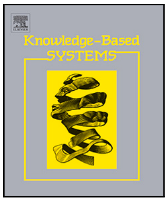
Knowledge-Based Systems 324 (2025) 113837

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Ord-MAP criterion: Extending MAP for ordinal classification

Rosario Delgado *

Department of Mathematics, Universitat Autònoma de Barcelona, Spain

ARTICLE INFO

Keywords:
Ordinal classification
MAP criterion
Probabilistic classifiers
Brier score
Continuous ranked probability score

ABSTRACT

Ordinal classification is a machine learning task where the goal is to predict labels with an inherent order. In traditional ordinal classification, the Maximum A Posteriori (MAP) criterion for assigning class labels fails to account for the ordinal structure of the target variable. We introduce Ord-MAP, a novel criterion for ordinal classification, as the most suitable extension of the binary MAP criterion to ordinal data. Unlike the usual approach, which selects the class with the highest probability — a logical choice for nominal classification — Ord-MAP identifies the first class whose cumulative probability exceeds 0.5, explicitly incorporating the class order and minimizing the expected misclassification cost under an order-sensitive loss function. This theoretical advancement addresses the fundamental limitation of existing methods by directly integrating the ordinal nature of the classes into the decision-making process.

The theoretical contribution of this study is complemented by a comprehensive empirical evaluation that includes both experiments with real-world datasets and controlled simulations, showing that Ord-MAP outperforms MAP in various scenarios, achieving statistically significant improvement in prediction. Simulations further demonstrate that this improvement is particularly noticeable for centrally located classes, with symmetric gains at both extremes of the ordinal scale. Additionally, as the Shannon entropy of the predicted probability distribution increases — indicating greater uncertainty — the difference in MAE between Ord-MAP and MAP also grows, with Ord-MAP consistently outperforming MAP under moderate to high entropy. These findings highlight the practical benefits and broad applicability of the Ord-MAP criterion, positioning it as a well-founded alternative for ordinal classification tasks.

1. Introduction

Ordinal classification, a specialized area of supervised learning, addresses problems where the target variable has a natural order but undefined distances between its levels, categories or classes. This structure frequently arises in applications such as medical diagnosis, socioeconomic studies, and quality assessments. When classes are unordered, they are represented on a *nominal scale*, as in categories such as car brands, job types, political parties, or sports. However, when classes exhibit a natural order, an *ordinal scale* is more appropriate. For instance, customer ratings on online shopping platforms — e.g., would not recommend, would recommend, or would highly recommend. This is an example of a Likert-type scale, which typically includes 3, 5, or 7 points that respondents use to express their level of agreement or disagreement with a given statement. Other typical examples include completely disagree, disagree, neutral, agree and completely agree, or always, often, and sometimes. Originally introduced by Rensis Likert [1], these scales are commonly used in sentiment analysis, satisfaction surveys, opinion mining, and information retrieval within recommender systems [2].

Although these scales provide a ranking of responses, they do not quantify the gaps between categories, which distinguishes them from *interval scales* (see [3]). A related but distinct approach is that of Ordered Qualitative Scales (OQS), as adopted in [4,5]. While OQS also refrain from assuming uniform spacing between categories, they incorporate the notion of ordinal proximity between adjacent categories, based on expert judgment. This proximity measure, however, is inherently subjective, as it relies on informed but qualitative assessments. Along similar lines, [6] propose the Multi-Thresholding meta-algorithm (MTh), an output-level methodology that implements an indirect cost-sensitive approach. MTh generalizes classical thresholding methods from binary to multiclass (or ordinal) classification by dynamically adjusting predicted class probabilities according to assumed misclassification costs, and then selecting the class with the highest adjusted value.

* Correspondence to: Department of Mathematics, Universitat Autònoma de Barcelona, Edifici C- Campus de la UAB. Av. de l'Eix Central s/n., 08193 Bellaterra (Cerdanyola del Vallès), Spain.

E-mail address: Rosario.Delgado@uab.cat.

https://doi.org/10.1016/j.knosys.2025.113837
Received 5 February 2025; Received in revised form 17 April 2025; Accepted 21 May 2025
Available online 6 June 2025

The Maximum A Posteriori (MAP) criterion, often referred to as the Bayesian criterion, is a fundamental tool in binary and multi-class classification, guiding decision-making based on probabilistic models. Rooted in Bayesian statistics and decision theory, it has played a pivotal role in shaping modern machine learning algorithms. The MAP criterion identifies the class (label) that maximizes the posterior probability given the observed evidence, providing a principled framework for classifying instances, but treats ordinal variables as categorical (nominal), potentially discarding valuable information encoded in the ordering of the classes. Formally, let $c_1, \ldots, c_r$ represent the $r \geq 2$ possible classes or labels. Suppose a probabilistic classifier assigns a posterior probability p_i to class c_i. According to the MAP criterion, the predicted class, c^*, is determined as follows: $c^* = c_k$ with

$$k = \arg \max_{i=1,\ldots,r} p_i. \tag{1}$$

The MAP criterion, while effective for nominal classification, fails in ordinal settings because it minimizes the expected loss associated with the 0–1 loss function, which treats all misclassifications equally, failing to account for the varying severity of errors in ordinal data:

$$\widetilde{\lambda}(c_j, c_i) = \begin{cases} 1 & \text{if } c_j \neq c_i \\ 0 & \text{if } c_j = c_i, \end{cases} \tag{2}$$

where c_j and c_i are the actual and predicted classes, respectively. Let p denote the probability distribution $p = (p_1, \ldots, p_r)$ over the classes provided by a classifier. If the classifier assigns class c_i, the expected loss $\widetilde{R}$ under the 0–1 loss function $\widetilde{\lambda}$ is defined as:

$$\widetilde{R}(p, c_i) = \sum_{j=1}^{r} \widetilde{\lambda}(c_j, c_i)\, p_j = \sum_{j=1,\ldots,r\,:\, j\neq i} p_j = 1 - p_i.$$

It is straightforward to observe that minimizing this expected loss is equivalent to selecting the class that maximizes the posterior probability, which is precisely what the MAP criterion does. The limitation of the MAP criterion in ordinal contexts, where misclassifications have varying severities that does not consider, is specially problematic. For example, predicting “low” as “high” is a more severe error than predicting it as “medium”. By ignoring these ordinal relationships, the MAP criterion often leads to suboptimal decisions and higher misclassification costs in ordinal tasks.

To address this limitation, we introduce the **Ord-MAP** criterion, a novel extension of the MAP criterion that leverages a loss function sensitive to the order of the classes, explicitly accounting for the ordinal structure. This framework reinterprets the MAP criterion for ordinal classification. In the binary case, the MAP criterion can be reformulated as selecting the smallest class where the cumulative probability exceeds 0.5:

$$k = \begin{cases} 1 & \text{if } p_1 \geq p_2 \\ 2 & \text{if } p_2 > p_1 \end{cases} = \begin{cases} 1 & \text{if } p_1 \geq 0.5 \\ 2 & \text{if } p_1 < 0.5 \ (p_1 + p_2 = 1 \geq 0.5) \end{cases}$$

(assuming that in case of a tie, where $p_1 = p_2$, we choose c_1, although any other tiebreaker rule could be applied). In binary classification, the MAP criterion selects the class whose probability exceeds 0.5. The traditional extension of this criterion to multi-class classification, which selects the class with the highest probability, works well for nominal problems but does not capture the ordinal nature of the data. In this paper, we propose Ord-MAP as the most appropriate extension for ordinal classification. Instead of focusing on the highest probability, Ord-MAP selects the first class whose **cumulative probability** surpasses 0.5, reflecting the order of the classes. In contrast, for nominal classification, the MAP criterion is extended as in (1). By minimizing the expected misclassification cost, Ord-MAP aligns predictions with the ordinal structure of the data. We provide rigorous theoretical justification for this criterion, demonstrating its consistency and optimality in ordinal settings.

