the discussion of Section 3.2, the honest predicted probabilities in (3.12) 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 (3.14) 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 to evaluate the performance of the ordered correlation forest (OCF) estimator. In the next subsection, I present the DGPs employed in the simulation. Then, I compare OCF with various alternative methods in terms of estimating conditional choice probabilities. Finally, I assess the ability of OCF in estimating and making inference about the covariates' marginal effects.

4.1 Data-Generating Processes

Latent outcomes are generated as in (2.1), with $\epsilon_i \sim logistic(0,1)$. Six raw covariates are generated as $W_{i,1}, W_{i,3}, W_{i,5} \sim \mathcal{N}(0,1)$ and $W_{i,2}, W_{i,4}, W_{i,6} \sim Bernoulli(0.4)$. Covariates are independent of one another and of ϵ_i. We consider $W_{i,5}$ and $W_{i,6}$ as "noise" covariates, as they enter the DPGs below with null coefficients.

I consider three designs that differ in the regression function $g(\cdot)$:

$$\textit{Design 1.}\ \ g(W_i) = W_i^T \beta$$

$$\textit{Design 2.}\ \ g(W_i) = \sum_{j=1}^{6} \sin\left(2W_{i,j}\right) \beta_j$$

$$\textit{Design 3.}\ \ g(W_i) = 2\sin\left(W_i^T \beta\right)$$

with $\beta = (1, 1, 1/2, 1/2, 0, 0)$ in all designs. *Design 1* represents a linear model where all the raw covariates enter without transformation, serving as a benchmark for assessing the performance of the estimators under a straightforward and interpretable setting. In *Design 2*, 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. *Design 3* introduces more complex non-linearities by departing from the additive model structure and employing a nonlinear regression model. For each design, I consider four sample sizes, $|\mathcal{S}| \in \{500, 1000, 2000, 4000\}$. Thus, I consider overall twelve different scenarios.

In each design, I obtain the observed outcomes Y_i by discretizing Y_i^* into three classes:

$$\zeta_{m-1} < Y_i^* \leq \zeta_m \implies Y_i = m, \ \ m = 1, 2, 3$$

I construct the threshold parameters ζ_1 and ζ_2 as follows. First, I fix two values $\zeta_1^q = 0.33$ and $\zeta_2^q = 0.66$. 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^*. This way, the threshold parameters are uniformly spaced, and the class widths are approximately equal.

4.2 Conditional Probabilities

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* ($OL1$). 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 standard ordered logit ($LOGIT$) model as a parametric benchmark for comparison.

To account for non-linearities in $g(\cdot)$, the parametric methods $LOGIT$, $ML1$, and $OL1$ employ different linear-in-parameter models such as (2.7). Three different specifications are considered. The first specification consists of a model with only the raw covariates, that is, with $X_i = W_i$. The second specification introduces third-order polynomials for continuous covariates, leading to a set of 12 covariates. The third specification enlarges this set by adding all the two-way interactions between the raw covariates, resulting in a total of 27 covariates. In contrast, the forest-based estimators MRF, ORF, OCF_A, and OCF_H are fed with only the raw covariates without adding any polynomials, interaction terms, or other transformations of the covariates, as these estimators can naturally handle non-linearities in $g(\cdot)$. 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 predictive performance of the estimators. This large number of observations helps minimize the 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:

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

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

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

with $\hat{p}_{m,r}(\cdot)$ the estimated conditional probabilities in the r-th replication, and $\hat{\mu}_{m,r}(w) = \sum_{j=1}^{m} \hat{p}_{j,r}(w)$ the estimated cumulative distribution function. Notice that, by simulation design, we can compute the true probabilities as in (2.6). I summarize these performance measures by averaging over the replications.[7]

Table 4.1 displays the results obtained with $R = 2,000$ replications. The simulation shows that OCF outperforms all other forest-based estimators uniformly across all considered scenarios. In particular, OCF consistently achieves lower MSE and MAE than MRF and ORF, and minimal disparities in RPS are observed. An exception arises in *Design 3* where, for the two smallest sample sizes, OCF and MRF show similar performance.

