Modified Ordered Random Forest*

Riccardo Di Francesco†

March 28, 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 provides a novel estimator, the *modified ordered random 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 and real data sets shows that the proposed estimator features a superior prediction performance than alternative estimators.

Keywords: Ordered choice models, choice probabilities, marginal effects, random forests, 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, and seminar participants at 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/morf. The associated vignette is at https://riccardo-df.github.io/morf/.

†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 that are 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.

Recent developments in statistical learning offer new ways to relax these assumptions. Nonparametric estimators, such as decision trees (Breiman, Friedman, Olshen, & Stone, 1984), random forests (Breiman, 2001), and boosting (Friedman, 2001), have been developed to accommodate continuous or discrete outcomes. For ordered non-numeric (i.e., categorical) outcomes, one possible adaptation involves expressing conditional probabilities as the difference between the cumulative probabilities of two adjacent classes, which transforms the problem into estimating the difference between two conditional expectations. Then, any machine learning algorithm can be used to estimate each expectation individually, and the difference between the estimated surfaces can be used to recover the conditional probabilities (Lechner & Okasa, 2019, combine this strategy with random forests). However, this estima-

tion strategy does not consider the potential correlation between the estimation errors of the cumulative probabilities. Accounting for this correlation may lead to improved estimation performance since errors that move in the same direction cancel out when taking the difference between the estimated surfaces. Therefore, it may be beneficial to tie the estimation of the two conditional expectations together.

This paper provides a novel estimator that overcomes this limitation, the *modified ordered random forest*. The proposed estimator modifies a standard random forest splitting criterion to build a collection of forests, with each forest estimating the conditional probability of a single class. Intuitively, each forest ties the estimation of the cumulative probabilities of two adjacent classes to correlate the estimation errors. This is achieved by using a splitting rule that penalizes splits that would induce a low or negative correlation. Once the forests are built, an unbiased estimator of conditional probabilities is used in each leaf. Model consistency is ensured, as the predictions always lie in the unit interval by construction. Evidence from synthetic and real data sets shows that the modified ordered random forest features a superior prediction performance than alternative estimators.

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, 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 predictions. The particular honesty implementation used by the modified ordered random 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, Tibshirani, & Wager, 2019). The weights, which are obtained for the predicted probabilities, can be properly transformed to obtain estimation and inference about the marginal effects (for a similar approach, see

Lechner & Okasa, 2019; Lechner & Mareckova, 2022).

The rest of the paper unfolds as follows. The next section provides a brief overview of the ordered choice model and discusses some of the estimators proposed in the literature. Section 3 presents the modified ordered random forest estimator, explaining estimation and inference about the statistical targets of interest. Section 4 investigates the finite sample performance of the estimator in a monte-carlo simulation. Section 5 uses real data sets to compare the modified ordered random forest with alternative estimators. Section 6 concludes.

2 Ordered Choice Models

Ordered choice models are a class of statistical models that are 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^*. This variable is assumed to be linearly related to the covariates through unknown coefficients β, and is subject to random error U_i:

$$Y_i^* = X_i^T \beta + U_i, \quad U_i \mid X_i \sim f(0, \sigma^2) \tag{2.1}$$

However, we observe only the 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:

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

Although the M classes have a natural ordering, their relative magnitude is unknown, thus limiting our ability to make precise quantitative comparisons.

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

$$p_m(X_i) := \mathbb{P}(Y_i = m \mid X_i) \tag{2.3}$$

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 based on the continuous or discrete nature of the covariate:

$$p'_{m,j}(x) := \frac{\partial p_m(x)}{\partial x_j} \tag{2.4}$$

$$p'_{m,j}(x) := p_m\left(\lceil x_j \rceil\right) - p_m\left(\lfloor x_j \rfloor\right) \tag{2.5}$$

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 $p'_{m,j}(\bar{x})$, 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 error term distribution f are generally imposed to derive a closed-form expression of the conditional probabilities. Popular choices of f are the standard normal distribution function and the standard logistic distribution function, which produce the ordered probit and ordered logit models. Estimation is generally performed using standard maximum likelihood methods.

Although easy to interpret and computationally efficient, 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. Thus, these models prevent a flexible analysis of ordered categorical outcomes.

Several alternatives have been proposed in the literature to overcome these limitations.

Boes and Winkelmann (2006) discuss generalizations of the standard ordered choice models, but they still rely on parametric and distributional assumptions, limiting the increase in flexibility they provide. Machine learning techniques, such as decision trees (Breiman et al., 1984), allow for a more flexible analysis of ordered categorical outcomes. Piccarreta (2008) discusses several criteria for constructing classification trees that account for the ordered structure of the outcome. However, trees exhibit a large sampling variance and aim to estimate only outcome categories, rather than conditional probabilities.