Beyond its theoretical foundation, we validate the effectiveness of Ord-MAP through extensive experiments on diverse datasets, covering diverse domains such as elections, socioeconomic surveys, sensory evaluations, cardiovascular health, and biomedical voice analysis for Parkinson’s disease progression. We employed several advanced ordinal classifiers, including ordinal logistic regression, cumulative link models, and ordinal forests. Notably, Ord-MAP consistently outperformed the traditional MAP criterion, achieving significant reductions in the Mean Absolute Error (MAE) metric, even when classifiers were specifically designed for ordinal data. In addition, we conducted controlled simulations, which further demonstrated that Ord-MAP provides substantial improvements, particularly for centrally located classes, with symmetric gains observed as the true class moves away from the extremes of the ordinal scale. Moreover, as Shannon entropy of the simulated probability distribution increases, the MAE difference between the predictions obtained using the Ord-MAP and the MAP criteria also grows, with Ord-MAP showing progressively better performance under higher entropy conditions, reinforcing its ability to effectively handle uncertainty in ordinal classification.

State-of-the-art

Ordinal classification methods have been applied to a wide range of real-world tasks, including identifying tracheal obstructions [7], monitoring product quality [8], tracking epidemic growth [9] and assessing asthma severity levels using EEG signals [10]. From a different perspective, recent work [3] has addressed the extension of ordinal classification metrics, such as the Mean Absolute Error (MAE), for interval-based classification tasks, highlighting the importance of adapting evaluation criteria to specific data structures and problem settings.

Existing approaches to ordinal classification can be broadly categorized (following [11]) into: (a) ignoring class order using standard nominal methods, (b) reducing ordinal problems into binary tasks, through techniques like One vs Next or One vs Followers, (c) threshold-based models defining class boundaries, and (d) adapting nominal classification algorithms to ordinal settings. For example, [12] presents a framework that constructs ordinal rankers using binary classifiers (approach (b)). Notable examples of methodologies that directly embed ordinal information into classical algorithms (approximation (d)) include the adaptation of k-NN models, decision trees, random forests and AdaBoost for ordinal contexts, as demonstrated in [13,14]. These adaptations consistently outperform non-ordinal classifiers, such as Logistic Regression, Naive Bayes, and XGBoost, in practical applications like supporting students with learning impairments.

Recent advancements focus on integrating innovative strategies tailored for ordinal data. For instance, [15] introduces a neural network with custom loss functions and a threshold-based decision rule, while [7,16] enhance decision tree-based algorithms by optimizing their splitting criteria for ordinal classification. This work lies in the intersection between approximation (d) and (c). Hybrid models have also gained traction, such as ORFEO (Ordinal classifier and Regressor Fusion for Estimating an Ordinal categorical target) [17]. ORFEO uses dual outputs — one for ordinal classification and another for regression — optimized through a combined loss function, enabling finer differentiation within categories. This model has found applications in marine and ocean engineering. Similarly, [18] introduces an ordinal classification model with soft label encoding, achieving significant reductions in misclassification costs for short-term energy flux prediction compared to nominal and traditional ordinal classifiers.

Tailored ordinal methods have also demonstrated their utility in educational settings. For example, [19] proposes FlexNSLVOrd, a model designed to predict student performance in distance learning courses. Its accuracy and interpretability across ordinal categories (Withdrawn, Fail, Pass, Distinction) make it a practical tool for educational assessments.

Explainability in ordinal classification has emerged as a crucial area of focus. [20] proposes a framework that combines inductive rules with fuzzy logic to predict the likelihood of an instance belonging to each ordinal class. Tested on weather forecasting datasets (e.g., wind speed, fog-induced low visibility), this approach balances competitive performance with enhanced transparency.

Finally, resource-limited scenarios have inspired novel solutions. For instance, [21] employs decision tree- and neural network-based ordinal classifiers to generate probability matrices, which are then utilized in optimization models to minimize misclassification costs under resource constraints.

Despite these advancements, many existing methods still rely on heuristic thresholds or binary decompositions, which can restrict their performance. The proposed Ord-MAP criterion overcomes these limitations by fully utilizing the probability distributions from probabilistic classifiers, providing a principled, scalable, and robust solution for ordinal classification.

Paper contributions

This paper introduces the **Ord-MAP** criterion, making the following key contributions to the field of ordinal classification:

(i) **A novel framework for ordinal classification.** We present a new methodology that extends the MAP criterion to address the specific challenges of ordinal classification effectively. Despite its simplicity, Ord-MAP provides a powerful solution that is easy to implement and understand, making it accessible for a wide range of applications.

(ii) **Theoretical justification.** We establish rigorous theoretical foundations for the Ord-MAP criterion, demonstrating its consistency with the principles and requirements of ordinal classification.

(iii) **Comprehensive empirical validation.** We conduct extensive experiments comparing the Ord-MAP criterion with several advanced ordinal classifiers, including ordinal logistic regression, cumulative link models, and ordinal forests. Our results, using the Mean Absolute Error (MAE) metric, reveal significant performance improvements across some real-world datasets with respect to the traditional MAP criterion.
We also include a simulation study using synthetic probability distributions, which highlights the consistent advantage of Ord-MAP, particularly for centrally located classes. The MAE difference between using the Ord-MAP and the MAP criteria grows as the Shannon entropy of the probability distribution increases, further emphasizing the robustness of Ord-MAP under higher uncertainty conditions.

(iv) **Explainability and Transparency.** We emphasize the interpretability of Ord-MAP predictions, making it particularly valuable for critical and sensitive domains such as healthcare and education.

Organization of the paper

The remainder of this paper is structured as follows: Section 2 recalls the MAP criterion and introduces some notations and fundamental concepts, as the *scoring rules*. Section 3 introduces the Ord-MAP criterion, and in Section 4 we state and prove the main theoretical result, which is an optimality property of the Ord-MAP criterion (Theorem 1). Section 5 describes the datasets and experimental design, outlines the diverse methodologies employed, and presents the results, highlighting the consistent improvements achieved by the Ord-MAP criterion over MAP across various scenarios. The corresponding tables and figures have been placed in Appendix B. Section 6 presents a simulation study based on synthetic ordinal probability distributions, providing additional insights into the behavior of the proposed criterion under controlled conditions. Finally, Section 7 discusses the implications of these findings and concludes with potential avenues for future research. In Appendix A we provide the proof of the properness for the two scores used in the paper: Brier score (MAP criterion) and discretized version of the Continuous Ranked Probability score (Ord-MAP criterion).

2. MAP criterion: Performance metrics and scoring

The *error rate*, also known as the *misclassification rate*, is a key evaluation metric in classification tasks, measuring the proportion of incorrect predictions out of the total instances in a dataset. It is complementary to accuracy, that is: *error rate* $= 1 - accuracy$. Given a dataset $\mathcal{D} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$, where $\mathbf{x}_i$ represents the feature vector for the ith instance and $y_i \in \{c_1, \ldots, c_r\}$ is its true label, the error rate is calculated as:

$$\text{Error rate } = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\lambda}(\widehat{y}_i, y_i)$$

where $\widehat{y}_i \in \{c_1, \ldots, c_r\}$ is the predicted label for the ith instance, n is the total number of instances, and $\widetilde{\lambda}$ is the 0–1 loss function defined in (2). As this loss function assigns a value of 1 for incorrect predictions and 0 for correct ones, it makes the *error rate* an aggregate measure of the individual losses across the dataset.