The simulation also shows that OCF maintains a competitive performance when compared to the parametric estimators $LOGIT$, $ML1$, and $OL1$. As expected, when fed with only the raw covariates, $LOGIT$ and $OL1$ perform best in *Design 1* since they correctly specify the parametric model and the distributional assumption of the error term. These estimators are closely followed by $ML1$ in terms of performance. However, the performance of $LOGIT$, $ML1$, and $OL1$ fed with only the raw covariates deteriorates when non-linearities in $g(\cdot)$ are introduced, causing them to rank among the worst estimators. For the smallest sample size, the performance gap with respect to OCF is relatively substantial in *Design 2* (between 26–36% in terms of MSE, between 8–16% in terms of MAE, and between 43–48% in terms of RPS) and moderate in *Design 3* (around 11% in terms of MSE and around 17% in terms of RPS, with minimal disparities in MAE observed). However, in larger samples, this performance gap becomes more pronounced, with the MSE of $LOGIT$, $ML1$, and $OL1$ reaching values up to 165% larger than that of OCF, MAE up to 61%, and RPS up to 232%.

Adding constructed covariates when the model for $g(\cdot)$ is linear and the raw covariates enter without transformation (*Design 1*) deteriorates the performance of $LOGIT$ and $OL1$, as this primarily inflates the variance of the estimation. This effect becomes less relevant in

[7] The objective of this subsection 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.