Other estimation strategies are based on the random forest algorithm introduced by Breiman (2001). Janitza, Tutz, and Boulesteix (2016) compare standard random forests to those based on conditional inference trees (Hothorn, Hornik, & Zeileis, 2006) for ordinal non-numeric outcomes using both simulated and real data. They find no significant differences in prediction accuracy. However, the latter approach requires more computational time, raising concerns about its practical relevance. Another alternative is the ordinal forest estimator proposed by Hornung (2020), which demonstrates better prediction performance. However, this estimator suffers from considerable computational time.

Finally, Lechner and Okasa (2019) introduce the ordered random forest estimator, which, in an extensive simulation study, outperforms conditional forests and ordinal forests in the most complex designs. This estimator constructs separate regression forests to estimate the cumulative probability of each class and uses the difference between the cumulative probabilities of two adjacent classes m and $m-1$ to estimate $p_m(\cdot)$. However, this estimation strategy ignores potential correlation in the estimation errors of the cumulative probabilities. Accounting for this correlation may lead to improved estimation performance since errors that move in the same direction cancel out when taking the difference. The methodology proposed in this paper, the modified ordered random forest, modifies the standard random forest algorithm to account for this correlation.

Lechner and Okasa (2019) also propose a nonparametric estimator of the marginal effects that approximates the infinitesimal change in x_j via its discrete counterpart and leverage the

weights induced by the forests (Athey et al., 2019) to estimate standard errors (see Lechner & Mareckova, 2022, for a similar approach). The modified ordered random forest uses these ideas to obtain estimation and inference about the marginal effects.

3 Modified Ordered Random Forest

In this section, I discuss the implementation of the modified ordered random forest (*MORF*) 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 *MORF* predictions. Finally, I show how to conduct approximate inference about the statistical targets of interest.

3.1 Estimation

Conditional choice probabilities can be expressed as the difference between the cumulative probabilities of two adjacent classes:

$$\begin{aligned} p_m(X_i) &= \mathbb{P}(Y_i \leq m \mid X_i) - \mathbb{P}(Y_i \leq m-1 \mid X_i) \\ &= \mathbb{E}[\mathbb{1}(Y_i \leq m) \mid X_i] - \mathbb{E}[\mathbb{1}(Y_i \leq m-1) \mid X_i] \\ &= \mu_m(X_i) - \mu_{m-1}(X_i) \end{aligned} \tag{3.1}$$

Then, a basic estimator of $p_m(\cdot)$ consists of estimating $\mu_m(\cdot)$ for all $m = 1, \ldots, M-1$ separately using any machine learner and picking the difference (see Lechner & Okasa, 2019, for an estimator that combines this strategy with random forests):[1]

$$\hat{p}_m^{basic}(X_i) = \hat{\mu}_m(X_i) - \hat{\mu}_{m-1}(X_i) \tag{3.2}$$

Negative predictions. However, this estimation strategy fails to take into account that the errors made in estimating $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ may be correlated. The mean squared error

[1] Estimation of the last cumulative probability is not needed as $\mu_M(x) = 1$ for all x by construction.

of a prediction $\hat{p}_m^{basic}(\cdot)$ at x can be decomposed as follows:[2]

$$\begin{aligned} MSE\left(\hat{p}_m^{basic}(x)\right) &= \mathbb{E}\left[\left\{\hat{p}_m^{basic}(x) - p_m(x)\right\}^2\right] \\ &= \mathbb{E}\left[\left\{\hat{\mu}_m(x) - \hat{\mu}_{m-1}(x) - \mu_m(x) + \mu_{m-1}(x)\right\}^2\right] \\ &= \mathbb{E}\left[\left\{\left(\hat{\mu}_m(x) - \mu_m(x)\right) - \left(\hat{\mu}_{m-1}(x) - \mu_{m-1}(x)\right)\right\}^2\right] \\ &= MSE\left(\hat{\mu}_m(x)\right) + MSE\left(\hat{\mu}_{m-1}(x)\right) - 2MCE\left(\hat{\mu}_m(x), \hat{\mu}_{m-1}(x)\right) \end{aligned} \tag{3.3}$$

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

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

The decomposition in (3.3) shows that $\hat{p}_m^{basic}(\cdot)$ is not an optimal estimator. Estimating $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ separately minimizes only the mean squared error terms and ignores the mean correlation error, which would improve the estimation performance if positive. Intuitively, if the estimation errors of the two conditional expectations go in the same direction, then these errors would cancel out when taking the difference $\hat{\mu}_m(\cdot) - \hat{\mu}_{m-1}(\cdot)$. Therefore, it may be advantageous to tie the estimation of $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$.