While the MAP criterion itself does not require further formalization, the introduction of a scoring rule can provide additional insights and facilitate its adaptation to ordinal classification to account for the inherent order among classes. To address this limitation, we introduce scoring rules in Section 3, enabling ordinal-specific evaluation.

2.1. Scoring rules

The concept of scoring rules was introduced by Savage in the seminal work [22], which provided the theoretical foundation for assessing probabilistic predictions. Savage emphasized the importance of coherent probability assessments in decision-making and statistical inference. Scoring rules evaluate the quality of probabilistic predictions by assigning a numerical scored based on the predicted probability distribution $p = (p_1, \ldots, p_r)$ and the true class. A widely used scoring rule for nominal classification is the Brier score [23], denoted by $\widetilde{S}$ and defined as:

$$\widetilde{S}(p, c_i) = -\sum_{j=1}^{r} \left(p_j - 1_{\{c_i = c_j\}}\right)^2. \tag{3}$$

where $1_{\{\ldots\}}$ is the indicator function. The Brier score quantifies the accuracy of probabilistic predictions and satisfies two desirable properties. First, let us establish some notations: for a fixed number of classes $r \geq 2$, let

$$\mathcal{P}_r = \{p = (p_1, \ldots, p_r) \; : \; p_i \geq 0, \; \sum_{i=1}^{r} p_i = 1\}, \tag{4}$$

be the simplex of valid probability distributions, which is a convex set. This means that if $p, q \in \mathcal{P}_r$ and $\alpha \in (0, 1)$, then $(1 - \alpha)\,p + \alpha\, q \in \mathcal{P}_r$. By the definition of score S, for a fixed $i \in \{1, \ldots, r\}$ we can define a function

$$\widetilde{S}(\cdot, c_i) \; : \; \mathcal{P}_r \longrightarrow \mathbb{R}^- = (-\infty, 0]$$

by $\widetilde{S}(p, c_i) = -\sum_{j=1}^{r} \left(p_j - 1_{\{c_i = c_j\}}\right)^2$ for any $p \in \mathcal{P}_r$.

- **Regularity**: The score $\widetilde{S}(p, c_i)$ is finite for all i, except possibly $\widetilde{S}(p, c_i) = +\infty$ or $-\infty$ if $p_i = 0$ (see Definition 1 in [24]). Indeed, $\widetilde{S}(p, c_i) \leq 0$ for all i and p, since it is defined through a negative finite sum of non-negative terms.
- **Propriety:** $\widetilde{S}$ is **proper** relative to $\mathcal{P}_r$. By definition (see (1) in [24]), this means that

  $$\widetilde{S}(p, p) \geq \widetilde{S}(p, q) \quad \forall p, q \in \mathcal{P}_r$$

  where by definition,

  $$\widetilde{S}(p, q) = \sum_{i=1}^{r} \widetilde{S}(p, c_i)\, q_i \quad \text{if} \quad q = (q_1, \ldots, q_r).$$

The rationale of this property is that the expected reward generated by the classifier if it quotes the probability distribution p over the classes, with respect to p (as the true probability distribution for the actual class), is greater than with respect to $q \neq p$. Proper scores incentivize more accurate classifiers by rewarding those that succeed in correctly predicting the actual class, and are designed such that the expected score is maximized when the predicted probability distribution over the classes provided by the classifier matches the actual probability distribution. Formal proof that $\widetilde{S}$ is **proper** can be found in Appendix A.

In recent years, other proper scoring rules have been developed to evaluate probabilistic forecasts for specific tasks. For example, [25] introduces scoring rules like the Weighted Interval Score (WIS) for quantile forecasts and the Ranked Probability Score (RPS) for integer forecasts. Additionally, [26] provides an extensive review of scoring functions, with a focus on hydrological applications, detailing their use for dichotomous events, categorical forecasts, and density distributions. These studies underscore the importance of propriety in ensuring meaningful evaluation metrics.

2.2. Relation between the MAP criterion, the 0–1 loss function, and the brier score

The Brier score is closely linked to the 0–1 loss function in two fundamental ways. First, consider a deterministic forecast, i.e., when $p = (0, \ldots, 0, 1^{j)}, 0 \ldots, 0)$ for some $j = 1, \ldots, r$, in which case it is straightforward to verify that $\widetilde{S}(p, c_i) = -2\,\widetilde{\lambda}(c_i, c_j)$. Indeed,

$$\widetilde{S}(p, c_i) = -\sum_{\ell=1}^{r} \left(p_\ell - 1_{\{c_i=c_\ell\}}\right)^2 = -\sum_{\ell\neq j} 1_{\{c_i=c_\ell\}} - \left(1 - 1_{\{c_i=c_j\}}\right)^2$$
$$= \begin{cases} 0 & \text{if } c_i = c_j \\ -2 & \text{if } c_i \neq c_j \end{cases}$$

This result reveals a direct correspondence between the Brier score and the 0–1 loss. Consequently, there is a close connection between Brier score and Error rate, understood as the average of the individual 0–1 losses over the dataset. Second, the Brier score is also related to the expected loss $\widetilde{R}$ (defined in terms of the loss function $\widetilde{\lambda}$), further reinforcing its intrinsic link with classification error through the 0–1 loss function. In fact, the MAP decision rule, which minimizes $\widetilde{R}$ (see Section 1), can also be characterized as the rule that maximizes the Brier score. To demonstrate this, consider two distinct indices $i, j \in \{1, \ldots, r\}$, with $i \neq j$. Then,

$$\widetilde{S}(p, c_i) = -\Big(\sum_{\ell\neq i,j} p_\ell^2 + p_j^2 + (1-p_i)^2 \Big) \text{ and}$$
$$\widetilde{S}(p, c_j) = -\Big(\sum_{\ell\neq i,j} p_\ell^2 + (1-p_j)^2 + p_i^2 \Big).$$

It follows that $\widetilde{S}(p, c_i) \leq \widetilde{S}(p, c_j) \iff p_j^2 + (1-p_i)^2 \geq p_i^2 + (1-p_j)^2 \iff p_i \leq p_j$. Thus, the MAP criterion — which selects the class with the highest probability, that is, the **mode** — not only minimizes the expected 0–1 loss $\widetilde{R}$, but also maximizes the Brier score $\widetilde{S}$. Fig. 1 provides a schematic illustration of the MAP criterion, showing how it relates to the 0–1 loss function $\widetilde{\lambda}$, the expected loss $\widetilde{R}$, and the Brier score $\widetilde{S}$, and emphasizing its role in optimal class assignment.

3. The ord-MAP criterion

The Ord-MAP criterion builds upon the discretized version of the Continuous Ranked Probability Score (CRPS), denoted S. The CRPS is a scoring rule well-suited for ordinal classification, as it leverages the cumulative distribution function to account for the inherent ordering of classes. This approach encourages meaningful predictions by rewarding classifiers that produce probability distributions aligned with actual outcomes.

This section is structured as follows: the discretized CRPS S is introduced in Section 3.1, followed by the formal definition of the Ord-MAP criterion in Section 3.2.

3.1. The continuous ranked probability score (CRPS)

Consider $1 < \cdots < r$ as the ordered classes, with $r \geq 2$ (emphasizing ordering over notation such as $c_1, \ldots, c_r$). Note that ordering becomes meaningful only if $r \geq 3$. Let $p_i \geq 0$ represent the posterior probability assigned to class i by a probabilistic classifier, and $p = (p_1, \ldots, p_r)$ denote the full probability distribution.