	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: $\overline{\mathbf{MSE}}$												
$LOGIT_{raw}$	0.005	0.002	0.001	0.001	0.050	0.048	0.046	0.046	0.060	0.058	0.056	0.056
$LOGIT_{poly}$	0.009	0.004	0.002	0.001	0.029	0.025	0.022	0.021	0.051	0.046	0.043	0.042
$LOGIT_{int}$	0.020	0.010	0.005	0.002	0.041	0.030	0.025	0.023	0.035	0.023	0.018	0.016
$ML1_{raw}$	0.014	0.011	0.009	0.008	0.051	0.048	0.047	0.046	0.058	0.054	0.051	0.050
$ML1_{poly}$	0.015	0.010	0.007	0.006	0.033	0.027	0.024	0.022	0.051	0.045	0.042	0.040
$ML1_{int}$	0.016	0.009	0.005	0.003	0.040	0.031	0.026	0.024	0.038	0.026	0.020	0.016
$OL1_{raw}$	0.009	0.005	0.002	0.001	0.054	0.050	0.047	0.046	0.061	0.056	0.053	0.052
$OL1_{poly}$	0.012	0.006	0.003	0.002	0.039	0.030	0.025	0.023	0.054	0.046	0.042	0.040
$OL1_{int}$	0.016	0.008	0.004	0.002	0.048	0.036	0.029	0.025	0.048	0.032	0.023	0.018
MRF	0.045	0.035	0.028	0.022	0.046	0.037	0.029	0.022	0.055	0.043	0.034	0.026
ORF	0.050	0.044	0.040	0.036	0.054	0.048	0.043	0.040	0.061	0.050	0.042	0.037
OCF_{A}	0.044	0.038	0.035	0.032	0.046	0.040	0.037	0.034	0.054	0.044	0.037	0.033
OCF_{H}	0.035	0.025	0.018	0.014	0.040	0.030	0.022	0.017	0.054	0.041	0.030	0.022
Panel 2: $\overline{\mathbf{MAE}}$												
$LOGIT_{raw}$	0.093	0.065	0.045	0.032	0.296	0.289	0.286	0.284	0.310	0.304	0.301	0.300
$LOGIT_{poly}$	0.118	0.082	0.057	0.040	0.206	0.189	0.181	0.176	0.274	0.261	0.255	0.252
$LOGIT_{int}$	0.176	0.120	0.083	0.058	0.244	0.208	0.190	0.181	0.226	0.184	0.163	0.153
$ML1_{raw}$	0.153	0.131	0.118	0.111	0.306	0.296	0.291	0.289	0.326	0.313	0.307	0.303
$ML1_{poly}$	0.160	0.128	0.109	0.097	0.241	0.217	0.204	0.196	0.300	0.280	0.269	0.261
$ML1_{int}$	0.171	0.128	0.094	0.068	0.267	0.235	0.216	0.204	0.262	0.217	0.189	0.171
$OL1_{raw}$	0.127	0.090	0.063	0.045	0.317	0.301	0.292	0.288	0.327	0.310	0.302	0.298
$OL1_{poly}$	0.145	0.102	0.073	0.052	0.261	0.225	0.204	0.191	0.306	0.278	0.263	0.253
$OL1_{int}$	0.169	0.121	0.086	0.062	0.294	0.249	0.220	0.202	0.288	0.230	0.192	0.168
MRF	0.285	0.253	0.224	0.196	0.292	0.260	0.230	0.201	0.316	0.277	0.243	0.212
ORF	0.303	0.283	0.268	0.256	0.318	0.298	0.283	0.271	0.331	0.298	0.274	0.255
OCF_{A}	0.284	0.266	0.252	0.241	0.292	0.275	0.262	0.251	0.312	0.281	0.258	0.241
OCF_{H}	0.257	0.216	0.184	0.160	0.273	0.235	0.203	0.178	0.325	0.278	0.236	0.199
Panel 3: $\overline{\mathbf{RPS}}$												
$LOGIT_{raw}$	0.002	0.001	0.001	0.001	0.023	0.022	0.022	0.021	0.027	0.026	0.025	0.025
$LOGIT_{poly}$	0.004	0.002	0.001	0.001	0.013	0.011	0.010	0.009	0.022	0.020	0.019	0.019
$LOGIT_{int}$	0.008	0.004	0.002	0.001	0.018	0.013	0.011	0.010	0.014	0.010	0.007	0.006
$ML1_{raw}$	0.004	0.003	0.002	0.002	0.024	0.022	0.022	0.021	0.026	0.024	0.023	0.022
$ML1_{poly}$	0.005	0.003	0.002	0.002	0.014	0.012	0.010	0.010	0.022	0.019	0.018	0.017
$ML1_{int}$	0.006	0.003	0.002	0.001	0.018	0.014	0.012	0.010	0.015	0.010	0.007	0.006
$OL1_{raw}$	0.003	0.001	0.001	0.001	0.024	0.023	0.022	0.021	0.027	0.025	0.024	0.024
$OL1_{poly}$	0.004	0.002	0.001	0.001	0.016	0.012	0.011	0.010	0.023	0.020	0.019	0.018
$OL1_{int}$	0.005	0.003	0.001	0.001	0.019	0.015	0.012	0.011	0.017	0.011	0.008	0.006
MRF	0.015	0.012	0.009	0.007	0.016	0.013	0.010	0.008	0.020	0.015	0.012	0.009
ORF	0.016	0.013	0.012	0.011	0.017	0.015	0.013	0.012	0.021	0.016	0.013	0.011
OCF_{A}	0.015	0.013	0.011	0.010	0.016	0.014	0.012	0.011	0.019	0.015	0.013	0.011
OCF_{H}	0.013	0.009	0.007	0.005	0.016	0.012	0.009	0.006	0.023	0.017	0.012	0.008

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 labels in the subscript of the parametric estimators $LOGIT$, $ML1$, and $OL1$ refer to the employed specification: *raw* for only raw covariates, *poly* for raw covariates plus third-order polynomials for continuous covariates, and *int* for raw covariates plus third-order polynomials for continuous covariates plus all two-way interactions between the raw covariates.

larger samples. In contrast, when non-linearities in $g(\cdot)$ are introduced - either via transformations of the covariates (*Design 2*) or by employing a non-linear regression model (*Design 3*) - adding polynomials and interactions of the covariates significantly improves the performance of the parametric estimators. In particular, in *Design 2*, $\boldsymbol{LOGIT}$, $\boldsymbol{ML}1$, and $\boldsymbol{OL}1$ achieve their best performance by introducing third-order polynomials, with their performance deteriorating when interactions between covariates are also included, although this deterioration is substantially attenuated in larger samples. However, in *Design 3*, where more complex non-linearities are introduced, including the interactions is necessary to achieve the best possible performance.