MORF modifies a standard random forest splitting criterion to overcome this limitation. It uses equation (3.3) as the splitting rule to build a collection of M forests, each estimating the conditional probability of a single class. As in the standard algorithm, individual trees are constructed by greedily minimizing the assumed loss function using axis-aligned splits. Each split is chosen to partition a *"parent"* node $\mathcal{P} \subseteq \mathcal{X}$ into two *"child nodes"* $C_1, C_2 \subset \mathcal{P}$ such as to minimize (3.3) as much as possible. The process is then repeated in the resulting nodes until some stopping criterion is met, e.g., a minimum number of observations in the *"leaves"* (i.e., terminal nodes) of the tree.

To use (3.3) as the splitting rule, we need to estimate its components. This, in turn,

[2]This decomposition applies to each estimation strategy that involves taking the difference between two surfaces. Therefore, the ideas of this paper apply, with the necessary adjustments, to all such estimation problems. Lechner and Mareckova (2022) exploit this decomposition to estimate heterogeneous causal effects.

requires an estimator of $\mu_m(\cdot)$ in each node. Let $Y_i^m \equiv \mathbb{1}(Y_i \leq m)$ be a dummy variable equal to one if the i-th unit's observed outcome is lower or equal to m. Then, an unbiased estimator of $\mu_m(\cdot)$ in a child node C of $\mathcal{P}$ consists of the average of Y_i^m in the node:

$$\hat{\mu}_m(X_i) = \frac{1}{|i : X_i \in C|} \sum_{i:X_i \in C} Y_i^m$$

This leads to estimating $MSE(\hat{\mu}_m(\cdot))$ in each node by:

$$\widehat{MSE}_C(\hat{\mu}_m(X_i)) = \frac{1}{|i : X_i \in C|} \sum_{i:X_i \in C} \left[Y_i^m - \hat{\mu}_m(X_i)\right]^2$$

Similarly, we can construct an estimator of $MCE(\hat{\mu}_m(\cdot), \hat{\mu}_{m-1}(\cdot))$:

$$\widehat{MCE}_C(\hat{\mu}_m(X_i), \hat{\mu}_{m-1}(X_i)) = \frac{1}{|i : X_i \in C|} \sum_{i:X_i \in C} Y_i^m Y_i^{m-1} - \hat{\mu}_m(X_i)\hat{\mu}_{m-1}(X_i)$$

Then, in the m-th forest, each parent node is partitioned into two child nodes $C_1, C_2 \subset \mathcal{P}$ that solve the following minimization problem:

$$\min_{C_1,C_2} \widehat{MSE}_{C_1}\left(\hat{p}_m^{basic}(X_i)\right) + \widehat{MSE}_{C_2}\left(\hat{p}_m^{basic}(X_i)\right) \tag{3.4}$$

with:

$$\widehat{MSE}_{C_j}\left(\hat{p}_m^{basic}(X_i)\right) = \widehat{MSE}_{C_j}(\hat{\mu}_m(X_i)) + \widehat{MSE}_{C_j}(\hat{\mu}_{m-1}(X_i)) - 2\widehat{MCE}_{C_j}(\hat{\mu}_m(X_i), \hat{\mu}_{m-1}(X_i))$$

Using this splitting rule to construct trees in the m-th forest, we favor splits that improve prediction performance as much as possible while penalizing splits that would induce a low or negative correlation between the errors made in estimating $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$. Intuitively, the m-th forest ties the estimation of $\mu_m(\cdot)$ and $\mu_{m-1}(\cdot)$ to make the mean correlation error positive.

Once the m-th forest has been constructed, each tree estimates $p_m(\cdot)$ at x by applying an unbiased estimator in the leaf where x falls. Then, the predictions from each tree are

averaged to get the forest predictions:[3]

$$\hat{p}_{m,b}^{MORF}(x) = \frac{1}{|L_{m,b}(x)|} \sum_{i \in L_{m,b}(x)} \left[Y_i^m - Y_i^{m-1} \right], \quad \hat{p}_m^{MORF}(x) = \frac{1}{B_m} \sum_{b=1}^{B_m} \hat{p}_{m,b}^{MORF}(x) \tag{3.5}$$

where $b = 1, \ldots, B_m$ indexes the trees in the m-th forest and $L_{m,b}(x)$ is the set of training observations falling in the same leaf of the b-th tree as the prediction point x. Equivalently, we can rewrite $\hat{p}_m^{MORF}(\cdot)$ at x as a weighted average of $Y_i^m - Y_i^{m-1}$ based on the weights induced by the forest (Athey et al., 2019). Letting $\mathcal{S}$ be the observed sample:

$$\begin{aligned} \hat{p}_m^{MORF}(x) &= \sum_{i \in \mathcal{S}} \hat{\alpha}_{m,i}(x) \left[Y_i^m - Y_i^{m-1} \right] \\ \hat{\alpha}_{m,b,i}(x) = \frac{\mathbb{1}\left(X_i \in L_{m,b}(x)\right)}{\left|L_{m,b}(x)\right|}, &\quad \hat{\alpha}_{m,i}(x) = \frac{1}{B_m} \sum_{b=1}^{B_m} \hat{\alpha}_{m,b,i}(x) \end{aligned} \tag{3.6}$$

with $\sum_{i \in \mathcal{S}} \alpha_{m,i}(x) = 1$ for all x. These weights define the forest-based adaptive neighborhood of x. They capture the frequency with which the i-th observation in $\mathcal{S}$ falls into the same leaf as x in the m-th forest, thus measuring the relevance of the i-th observation to fitting $p_m(\cdot)$ at x.