Definition 1. The discretized version of the CRPS for a probability distribution $p = (p_1, \ldots, p_r)$ and an actual class $k = 1, \ldots, r$, is defined as:

$$S(p, k) = -\sum_{j=1}^{r} \Big((p_1 + \cdots + p_j) - 1_{\{k\leq j\}} \Big)^2. \tag{5}$$

The score $S(p, k)$ reflects the reward for quoting p when the actual class is k. Due to $p_1 + \cdots + p_r = 1$ and $1_{\{k\leq r\}} = 1$ for any k, the last term in the summation (5) is always zero, allowing the sum to be computed up to $j = r - 1$.

Remark 1. The CRPS, in its continuous form, is widely used for probabilistic forecasts of continuous variables (see for instance [27]). It evaluates the negative integral of squared differences between the predicted cumulative distribution function $F(x)$ and the cumulative empirical distribution function of the observation y:

$$-\int_{\mathbb{R}} \Big(F(x) - 1_{\{y\leq x\}} \Big)^2 \, dx.$$

Properties of the discretized CRPS

To validate the usefulness of S, the **Regularity** and **Propriety** properties are required. The first of these properties is trivially justified, as in the case of the Brier score (Section 2). The second is proven in Appendix A.

3.2. The ord-MAP criterion

We now define the **Ord-MAP** criterion, an extension of the MAP criterion adapted for ordinal classification:

Definition 2. Given the probability distribution $p = (p_1, \ldots, p_r)$, the **Ord-MAP** criterion selects the class k^* that maximizes the discretized version of CRPS:

$$k^* = \arg\max_{k=1,\ldots,r} S(p, k), \quad \text{where } S(p, k) \text{ is introduced in (5).}$$

Remark 2. When $r = 2$ (binary case), Ord-MAP is equivalent to MAP, since the ordering between two classes is trivial. Indeed, for $k = 1, 2$, we have by (5) that

$$S(p, k) = -\big(p_1 - 1_{\{k\leq 1\}} \big)^2 - \big((p_1 + p_2) - 1_{\{k\leq 2\}} \big)^2 = -\big(p_1 - 1_{\{k\leq 1\}} \big)^2,$$

which can be written as

$$S(p, k) = \begin{cases} -(p_1 - 1)^2 = -p_2^2 & \text{if } k = 1 \\ -p_1^2 & \text{if } k = 2 \end{cases}$$

and thus,

$$\arg\max_{k=1,2} S(p, k) = \begin{cases} 1 & \text{if } p_2 < p_1 \\ 2 & \text{if } p_2 > p_1, \end{cases}$$

which coincides with the MAP criterion (setting aside the issue of a tie that would occur if $p_1 = p_2$, in which case a tiebreaker rule must be provided). Equivalently, by (3) we can easily check that

$$\widetilde{S}(p, k) = -(p_1 - 1_{\{k=1\}})^2 - (p_2 - 1_{\{k=2\}})^2$$
$$= \begin{cases} -(p_1 - 1)^2 - p_2^2 = -2\,p_2^2 & \text{if } k = 1 \\ -p_1^2 - (p_2 - 1)^2 = -2\,p_1^2 & \text{if } k = 2 \end{cases}$$

and also

$$\arg\max_{k=1,2} \widetilde{S}(p, k) = \begin{cases} 1 & \text{if } p_2 < p_1 \\ 2 & \text{if } p_2 > p_1. \end{cases}$$

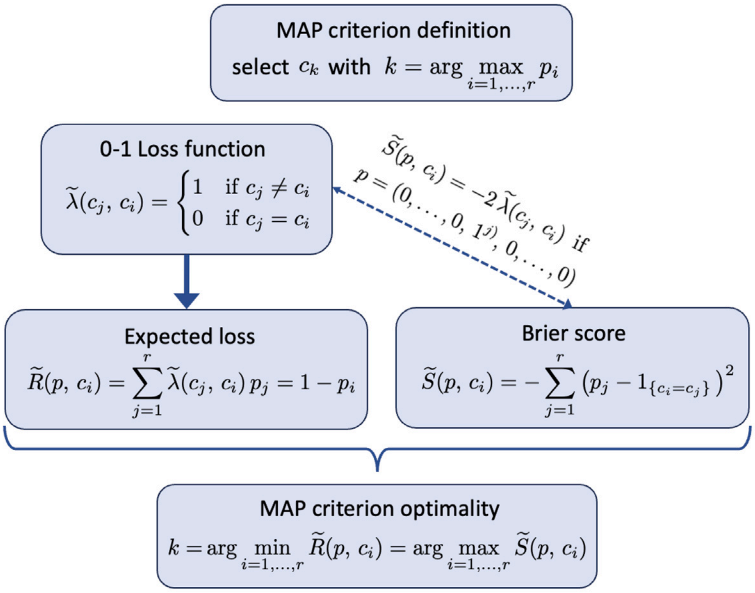

MAP criterion definition
select c_k with $k = \arg\max_{i=1,\dots,r} p_i$
0-1 Loss function
$\widetilde{\lambda}(c_j, c_i) = \begin{cases} 1 & \text{if } c_j \neq c_i \\ 0 & \text{if } c_j = c_i \end{cases}$
$\widetilde{S}(p, c_i) = -2\widetilde{\lambda}(c_j, c_i)$ if $p = (0, \dots, 0, 1^{j)}, 0, \dots, 0)$
Expected loss
$\widetilde{R}(p, c_i) = \sum_{j=1}^{r} \widetilde{\lambda}(c_j, c_i)\, p_j = 1 - p_i$
Brier score
$\widetilde{S}(p, c_i) = -\sum_{j=1}^{r} \left(p_j - 1_{\{c_i = c_j\}}\right)^2$
MAP criterion optimality
$k = \arg\min_{i=1,\dots,r} \widetilde{R}(p, c_i) = \arg\max_{i=1,\dots,r} \widetilde{S}(p, c_i)$

Fig. 1. Schematic representation of the MAP criterion for nominal classification, given the class probability distribution $p = (p_1, \dots, p_r)$, highlighting its relationship with expected 0–1 loss and the Brier score.

Remark 3. If $r > 2$, Ord-MAP is sensitive to class ordering, distinguishing it from MAP. The relationship between the two criteria and their scoring functions is as follows, in the simplest case $r = 3$: by (5), for $k = 1, 2, 3$,

$$S(p, k) = -\left(p_1 - 1_{\{k \le 1\}}\right)^2 - \left((p_1 + p_2) - 1_{\{k \le 2\}}\right)^2,$$

while

$$\begin{aligned} \widetilde{S}(p, k) &= -(p_1 - 1_{\{k=1\}})^2 - (p_2 - 1_{\{k=2\}})^2 - (p_3 - 1_{\{k=3\}})^2 \\ &= -(p_1 - 1_{\{k=1\}})^2 - (p_2 - 1_{\{k=2\}})^2 - ((p_1 + p_2) - 1_{\{k \le 2\}})^2. \end{aligned}$$

Then, it is evident that $\widetilde{S}(p, k) \le S(p, k)$. Moreover, the inequality is strict for $k = 2$ if $p_2 < 1$, and for $k \neq 2$ if $p_2 > 0$. Indeed,

$$S(p, k) - \widetilde{S}(p, k) = (p_2 - 1_{\{k=2\}})^2 = \begin{cases} (p_2 - 1)^2 & \text{if } k = 2 \\ p_2^2 & \text{if } k \neq 2. \end{cases}$$

In Section 4, we will introduce the loss function λ associated with S, along with its inherently linked performance metric (MAE), and expected loss R. Theorem 1 gives the theoretical justification and practical implementation of the Ord-MAP criterion.