When constructed covariates are included in their specifications, $\boldsymbol{LOGIT}$, $\boldsymbol{ML}1$, and $\boldsymbol{OL}1$ generally exhibit lower MSE, MAE, and RPS than OCF. However, in *Design 2*, OCF outperforms all the parametric methods in larger samples, with advantages over the best parametric specification ranging between 22–31% in terms of MSE, 7–10% in terms of MAE, and 47–55% in terms of RPS. An exception arises in the largest sample where $\boldsymbol{LOGIT}$ and OCF tie in terms of MAE. Moreover, in *Design 3*, the performance gap appears to narrow as the sample size increases, suggesting that OCF might outperform the parametric methods if enough observations are used to train the model.

Finally, we compare the adaptive and the honest versions of OCF to quantify the price to pay for valid inference. Surprisingly, the honest version $\boldsymbol{OCF}_H$ performs better than the adaptive version $\boldsymbol{OCF}_A$ in almost all scenarios despite using half of the observations to construct the forests. Honesty reduces the bias of the forests' estimates but generally comes at the expense of a higher variance. In this simulation, the reduction in bias appears to outweigh the increase in variance, resulting in improved prediction performance.

4.3 Marginal Effects

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 (3.14). Additionally, I use the sample analog of equation (3.15) to get standard errors for the estimated effects.[8]

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 95% confidence intervals. Notice that, by simulation design, we can compute the true marginal effects. I summarize these performance measures by averaging across all marginal effects.

Table 4.2 displays the results obtained with $2,000$ replications. Overall, the simulation shows the ability of OCF to conduct asymptotically valid inference about marginal effects. The estimated squared bias consistently remains close to zero, indicating that the estimator is approximately unbiased. In smaller samples, the actual coverage rates of the confidence intervals tend to fall below the nominal rate and can be as low as 70%. However, as the sample size increases, the coverage rates gradually converge to the nominal level. In *Design 2* and *Design 3*, more observations are required to reach the nominal rate compared to *Design 1*.

When we compare these results with those presented in Table 4.1, an interesting pattern

	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												
$Bias^2$	0.002	0.001	0.001	0.001	0.010	0.009	0.010	0.011	0.011	0.009	0.009	0.010
Var	0.010	0.012	0.014	0.017	0.012	0.015	0.018	0.022	0.011	0.013	0.016	0.019
Coverage 95%	0.84	0.90	0.93	0.95	0.80	0.85	0.90	0.92	0.71	0.79	0.86	0.91
Panel 2: Marginal effects at median												
$Bias^2$	0.002	0.001	0.001	0.001	0.006	0.003	0.001	0.001	0.013	0.009	0.005	0.002
Var	0.010	0.012	0.014	0.017	0.012	0.015	0.018	0.022	0.011	0.013	0.016	0.019
Coverage 95%	0.85	0.90	0.94	0.95	0.81	0.88	0.93	0.95	0.70	0.76	0.83	0.89

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.

[8] Estimating the mean or the median marginal effect and its standard error would involve computing the weights $\check{\alpha}^{tr}_{m,i}(\cdot,\cdot)$ for each prediction point w, which would result in an impractically long computational time for a Monte Carlo exercise. Therefore, I restrict the analysis solely to the marginal effects at the mean and median of the covariates.

emerges: the actual coverage rates of the confidence intervals tend to be worse when the predictive performance of OCF_H is lower. This pattern aligns with the fact that OCF estimates marginal effects by post-processing its conditional probability predictions (see equations 3.9–3.10), meaning that the quality of conditional probability estimation directly impacts the accuracy of marginal effects estimation. As evident in Table 4.1, OCF_H performs best in *Design 1* and exhibits relatively lower performance in *Design 3* compared to *Design 2*. Consequently, for any given sample size, we observe a larger estimated squared bias in *Design 2* and *Design 3* relative to *Design 1*, which explains why more observations are needed to reach the nominal rate in these designs. However, as the sample size increases, the predictive performance of OCF improves, and thus the estimated bias decays asymptotically. Therefore, in larger samples, the estimated confidence intervals are more likely to be centered around the true estimand, resulting in actual coverage rates that converge to the nominal level.