Estimation of marginal effects proceeds as proposed by Lechner and Okasa (2019). For discrete covariates, we can plug an estimate $\hat{p}_m^{MORF}(\cdot)$ of $p_m(\cdot)$ into equation (2.5) to have a straightforward estimator of $p'_{m,j}(\cdot)$. For continuous covariates, we use a nonparametric approximation of the infinitesimal change in x_j:

$$\hat{p}'^{MORF}_{m,b}(x) = \frac{\hat{p}_m^{MORF}(\widehat{\lceil x_j \rceil}) - \hat{p}_m^{MORF}(\widehat{\lfloor x_j \rfloor})}{\bar{x}_j - \underline{x}_j} \tag{3.7}$$

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. $\bar{x}_j$ and $\underline{x}_j$

[3] One might need a normalization step to ensure that $\sum_{m=1}^{M} \hat{p}_m^{MORF}(x) = 1$. This would not be required if, rather than estimating $p_m(\cdot)$, $m = 1, \ldots M$ separately, one builds a single forest by constructing trees that minimize $\sum_{m=1}^{M} \widehat{MSE}(\hat{p}_m^{basic}(x))$. However, fitting M separate forests allows the researcher more flexibility in tuning the estimator.

are enforced to respect the support of x_j. Following equation (3.6), we can transform the weights $\hat{\alpha}_{m,i}(\cdot)$ to rewrite (3.7) as a weighted average of $Y_i^m - Y_i^{m-1}$:

$$\begin{aligned}\hat{p}'^{MORF}_{m,b}(x) &= \frac{1}{\bar{x}_j - \underline{x}_j}\left\{\sum_{i\in\mathcal{S}}\hat{\alpha}_{m,i}(\widehat{\lceil x_j \rceil})\left[Y_i^m - Y_i^{m-1}\right] - \sum_{i\in\mathcal{S}}\hat{\alpha}_{m,i}(\widehat{\lfloor x_j \rfloor})\left[Y_i^m - Y_i^{m-1}\right]\right\} \\ &= \frac{1}{\bar{x}_j - \underline{x}_j}\sum_{i\in\mathcal{S}}\tilde{\alpha}_{m,i}(\widehat{\lceil x_j \rceil}, \widehat{\lfloor x_j \rfloor})\left[Y_i^m - Y_i^{m-1}\right]\end{aligned} \tag{3.8}$$

with $\tilde{\alpha}_{m,i}(\widehat{\lceil x_j \rceil}, \widehat{\lfloor x_j \rfloor}) = \hat{\alpha}_{m,i}(\widehat{\lceil x_j \rceil}) - \hat{\alpha}_{m,i}(\widehat{\lfloor x_j \rfloor})$ the transformed weights.

3.2 Asymptotic Properties

Wager and Athey (2018) provide proof of the consistency and asymptotic normality of random forest predictions. In addition to some regularity and technical conditions, there are several requirements on how the individual trees are built. In the following, I define these conditions.

First, we require 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 as follows. First, for each tree, they draw a subsample from the original sample $\mathcal{S}$. Second, they split the subsample into two halves, one used to grow the tree and the other to compute leaf predictions (see also Athey et al., 2019). Following the approach proposed by Lechner and Mareckova (2022), ***MORF*** uses a different strategy. First, we split the original sample $\mathcal{S}$ into a training sample $\mathcal{S}^{tr}$ and an honest sample $\mathcal{S}^{hon}$. Then, trees are constructed using subsamples drawn from $\mathcal{S}^{tr}$, and leaf predictions are computed using observations in $\mathcal{S}^{hon}$ (Lechner & Okasa, 2019, also use this approach). This strategy enables ***MORF*** to conduct valid weight-based inference about the marginal effects (see Section 3.3).

Second, we require the leaves of the trees to become small in all dimensions of the covariate space as the sample size increases. This is a necessary condition to achieve consistency of the predictions and is achieved by enforcing some randomness in the tree-growing procedure and imposing a regularity condition on how fast 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 satisfy α-regularity, *MORF* ignores splits that would violate this condition. The best split is always chosen among the candidate splits that, if chosen, would leave at least a fraction α of the observations in the parent node on each side of the split. This way, we can rule out the influence of a particular 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 derive the asymptotic normality, we require trees to 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, we can establish consistency and asymptotic normality of the predictions. For completeness, I here report the main theorem by Wager and Athey (2018).