4. Optimality of the ord-MAP criterion

We mentioned in the introduction that the MAP criterion is optimal in minimizing the expected loss under the 0–1 loss function. Can we establish a similar property for the Ord-MAP criterion? The answer is yes, with a different loss function, denoted by λ, defined as follows: if $j, k \in \{1, \dots, r\}$,

$$\lambda(j, k) = |j - k|.$$

where $|x|$ denotes the absolute value of $x \in \mathbb{R}$. The *Mean Absolute Error* (MAE) metric, widely used for ordinal classification, is defined as the average absolute difference between the ground truth labels y_i and the predicted labels $\hat{y}_i$, both in $\{1, \dots, r\}$:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} \lambda(\hat{y}_i, y_i)$$

where n is the total number of instances in the dataset. For example, in [8] the authors experimentally show that for the studied unbalanced dataset, MSE (Mean Squared Error) and MAE perform the best, but while MSE is better in situations where the severity of the errors is more important, MAE shows to be better in situations where the tolerance for small errors is lower. In [28] the authors use MAE as a performance metric for monotonic ordinal in classification, to show the usefulness of selecting the training set to obtain more accurate and efficient models. More recently, MAE is employed in tasks such as medical diagnosis and face age prediction [29], and to estimate the performance of a deep convolutional neural network model for ordinal classification [30].

λ measures the loss for individual predictions, assigning a loss which is the difference in absolute value between the predicted and the observed labels. Then, MAE penalizes classification errors proportionally to the distance between the categories, so the lower the metric value, the better the performance of the classifier. Unlike MAE, which is computed directly from the confusion matrix and reflects the average absolute deviation between predicted and true class labels, other metrics are based on the concordance of pairwise comparisons, as the c-index (see [31]) and the Kendall's tau-b [32]. These metrics have the advantage of being purely ordinal and do not assume equal spacing between classes. However, they do not offer a direct measure of the magnitude of prediction errors as MAE does. Moreover, as we will see in Section 4.1, MAE is intrinsically linked to the loss function λ and the (discrete) CRPS score S used in our approach — just as the Error Rate is closely tied to the 0–1 loss function and the Brier score in the nominal setting (Section 2.2). This alignment ensures a consistent theoretical framework connecting prediction, loss, and evaluation, underpinning the theoretical coherence of using MAE as evaluation metric.

The expected loss R is defined as the expected value of λ given the posterior probability distribution $p = (p_1, \dots, p_r)$ provided by the classifier:

$$R(p, k) = \sum_{j=1}^{r} \lambda(j, k)\, p_j = \sum_{j=1,\dots,r\,:\, j \neq k} |j - k|\, p_j$$

This expected loss represents the expected penalty associated with selecting class k as the prediction.

4.1. Relation between the ord-MAP criterion, the λ loss function, and the (discrete) CRPS score S

Analogous to the relationship between the Brier score and the 0–1 loss function, the (discrete) CRPS score S is closely linked to the λ

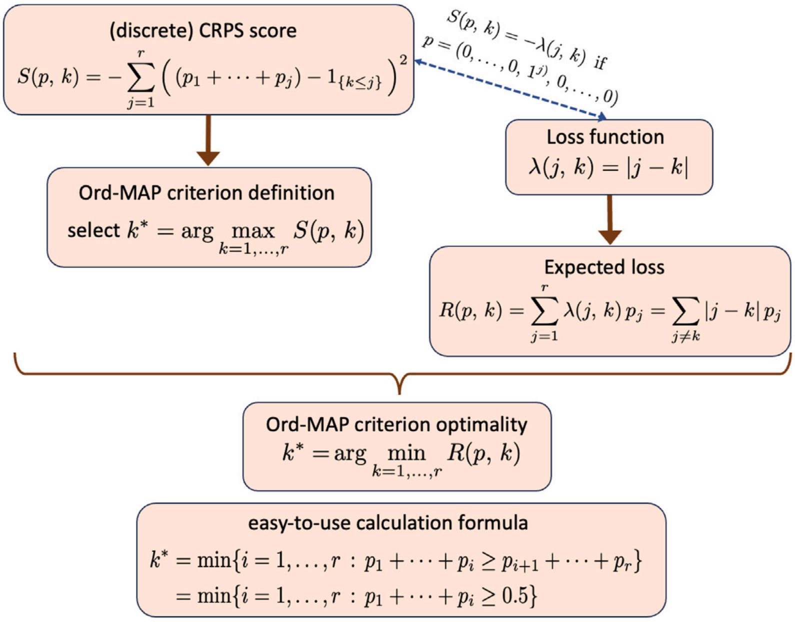

(discrete) CRPS score
$S(p,k) = -\sum_{j=1}^{r}\Big((p_1+\cdots+p_j) - 1_{\{k\le j\}}\Big)^2$
$S(p,k) = -\lambda(j,k)$ if $p = (0,\dots,0,1^{j)},0,\dots,0)$
Loss function
$\lambda(j,k) = |j-k|$
Ord-MAP criterion definition
select $k^* = \arg\max_{k=1,\dots,r} S(p,k)$
Expected loss
$R(p,k) = \sum_{j=1}^{r}\lambda(j,k)\,p_j = \sum_{j\ne k}|j-k|\,p_j$
Ord-MAP criterion optimality
$k^* = \arg\min_{k=1,\dots,r} R(p,k)$
easy-to-use calculation formula
$k^* = \min\{i=1,\dots,r : p_1+\cdots+p_i \ge p_{i+1}+\cdots+p_r\}$
$= \min\{i=1,\dots,r : p_1+\cdots+p_i \ge 0.5\}$

Fig. 2. Schematic representation of the Ord-MAP criterion for ordinal classification, given the class probability distribution $p = (p_1,\dots,p_r)$, highlighting its relationship with the expected loss and the (discrete) CRPS score.

loss function in two key ways. First, for a deterministic forecast $p = (0,\dots,0,1^{j)},0\dots,0)$, it can be shown that $S(p,k) = -\lambda(j,k)$ as follows:

$$
\begin{aligned}
S(p,k) &= -\sum_{\ell=1}^{r}\Big((p_1+\cdots+p_\ell) - 1_{\{k\le\ell\}}\Big)^2 \\
&= -\sum_{\ell=1}^{j-1} 1_{\{k\le\ell\}} - \sum_{\ell=j}^{r}\big(1-1_{\{k\le\ell\}}\big)^2 \\
&= \begin{cases} -\sum_{\ell=k}^{j-1} 1 = k-j = -|j-k| & \text{if } k<j \\ 0 & \text{if } k=j \\ -\sum_{\ell=j}^{k-1} 1 = j-k = -|j-k| & \text{if } k>j. \end{cases}
\end{aligned}
$$

(As usual, if the upper summation limit is smaller than the lower limit, the summation is understood to be omitted from the expression.) This result reveals a direct correspondence between the score S and the loss function λ, as well as with the MAE metric, which corresponds to the average λ-loss across the dataset. Second, the score S is also related to the expected loss R defined in terms of λ since the Ord-MAP decision rule — defined by maximization of S — also minimizes R, as formally stated in Section 4.2 below.