5 Empirical Results

This section uses real data to compare the predictive performance of the ordered correlation forest estimator with the same estimators of Section 4.2.

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 5.1 provides a summary of the data sets. For further details on the background of each data set, the reader is referred to

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 5.1: Summary of data sets, sorted in increasing order of sample size.

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 4.2 using the held-out $\mathcal{S}^f$:

$$\text{MSE}_f = \frac{1}{|\mathcal{S}^f|} \sum_{i \in \mathcal{S}^f} \sum_{m=1}^{M} \left[\mathbb{1}\left(Y_i = m\right) - \hat{p}_{m,f}\left(W_i\right) \right]^2 \tag{5.1}$$

$$\text{MAE}_f = \frac{1}{|\mathcal{S}^f|} \sum_{i \in \mathcal{S}^f} \sum_{m=1}^{M} \left| \mathbb{1}\left(Y_i = m\right) - \hat{p}_{m,f}\left(W_i\right) \right| \tag{5.2}$$

$$\text{RPS}_f = \frac{1}{|\mathcal{S}^f|} \sum_{i \in \mathcal{S}^f} \frac{1}{M-1} \sum_{m=1}^{M} \left[\mathbb{1}\left(Y_i \leq m\right) - \hat{\mu}_{m,f}\left(W_i\right) \right]^2 \tag{5.3}$$

with $\hat{p}_{m,f}(\cdot)$ the estimated conditional probabilities using all the other folds except for $\mathcal{S}^f$, and $\hat{\mu}_{m,f}(w) = \sum_{j=1}^{m} \hat{p}_{j,f}(w)$ the estimated cumulative distribution function. Finally, I repeat this process ten times. This approach eliminates the dependence of the results on a particular training-validation sample split.

Figure 5.1 reports the results by displaying boxplots showing the median and interquartile range of the estimated MSE, MAE, and RPS, together with their minima and maxima.[9]

Overall, the results indicate that OCF performs competitively compared to the other estimators, with no substantial differences in performance observed in most data sets. In the smallest data set under consideration (*vlbw*), ***LOGIT***, ***ML***1, and ***OL***1 perform marginally better than ***MRF***, ***ORF***, and OCF in terms of MSE, with similar MAE and RPS observed. However, in the largest data set (*wines*), OCF emerges as one of the best estimators together with ***MRF*** and ***ORF***. This result highlights the advantage of forest-based methods over parametric methods in larger samples.

[9] The cross-validation exercise yields a smaller sample size compared to the simulation results presented in Section 4.2. 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.