Theorem 3.1. *Suppose that we have n independent and identically distributed training examples $(X_i, Y_i) \in [0,1]^k \times \mathcal{R}$. Suppose moreover that the covariates are independently and uniformly distributed $X_i \sim U([0,1]^k)$, that $\theta(x) = \mathbb{E}[Y|X = x]$ and $\theta_2(x) = \mathbb{E}\left[Y^2|X = x\right]$ are*

Lipschitz-continuous, and finally that $Var\left[Y|X=x\right] > 0$ *and* $\mathbb{E}\left[|Y - \mathbb{E}\left[Y|X=x\right]|^{2+\delta}|X=x\right] \leq M$ *for some constants* $\delta, M > 0$*, uniformly over all* $x \in [0,1]^k$*. Given this data-generating process, let* $\mathcal{T}$ *be an honest,* α*-regular with* $\alpha \leq 0.2$*, and symmetric random-split tree in the sense of Definitions 1–4, and let* $\hat{\theta}_n(x)$ *be the estimate for* $\theta(x)$ *given by a random forest with base learner* $\mathcal{T}$ *and a subsample size* s_n*. Finally, suppose that the subsample size* s_n *scales as:*

$$s_n \asymp n^{\beta} \text{ for some } \beta_{min} := 1 - \left(1 + \frac{k}{\pi} \frac{log\left(\alpha^{-1}\right)}{log\left((1-\alpha)^{-1}\right)}\right) < \beta < 1$$

Then, random forest predictions are asymptotically Gaussian:

$$\frac{\hat{\theta}_n(x) - \theta(x)}{\sigma_n(x)} \Longrightarrow \mathcal{N}(0,1) \text{ for a sequence } \sigma_n(x) \to 0$$

3.3 Inference

To conduct valid inference about *MORF* predictions, we need each forest to satisfy the conditions of Theorem 3.1. Let $\mathcal{S}^{tr}$ and $\mathcal{S}^{hon}$ be a partition of the observed sample $\mathcal{S}$. Also, let $\hat{\alpha}^{H}_{m,i}(\cdot)$ be the weights induced by a forest composed of α-regular with $\alpha \leq 0.2$ and symmetric random-split trees in the sense of Definitions 1–4 constructed using only $\mathcal{S}^{tr}$. Then, we can obtain an honest prediction $\hat{p}_m^{MORF_H}(\cdot)$ at x by the following weighted average of observations in $\mathcal{S}^{hon}$:

$$\hat{p}_m^{MORF_H}(x) = \sum_{i \in \mathcal{S}^{hon}} \hat{\alpha}^{H}_{m,i}(x) \left[Y_i^m - Y_i^{m-1}\right] \tag{3.9}$$

Because observations in $\mathcal{S}^{hon}$ have not been used to construct the trees, honesty is satisfied, and thus the prediction is consistent and asymptotically normal. Therefore, given an estimate of the variance of $\hat{p}_m^{MORF_H}(x)$, it is possible to conduct valid inference using standard approaches, e.g., by constructing conventional confidence intervals. One can use the infinitesimal jackknife (Wager, Hastie, & Efron, 2014) to consistently estimate the asymptotic

variance of $\hat{p}_m^{MORF_H}(x)$. However, it is not clear how to generalize this strategy to estimate the variance of the marginal effects $\hat{p}'^{MORF_H}_{m,j}(x)$. Thus, following the proposal of Lechner and Mareckova (2022), ***MORF*** adopts a different strategy exploiting the fact that, under this particular honesty implementation and with i.i.d. sampling, $Y_i^m - Y_i^{m-1}$ and $\hat{\alpha}^H_{m,j}(x)$ are independent for all $i, j \in \mathcal{S}^{hon} : i \neq j$ and for all x. This allows us to have a simple formula for the variance of $\hat{p}_m^{MORF_H}(\cdot)$ at x:

$$\mathbb{V}\left(\hat{p}_m^{MORF_H}(x)\right) = |\mathcal{S}^{hon}|\, \mathbb{V}\left(\hat{\alpha}^H_{m,i}(x)\left[Y_i^m - Y_i^{m-1}\right]\right) \tag{3.10}$$

with $|\mathcal{S}^{hon}|$ the number of observations in the honest sample. Similarly, the variance of the marginal effect at x writes as:

$$\mathbb{V}\left(\hat{p}'^{MORF_H}_{m,j}(x)\right) = \frac{|\mathcal{S}^{hon}|}{\left(\bar{x}_j - \underline{x}_j\right)^2}\, \mathbb{V}\left(\tilde{\alpha}^H_{m,i}(\bar{x}_j, \underline{x}_j)\left[Y_i^m - Y_i^{m-1}\right]\right) \tag{3.11}$$

One can estimate equations (3.10)–(3.11) using sample analogs.

4 Simulation Results

In this section, I compare the modified ordered random forest (***MORF***) with the ordered random forest (***ORF***) estimator (Lechner & Okasa, 2019) using synthetic data. I also consider the ordered logit (***OL***) model as a benchmark for the comparison.