The main result in Section 4.2, which is Theorem 1, provides a practical implementation of the Ord-MAP rule: **the optimal class is that corresponding to the median of the posterior probability distribution**, in contrast to the MAP rule, which selects the **mode**. This result states that Ord-MAP not only maximizes the (discrete) CRPS score, but also minimizes the expected λ-loss (see Fig. 2).

4.2. The main result

In this section we state the main theoretical result, which is Theorem 1. Its proof relies on Lemma 1. We provide the proof of both results.

Theorem 1. *Let $r \ge 2$ and $p \in \mathcal{P}_r$. The class k^* that maximizes the score S (Definition 2) also minimizes the expected loss R:*

$$k^* = \arg\min_{k=1,\dots,r} R(p,k).$$

Furthermore, k^* can be computed using the following formula:

$$k^* = \min\{i=1,\dots,r \;:\; p_1+\cdots+p_i \ge p_{i+1}+\cdots+p_r\}, \tag{6}$$

where if $i = r$, it is understood that $p_{i+1}+\cdots+p_r = 0$. Equivalently,

$$k^* = \min\{i=1,\dots,r \;:\; p_1+\cdots+p_i \ge 0.5\}. \tag{7}$$

(Note that, since $p_1+\cdots+p_r = 1 \ge 0.5 \ge 0$, both sets $\{i=1,\dots,r \;:\; p_1+\cdots+p_i \ge p_{i+1}+\cdots+p_r\}$ in (6) and $\{i=1,\dots,r \;:\; p_1+\cdots+p_i \ge 0.5\}$ in (7), are non-empty as they both contain r. Moreover, since they are finite, their minimum values are well-defined.)

The proof of Theorem 1 will proceed from the following lemma.

Lemma 1. *Given fixed $p = (p_1,\dots,p_r) \in \mathcal{P}_r$, let $\varphi(p,s)$ be a non-positive function of $s = 1,\dots,r$ depending on p. For any $k = 1,\dots,r-1$ and $\delta = 1,\dots,r-k$, define*

$$\Delta(p,k,\delta) = \varphi(p,k+\delta) - \varphi(p,k).$$

Assume that there exists $h = 1,\dots,r$ such that the following hypotheses hold:

- (h_1) *If $h < r$, $\Delta(p,h,\delta) \le 0$ for all $\delta = 1,\dots,r-h$.*
- (h_2) *If $h > 1$, $\Delta(p,\ell,1) > 0$ for all $\ell = 1,\dots,h-1$.*

Therefore,

$$\arg\max_{\ell=1,\dots,r} \varphi(p,\ell) = h.$$

(With the convention that we take the minimum of the points where the maximum is achieved, if there are multiple such points.)

Proof (*Lemma* 1)**.** Hypothesis (h_1) implies that if $h = 1,\dots,r-1$,

$$\varphi(p,h) \ge \varphi(p,\ell) \quad \forall \ell = h+1,\dots,r$$

(by setting $\ell = h+\delta$). According to hypothesis (h_2), if $h = 2,\dots,r$, we have $\varphi(p,h) > \varphi(p,h-1)$ (setting $\ell = h-1$), $\varphi(p,h-1) > \varphi(p,h-2)$ (setting $\ell = h-2$), and so forth, until we reach $\varphi(p,2) > \varphi(p,1)$ (setting $\ell = 1$). Concatenating these conclusions, we obtain

$$\varphi(p,h) > \varphi(p,\ell) \quad \forall \ell = 1,\dots,h-1.$$

This directly implies the thesis of the lemma. □

Proof (*Theorem* 1)**.** We assume without loss of generality that $r > 2$, since for $r = 2$, the Ord-MAP and MAP criteria coincide. Indeed, if $r = 2$,

$$R(p,k) = \sum_{j=1,2,\, j\neq k} |j-k|\, p_j = \begin{cases} p_1 & \text{if } k = 2 \\ p_2 & \text{if } k = 1. \end{cases}$$

Then, the class k^* that minimizes R is:

$$k^* = \begin{cases} 1 & \text{if } p_1 > p_2 \\ 2 & \text{if } p_1 < p_2, \end{cases}$$

which coincides with the class that maximizes S, by Remark 2, and also verifies (6).

To prove the theorem, we will apply Lemma 1 to both the score S and minus the expected loss $-R$, considered as functions φ. Indeed, by verifying that both S and $-R$ satisfy hypotheses (h_1) and (h_2) with $h = k^*$ given by (6), Lemma 1 ensures that S achieves the maximum and R the minimum, at the same class, which is k^*.

Step 1: $\varphi(p,\, k) = S(p,\, k)$ satisfies (h_1) and (h_2) in Lemma 1 with $h = k^*$.
Indeed, by definition,

$$S(p,\, k) = -\sum_{j=1}^{r}\Big((p_1 + \cdots + p_j) - 1_{\{k\leq j\}} \Big)^2$$
$$= -\sum_{j=1}^{k-1}(p_1 + \cdots + p_j)^2 - \sum_{j=k}^{r-1}(p_{j+1} + \cdots + p_r)^2.$$

As a consequence,

$$\Delta(p,k,\delta) = S(p,k{+}\delta){-}S(p,\,k) = -\sum_{j=k}^{k+\delta-1}(p_1{+}\cdots{+}p_j)^2{+}\sum_{j=k}^{k+\delta-1}(p_{j+1}{+}\cdots{+}p_r)^2.$$

First, note that if $j < r$ belongs to the set $\{i = 1, \ldots, r \,:\, p_1 + \cdots + p_i \geq p_{i+1} + \cdots + p_r\}$, then any h in the range $j+1, \ldots, r$ also belongs to this set. Therefore, by definition of k^* (see (6)):

- $p_1 + \cdots + p_j \geq p_{j+1} + \cdots + p_r$ for all $j \geq k^*$ if $k^* < r$. This implies $\Delta(p,\, k^*,\, \delta) \leq 0$ for all $\delta = 1, \ldots, r{-}k^*$. That is, (h_1) is accomplished with $h = k^*$.
- $p_1 + \cdots + p_j < p_{j+1} + \cdots + p_r$ for all $j < k^*$ if $k^* > 1$, since k^* it is the minimum accomplishing the reverse inequality. This implies $\Delta(p,\, \ell,\, 1) > 0$ for all $\ell < k^*$, and therefore, (h_2) is satisfied with $h = k^*$.

Step 2: $\varphi(p,\, k) = -R(p,\, k)$ satisfies (h_1) and (h_2) in Lemma 1 with $h = k^*$.
Again, by definition,

$$-R(p,\, k) = -\sum_{j=1}^{r} |k-j|\, p_j = -\sum_{j=1}^{k-1}(k-j)\, p_j - \sum_{j=k+1}^{r}(j-k)\, p_j, \quad \text{and}$$

$$\Delta(p,\, k,\, \delta) = -R(p,\, k+\delta) + R(p,\, k)$$
$$= -\sum_{j=1}^{k+\delta-1}(k+\delta-j)p_j + \sum_{j=1}^{k-1}(k-j)p_j - \sum_{j=k+\delta+1}^{r}(j-(k+\delta))p_j$$
$$+ \sum_{j=k+1}^{r}(j-k)p_j$$
$$= -\sum_{j=1}^{k-1}\delta\, p_j - \sum_{j=k}^{k+\delta-1}(k+\delta-j)\, p_j + \Big(\sum_{j=k+\delta+1}^{r}\delta\, p_j + \sum_{j=k+1}^{k+\delta}(j-k)\, p_j\Big).$$