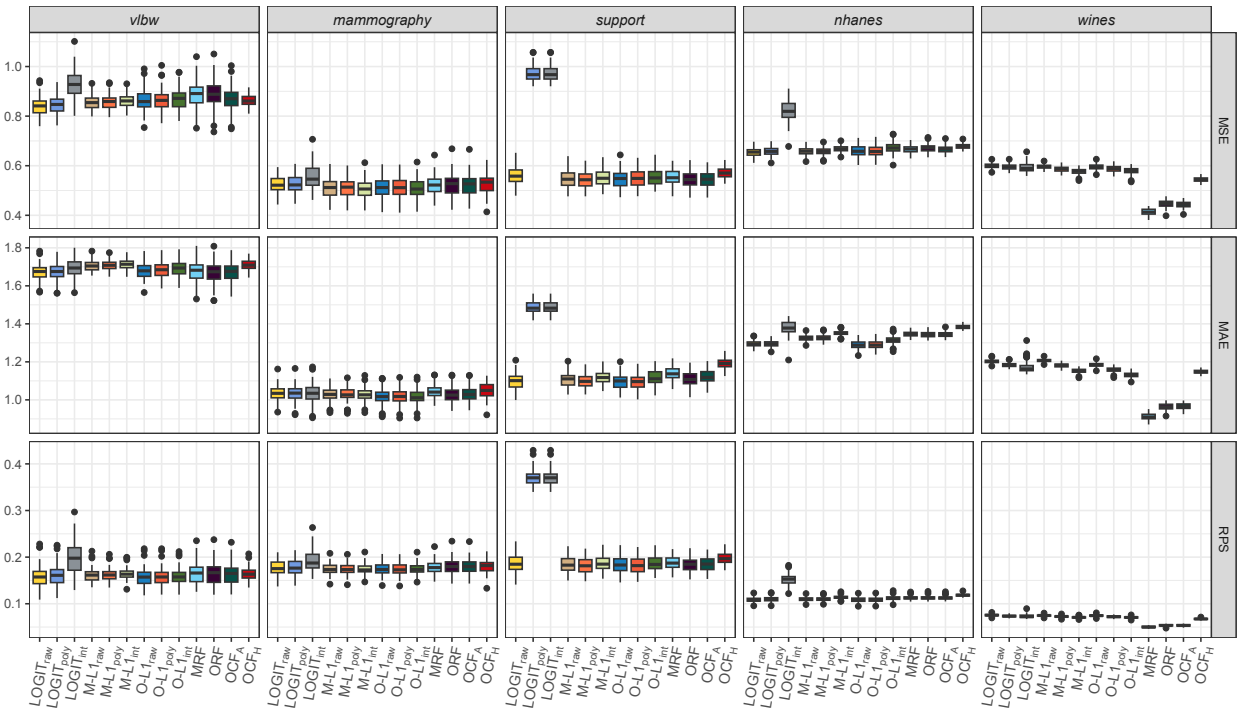

Figure 5.1: Prediction performance on real data sets. Each row contains boxplots showing the median and interquartile range of the estimated mean squared error (upper row), mean absolute error (mid row), and ranked probability score (lower row). Each column refers to a different data set, with the data set name displayed at the top of each column. Data sets are sorted according to their sample size.

The addition of constructed covariates deteriorates the performance of $LOGIT$ while it does not substantially change the performance of $ML1$ and $OL1$, except in the largest data set (*wines*) where it leads to improved performance for all parametric estimators. The deterioration in the performance of $LOGIT$ is particularly pronounced in the *support* and *nhanes* data sets. In these data sets, including third-order polynomials for continuous covariates and all the two-way interactions between the raw covariates results in a total of 324 and 1394 covariates, respectively. Given the moderate sample sizes, the inclusion of the additional covariates substantially increases the variance of the estimation, causing $LOGIT$ to perform worse compared to a specification with only the raw covariates. In contrast, the performance of $ML1$ and $OL1$ is not substantially affected by the inclusion of the additional covariates, as these estimators employ regularization techniques that help contain the increase in variance.

In contrast to the simulation results, the honest version of OCF does not outperform the adaptive version. In the two smaller data sets (*vlbw* and *mammography*), OCF_A and OCF_H exhibit similar MSE and RPS, with the MAE of OCF_H being slightly larger than that of OCF_A. However, as the sample size increases, a performance gap in favor of OCF_A emerges in terms of all performance measures. This gap becomes substantial in the largest data set.

6 Conclusion

This paper proposes a novel machine learning estimator specifically optimized for handling ordered non-numeric outcomes. The proposed estimator adapts a standard random forest splitting criterion (Breiman, 2001) to the mean squared error relevant to the specific estimation problem at hand, thus mitigating the biases that traditional methods can introduce. The new splitting rule is then used to build a collection of forests, each estimating the conditional probability of a single class. A nonparametric approximation of derivatives is employed to estimate the covariates' marginal effects (Lechner & Okasa, 2019).

Under an "honesty" condition (Athey & Imbens, 2016), the estimator inherits the asymptotic properties of random forests, namely the consistency and asymptotic normality of their predictions (Wager & Athey, 2018). The particular honesty implementation used by the ordered correlation forest allows us to obtain standard errors for the covariates' marginal effects by leveraging the weight-based representation of the random forest predictions (Athey et al., 2019). The estimated standard errors can then be used to construct asymptotically valid symmetric confidence intervals.

Evidence from synthetic data shows that the proposed estimator features a superior prediction performance than alternative forest-based 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.

Belloni, A., & Chernozhukov, V. (2011). *High dimensional sparse econometric models: An introduction*. Springer.

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.