I choose the DGPs to replicate part of the simulation study in Lechner and Okasa (2019). I consider two different designs that differ in the model for the latent outcome variable:

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

with $U_i \sim logistic\,(0, 1)$ in both designs. *Design 1* and *Design 2* share all the other settings

described below. For each design, I consider four sample sizes, $|\mathcal{S}| \in \{500,\ 1000, 2000, 4000\}$. Thus, I consider overall eight different scenarios.

I obtain the observed outcomes Y_i by discretizing Y_i^*:

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

I construct the threshold parameters $\zeta_1, \dots, \zeta_8$ as follows. First, I draw eight values $\zeta_m^q \sim U(0.09, 0.91)$ and sort them in ascending order, so that $\zeta_m^q \leq \zeta_{m+1}^q$. **I ensure that some distance is there.** 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 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 $\beta_1, \dots, \beta_5 = 1$, $\beta_6, \dots, \beta_{10} = 0.75$, and $\beta_{11}, \dots, \beta_{15} = 0.5$. The remaining covariates have null coefficients, that is, they are "noise" covariates. The variance-covariance matrix Σ is block diagonal **we allow some correlation among signal and noise covariates, but signal are not correlated with noise.**:

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

After drawing a sample $\mathcal{S}$, I estimate the conditional choice probabilities using OL, ORF, and two versions of $MORF$, the "adaptive" version $MORF_A$ and the "honest" version $MORF_H$. 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. I feed OL with all covariates without adding any polynomials, interaction terms, or other transformations of the covariates. Thus, OL is correctly specified in *Design 1* and misspecified in *Design 2*. ORF, $MORF_A$ and $MORF_H$ use the same tuning parameters (e.g., the number of trees or the minimum number of observations in each leaf). To implement $MORF_H$, I 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 quality of the approximation. This large number of observations helps minimize the sampling variance. Two performance measures are computed: the average mean squared error and the average ranked probability score, with the averaging carried out over the validation sample:

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

$$ARPS_r = \frac{1}{|\mathcal{S}^{val}|} \sum_{i \in \mathcal{S}^{val}} \frac{1}{M-1} \sum_{m=1}^{M} \left[\mu_m(X_i) - \hat{\mu}_{m,r}(X_i) \right]^2 \tag{4.2}$$

with $\hat{p}_{m,r}(\cdot)$ and $\hat{\mu}_{m,r}(\cdot)$ the estimated functions in the r-th replication. Notice that, by simulation design, we can compute the true probabilities (see Section 2).

Table 4.1 displays the results obtained with 1,000 replications. Broadly speaking, $MORF_A$ performs well relative to ORF in terms of mean squared error, while the two estimators do not have substantial differences in ranked probability score. In *Design 1*, OL is the best-performing estimator for all the sample sizes considered. This is not surprising, as OL correctly specifies the parametric model and the distributional assumption of the error term. However, in *Design 2*, the performance of OL deteriorates and this estimator is rated the

	Design 1				*Design 2*			
	500	1,000	2,000	4,000	500	1,000	2,000	4,000
Panel 1: $\overline{AMSE}$								
OL	0.024	0.012	0.006	0.003	0.171	0.166	0.163	0.162
ORF	0.133	0.121	0.110	0.100	0.120	0.107	0.095	0.085
$MORF_A$	0.127	0.115	0.104	0.095	0.107	0.095	0.085	0.077
$MORF_H$	0.167	0.150	0.137	0.126	0.138	0.125	0.112	0.102
Panel 2: $\overline{ARPS}$								
OL	0.003	0.002	0.001	0.000	0.091	0.088	0.086	0.085
ORF	0.026	0.023	0.020	0.018	0.034	0.030	0.027	0.024
$MORF_A$	0.027	0.024	0.021	0.019	0.037	0.031	0.028	0.025
$MORF_H$	0.041	0.035	0.031	0.028	0.062	0.052	0.044	0.037

Table 4.1: Comparison with OL and ORF. The two panels report the average over the replications of $AMSE_r$ ($\overline{AMSE}$) and $ARPS_r$ ($\overline{ARPS}$).

worst. $MORF_A$ is the second-best estimator in *Design 1* for all sample sizes, with the MSE of ORF about 5% larger than the MSE of $MORF_A$. The cost of honesty, due to the sample splitting, is about 32% in terms of mean squared error and 50% in terms of ranked probability score.

In *Design 2*, the superior performance of the modified ordered random forest estimator becomes more noticeable. $MORF_A$ features the best predictive performance, with the MSE of ORF about 11% larger than the MSE of $MORF_A$. As in *Design 1*, the advantage of $MORF_A$ does not vanish as the sample size diverges. The cost of honesty is now larger, ranging between 28% and 32% in terms of mean squared error and between 51% and 69% in terms of ranked probability score.