Finally, this can be written

$$\Delta(p,\, k,\, \delta) = -\delta\Big(\sum_{j=1}^{k-1}p_j - \sum_{j=k+\delta-1}^{r}p_j\Big)$$
$$- \Big(\delta\,(p_k - p_{k+\delta}) + \sum_{j=k+1}^{k+\delta-1}\big(\delta-2\,(j-k)\big)\, p_j\Big)$$
$$= -\delta\Big(\sum_{j=1}^{k}p_j - \sum_{j=k+\delta}^{r}p_j\Big)$$
$$- \Big(\sum_{j=k+1}^{k+\lfloor\frac{\delta}{2}\rfloor}\big(\delta-2\,(j-k)\big)\, p_j + \sum_{j=k+\lfloor\frac{\delta}{2}\rfloor+1}^{k+\delta-1}\big(\delta-2\,(j-k)\big)\, p_j\Big). \qquad (8)$$

Note that in the last summands, when they are meaningful, $\delta - 2\,(j-k) \geq 0$ when $j = k+1, \ldots, k+\lfloor\frac{\delta}{2}\rfloor$, while $\delta - 2\,(j-k) < 0$ when $j = k + \lfloor\frac{\delta}{2}\rfloor + 1, \ldots,\, k+\delta-1$. In particular, they are not meaningful when $\delta = 1$.

By definition of k^*, if $k^* < r$, $p_1{+}\cdots{+}p_{k^*} \geq p_{k^*+1}{+}\cdots{+}p_r$. Multiplying by δ,

$$\delta\,(p_1 + \cdots + p_{k^*}) \geq \delta\,(p_{k^*+1} + \cdots + p_{k^*+\delta-1}) + \delta\,(p_{k^*+\delta} + \cdots + p_r)$$
$$\geq \sum_{j=k^*+\lfloor\frac{\delta}{2}\rfloor+1}^{k^*+\delta-1}\delta\, p_j + \delta\,(p_{k^*+\delta} + \cdots + p_r)$$
$$\geq \sum_{j=k^*+\lfloor\frac{\delta}{2}\rfloor+1}^{k^*+\delta-1} -\big(\delta-2\,(j-k^*)\big)\, p_j + \delta\,(p_{k^*+\delta} + \cdots + p_r)$$

since $\delta \geq -(\delta{+}2\,k^*{-}2\,j) \Leftrightarrow j \leq k^*{+}\delta$. That is, we have demonstrated

$$\delta\Big(\sum_{j=1}^{k^*}p_j - \sum_{j=k^*+\delta}^{r}p_j\Big) + \sum_{j=k^*+\lfloor\frac{\delta}{2}\rfloor+1}^{k^*+\delta-1}\big(\delta-2\,(j-k^*)\big)\, p_j \geq 0.$$

Therefore, by (8), we have proved that $\Delta(p,\, k^*,\, \delta) \leq 0$ if $k^* < r$, for all $\delta = 1, \ldots,\, r - k^*$, which is hypothesis (h_1), considering that

$$\sum_{j=k^*+1}^{k^*+\lfloor\frac{\delta}{2}\rfloor}\big(\delta-2\,(j-k^*)\big)\, p_j \geq 0.$$

On the other hand, for $k^* > 1$ and $h = 1, \ldots,\, k^*-1$, by definition of k^*, $p_1 + \cdots + p_h < p_{h+1} + \cdots + p_r$. Consequently, by (8), taking into account that the last two summands are not meaningful if $\delta = 1$ and therefore disappear from the expression, we obtain that

$$\Delta(p,\, \ell,\, 1) = -\sum_{j=1}^{\ell}p_j + \sum_{j=\ell+1}^{r}p_j > 0,$$

which is hypothesis (h_2). □

5. Experimentation with real-world data

This section describes the experimental phase conducted to assess the performance of the Ord-MAP criterion in ordinal classification tasks. Five real-world datasets were used for evaluation.

5.1. Datasets

(a) **World Values Surveys (WVS) dataset**. The WVS dataset, sourced from the carData R package,[1] comprises 5381 observations across six variables: religion, university degree, country, age and gender, and the target variable, Poverty. The target is an ordinal variable with three ordered categories: Too Little, About Right, Too Much.

(b) **Wine dataset**. The Wine dataset, available in the ordinal R package[2] contains 72 observations from wine assessments conducted by nine judges. Predictor variables include temperature, contact between juice and the skins, perceived bitterness, and identifiers for the bottle and the judge. The target variable, Rating, is ordinal with five levels (1–5).

[1] Datasets to accompany J. Fox and S. Weisberg, An R Companion to Applied Regression, Third Edition, Sage (2019). https://doi.org/10.32614/CRAN.package.carData

[2] https://doi.org/10.32614/CRAN.package.ordinal

(c) **Hearth dataset**. This dataset, included in the `ordinalForest` R package,[3] provides data on 294 patients who underwent angiography at the Hungarian Institute of Cardiology, Budapest, from 1983 to 1987. It consists of 10 covariates, including age, sex, blood pressure, cholesterol levels, and other medical features, and an ordinal target, `Class`, which reflects the severity of coronary artery disease. The target variable has five categories: 1 (`no disease`), 2 (`degree 1`), 3 (`degree 2`), 4 (`degree 3`), 5 (`degree 4`).

(d) **Parkinson dataset**. This dataset, available from the UC Irvine Machine Learning Repository,[4] includes biomedical voice measurements collected from 42 individuals with early-stage Parkinson's disease over six months for tele-monitoring purposes. It contains 5875 voice recordings described by 16 biomedical voice features (variables 7–22), and has been used in [3] for experimental purposes in the context of interval scale classification, which falls outside the scope of this research. We focus on two target variables, which are continuous and have been discretized to 4, 5 and 6 categories:

- `v5:motor_UPDRS` (clinician's motor Unified Parkinson's Disease Rating Scale). This scale is a widely used score to track disease progression.
  4 categories: < 15, $[15, 20)$, $[20, 30)$, ≥ 30
  5 categories: < 13, $[13, 18)$, $[18, 24)$, $[24, 29)$, ≥ 29
  6 categories: < 10, $[10, 15)$, $[15, 20)$, $[20, 25)$, $[25, 30)$, ≥ 30
- `v6:total_UPDRS` (clinician's total Unified Parkinson's Disease Rating Scale).
  4 categories: < 22, $[22, 30)$, $[30, 40)$, ≥ 40
  5 categories: < 20, $[20, 25)$, $[25, 30)$, $[30, 40)$, ≥ 40
  6 categories: < 20, $[20, 25)$, $[25, 30)$, $[30, 35)$, $[35, 45)$, ≥ 45

(e) **2011 Canadian National Election Study (CES11) dataset**. This dataset is also sourced from the `carData` R package, and is drawn from the 2011 Canadian National Election Study. It comprises 2231 observations of the following variables: province, weight, gender, abortion, education, urban, and the target variable, `importance`, corresponding to the importance assigned to religion, which is ordinal with four ordered categories: `not`, `notvery`, `somewhat`, `very`.

5.2. Experimental design

To evaluate the performance of the Ord-MAP criterion, we implemented different experimental procedures and apply them to any dataset:

- **Procedure 1:** Ordinal logistic regression using the `polr` function from the `MASS` R package. This function fits a regression model to an ordered factor response: the proportional odds logistic regression model or its variants (probit, log–log, complementary log–log, and Cauchy). A total of five models were constructed, each employing a different link function (probit, loglog, cloglog and cauchit, respectively). The thresholds (cut-points) were treated as flexible by default, imposing no restrictions on the distances between them.
- **Procedure 2:** Cumulative link models (CLMs), implemented via the `clm` function from the `ordinal` R package. CLMs support a range of link functions (logit, probit, log–log, complementary log–log, and Cauchy) and four threshold structures:
  - **Flexible**: No constraints on the threshold spacing (similar to `polr`).
  - **Equidistant**: Equal spacing between thresholds.
  - **Symmetric**: Thresholds symmetrically distributed around a central value.
  - **Symmetric2**: Similar to *symmetric* but with the latent mean of the reference group fixed at zero.