	Design 1				*Design 2*				*Design 3*			
	500	1000	2000	4000	500	1000	2000	4000	500	1000	2000	4000
Panel 1: $\overline{AMSE}$												
OL	0.027	0.014	0.003	0.002	0.062	0.083	0.061	0.017	0.282	0.255	0.256	0.249
$RANGER$	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
ORF	0.099	0.094	0.046	0.078	0.049	0.059	0.051	0.025	0.111	0.084	0.1	0.078
$MORF_A$	0.098	0.092	0.046	0.076	0.048	0.056	0.048	0.024	0.109	0.081	0.098	0.076
$MORF_H$	0.147	0.13	0.062	0.108	0.076	0.09	0.075	0.037	0.173	0.121	0.133	0.106
Panel 2: $\overline{ARPS}$												
OL	0.007	0.004	0.001	0	0.029	0.024	0.017	0.007	0.108	0.112	0.088	0.101
$RANGER$	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
ORF	0.025	0.025	0.016	0.019	0.021	0.016	0.013	0.01	0.035	0.031	0.028	0.026
$MORF_A$	0.024	0.024	0.016	0.019	0.021	0.015	0.013	0.01	0.035	0.031	0.028	0.026
$MORF_H$	0.038	0.036	0.022	0.028	0.034	0.025	0.02	0.016	0.061	0.049	0.04	0.037
Panel 3: $\overline{CE}$												
OL	0.159	0.167	0.101	0.155	0.265	0.334	0.268	0.152	0.481	0.413	0.394	0.448
$RANGER$	0.206	0.222	0.139	0.206	0.253	0.317	0.255	0.155	0.32	0.283	0.304	0.293
ORF	0.206	0.226	0.14	0.208	0.254	0.323	0.261	0.154	0.318	0.282	0.304	0.296
$MORF_A$	0.206	0.223	0.139	0.207	0.254	0.317	0.256	0.155	0.316	0.281	0.3	0.292
$MORF_H$	0.227	0.241	0.153	0.228	0.257	0.345	0.275	0.16	0.356	0.304	0.337	0.315

Table 4.2: Comparison with OL and ORF. The two panels report the average over the replications of $AMSE_r$ ($\overline{AMSE}$) and $ARPS_r$ ($\overline{ARPS}$) with $s = PUTSCALEHERE$.

Data Sets						
Data set	Sample Size	Outcome	Class range			N. Covariates
vlbw	218	Apgar score	1 (At risk)	–	9 (Normal)	10
mammography	412	Last mammography	1 (Never)	–	3 (Over a year)	5
nhanes	1,914	Health status	1 (Excellent)	–	5 (Poor)	26
wine_quality	6,497	Quality	1 (Moderate)	–	6 (High)	11

Table 5.1: Summary of data sets.

5 Empirical Results

In this section, I evaluate the performance of the modified ordered random forest ($MORF$) estimator using four real data sets. The data sets are the same used by Janitza et al. (2016), Hornung (2020). and Lechner and Okasa (2019) in their simulations studies.

The wine quality data set contains data on red and white variants of the Portuguese wine "Vinho Verde" and has been constructed by Cortez, Cerdeira, Almeida, Matos, and Reis (2009). The observed outcome is the wine quality, measured on a ten-point scale. The quality of each wine is evaluated by a minimum of three human experts via blind tastes, and the final quality consists of the median of these evaluations. No wine in the data set has been assessed as extremely bad or extremely good, hence the actual number of classes is seven. The data set also contains eleven covariates that provide information on the physicochemical characteristics of each wine that are easily available at the wine certification step. See Table 5.1 for a summary of the data sets and Appendix **??** for a description of all the variables used in the analysis.

I focus on comparing the same estimators investigated in Section 4, i.e., the ordered logit (OL) model, the ordered random forest (ORF) estimator (Lechner & Okasa, 2019), and two versions of $MORF$, the "adaptive" version $MORF_A$ and the "honest" version $MORF_H$. To evaluate the prediction accuracy of each estimator, I consider the same performance measures of Section 4, i.e., the average mean squared error and the average ranked probability score, with the averaging carried out over the validation sample. Because the true probabilities are not observed in real data sets, I substitute the binary variables $\mathbb{1}(Y_i = m)$ and Y_i^m for $p_m(X_i)$

and $\mu_m(X_i)$ in equations 4.1–4.2 to compute the two measures.

I split the sample into ten folds of approximately the same size. Then, for $k = 1, \ldots, 10$, I fit all the estimators using observations in all folds but in the k-th, and compute the performance measures using only the k-th fold. Each fold thus serves the role of validation sample exactly once, and the dependence of the results on a particular training-validation sample split is avoided.

Figure 5.1 reports the results of this cross-validation exercise. It displays boxplots showing the median and interquartile range of the average mean squared error (upper panels) and the average ranked probability score (lower panels), together with their minima and maxima. Overall, $MORF_A$ performs well relative to the other estimators, and the cost of honesty differs across data sets. In the IMDB data set, the prediction accuracy of ORF, $MORF_A$, and $MORF_H$ are similar. OL has the lowest prediction accuracy, as a consequence of the parametric and distributional assumptions that limit its flexibility. Contrary to the findings of Section 4, the cost of honesty is quite small, about 4% in terms of both performance measures.

In the wine data set, $MORF_A$ features the best prediction accuracy in terms of both mean squared error and ranked probability score. The average MSE and RPS of ORF are about 19% and 21% larger than the average MSE and RPS of $MORF_A$. The cost of honesty is substantial, about 47% in terms of mean squared error, leading $MORF_H$ to be rated as the worst estimator.

6 Conclusion

This paper proposes a completely nonparametric estimator of the ordered choice model, the *modified ordered random 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. Evidence from synthetic and real data sets shows that the proposed estimator

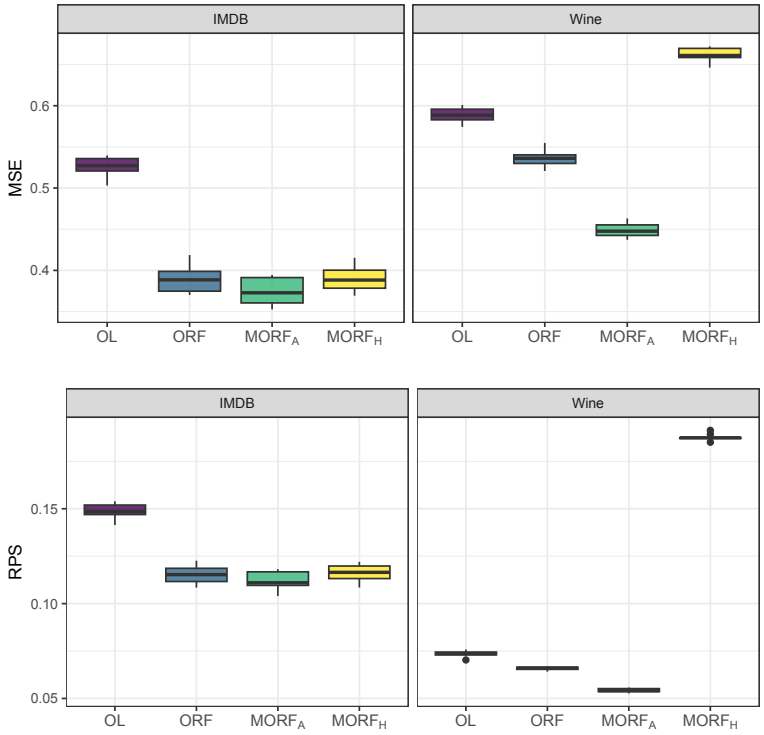

Figure 5.1: Prediction performance on real data sets. Boxplots showing the median and the interquartile range of the average mean squared error (upper panels) and the average ranked probability score (lower panels) are reported. The name of the data set is displayed on top of each panel.

features a superior prediction performance than alternative estimators.

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 marginal effects of each covariate on the estimated probabilities.

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. *The Annals of Statistics*, *47*(2), 1148–1178.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Wadsworth; Brooks.

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

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, *47*(4), 547–553.

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.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of statistics*, 1189–1232.

Frijters, P., Haisken-DeNew, J. P., & Shields, M. A. (2004). Money does matter! evidence from increasing real income and life satisfaction in east germany following reunification. *American Economic Review*, *94*(3), 730–740.

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., & 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.

Piccarreta, R. (2008). Classification trees for ordinal variables. *Computational Statistics*, *23*(3), 407–427.

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 Data

Label	Description
IMDB.	
`rating`	One of bad, good, awesome (outcome).
`year`	Year of release.
`budget`	Total budget in US dollars.
`length`	Length in minutes.
`votes`	Number of IMDB users who rated this movie.
`mpa`	MPA rating.
`genre`	One of action, animation, comedy, drama, documentary, romance, short, mixed.
Wine quality.	
`quality`	Wine quality, as assessed by human experts (outcome).
`fixed.acidity`	Tartaric acid (g/dm^3).
`volatile.acidity`	Acetic acid (g/dm^3).
`citric.acid`	Citric acid (g/dm^3).
`residual.sugar`	Residual sugar (g/dm^3).
`chlorides`	Sodium chloride (g/dm^3).
`free.sulfur.dioxide`	Free sulfur dioxide (mg/dm^3).
`total.sulfur.dioxide`	Total sulfur dioxide (mg/dm^3).
`density`	Density (g/cm^3).
`pH`	pH.
`sulphates`	Potassium sulphate (g/dm^3).
`alcohol`	Alcohol (% by volume).

Table A.1: Description of data sets.