  This setup allowed us to explore the interaction between data characteristics captured by link functions and threshold structures. A total of 20 models (5 × 4 combinations) were trained.
- **Procedure 3:** Ordinal forest method described in [33] with `ordfor` function from the `ordinalForest` R package. This method is tailored for predicting ordinal outcomes. We used default values for the **hyperparameters** `nsets`, `ntreeperdiv`, `ntreefinal`, and `npermtrial`, except for `nbest`, which was tuned carefully and determines the number of top-performing score sets used in optimizing the final score set. Improper specification of this parameter can degrade performance: overly large values introduce suboptimal score sets with excessive deviation from the optimal score set. Overly small values increase variance in the optimized score set.
  Following [33], we set `nsets=1000`, `ntreeperdiv=100`, `npermtrial=100`, `ntreefinal=1000`, and explored 5 values of `nbest`, ranging from 8 to 12, including the default value of 10. This approach balanced score set heterogeneity and estimation variance. As the goal was to obtain class probability predictions, we used `perffunction="probability"` to optimize class probability predictions using the ranked probability score.
- **Procedure 4:** Finally, we evaluate the impact of the Ord-MAP criterion versus the traditional MAP criterion when tuning random forest models, using both Accuracy and MAE as optimization metrics. Models were trained using the `train` function from the `caret` R package.[5] Key aspects of the experimental setup:
  - **Random forest setup**: A small ensemble of three trees was used.
  - **Hyper-parameter tuning**: Conducted via 3-fold cross-validation with a random search of 10 iterations.
  - **Custom MAE metric**: While `caret` defaults to Accuracy as the performance metric, using the `summaryFunction` argument in `trainControl` we introduced MAE as a custom metric, setting it as the minimization objective.

  This design enabled a robust evaluation of the Ord-MAP criterion's impact across two metrics: Accuracy and MAE, used in guiding the grid search process for hyper-parameter tuning to improve classifier performance.

5.3. Evaluation methodology

The performance of the criteria was validated using 10-fold cross-validation. The dataset were randomly partitioned into 10 folds, with each fold serving once as a test set while the remaining folds formed the training set. For each test set, predicted probability distributions from each trained model were obtained using the generic `predict` function in R. The Mean Absolute Error (MAE) was then calculated for predictions generated by both the MAP and Ord-MAP criteria.

To statistically compare the MAE values of the two criteria, we conducted hypothesis testing. Regardless of the assumption of normality, paired Wilcoxon signed-rank tests were performed to evaluate differences between the criteria. The significance levels of the results

[3] https://doi.org/10.32614/CRAN.package.ordinalForest

[4] https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring

[5] Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. Journal of Statistical Software, Vol. 28(5), pp. 1–26. https://www.jstatsoft.org/index.php/jss/article/view/v028i05

are denoted as follows: $^{\bullet}$ for $p < 0.10$ (10%), * for $p < 0.05$ (5%), ** for $p < 0.01$ (1%), and *** for $p < 0.001$ (1‰).

By applying this approach, we ensured a robust assessment of the Ord-MAP criterion across multiple modeling frameworks, making it possible to compare its predictive performance with that of MAP in terms of the MAE metric.

5.4. Results

To assess the effectiveness of the Ord-MAP criterion, we compared the MAE values obtained with the MAP and Ord-MAP criteria for each procedure and dataset combination.

Model estimation failed to converge for some datasets, both in **Procedure 1** (using the `polr` function from the `MASS` package) and in **Procedure 2** (using the `clm` function from the `ordinal` package). These convergence issues are not uncommon in maximum likelihood estimation of ordinal regression models and typically arise when the optimization algorithm encounters non-finite values — such as NA, Inf, or NaN — in the evaluation of the log-likelihood function or its gradient. Such problems can be exacerbated when using asymmetric link functions (such as loglog and cloglog), which may behave unstably in the presence of sparse outcome categories or extreme predictor values. Although these issues can sometimes be resolved through regularization, model simplification, alternative link choices, or better initialization strategies, in this study we deliberately applied each method to the original datasets without any dataset-specific tuning or adaptation to favor the convergence or performance of any particular method, as our goal was to assess the comparative performance of the models across a wide variety of conditions without biasing the results in favor of any particular setting. Additionally, the probability distributions assigned by the **Procedure 3** for the target variables of the Parkinson and the CES11 datasets, are so heavily skewed toward the true class that both MAP and Ord-MAP yielded identical predictions, always correct, making it impossible to observe differences.

Tables B.4–B.8 show p-values testing whether the MAE metric using MAP is higher (worse) than using Ord-MAP, for the different datasets. Instances of convergence failure are marked in the results tables with the symbol ✗. Figs. B.8–B.14 illustrate the comparative behavior of the Ord-MAP criterion versus the standard MAP through boxplots of differences in MAE (MAP − Ord-MAP) for each procedure and dataset, corroborating the results reported in Tables B.4–B.8. A red dashed line indicating zero has been added to each plot to serve as a visual reference: boxplots clearly positioned above this line reflect outcomes in favor of the Ord-MAP criterion, whereas those below it indicate better performance of the standard MAP.

5.5. Analysis of the results

This final subsection discusses the main experimental findings, summarized in Tables 1–3, which highlight the superior performance of the proposed Ord-MAP criterion over MAP, with the MAE performance metric. To aid interpretation:

Table 1
Summary table of the results: Procedure 1. First row: number of significant results/in favor of Ord-MAP/in favor of MAP. Second row: exact p-value. Only significant p-values (< 0.10) are reported, and **all** of them indicate better performance with Ord-MAP.

Link function				
logistic	probit	loglog	cloglog	cauchit
9/9/0 0.001953**	8/8/0 0.003906**	5/5/0 0.031250*	6/6/0 0.015625*	7/7/0 0.007813**

Table 2
Summary table of the results: Procedure 2. At each cell, first row: number of significant results/in favor of Ord-MAP/in favor of MAP. Second row: exact p-value. Only significant p-values (< 0.10) are reported, and **all** of them indicate better performance with Ord-MAP.

		Link function				
		logistic	probit	loglog	cloglog	cauchit
Threshold structure	Flexible	8/8/0 0.003906**	3/3/0	2/2/0	3/3/0	7/7/0 0.007813**
	Symmetric	8/8/0 0.003906**	2/2/0	2/2/0	3/3/0	5/5/0 0.03125*
	symmetric2	7/7/0 0.007813**	2/2/0	2/2/0	1/1/0	8/8/0 0.003906**
	Equidistant	8/8/0 0.003906**	2/2/0	2/2/0	2/2/0	5/5/0 0.03125*

- Table 1 summarizes the results from a more detailed table (Table B.4), where for each link function, the number of statistically significant outcomes is reported (out of a maximum of 10, corresponding to the various datasets and target variables), along with how many of those favor the Ord-MAP criterion. All significant results support Ord-MAP. The associated **exact p-values**,[6]refer to the probabilities of obtaining the observed number of favorable outcomes under the null hypothesis (i.e., assuming no real difference between Ord-MAP and MAP, and each result being equally likely with probability $1/2$). For instance, for the logistic link function, we obtained 9 significant outcomes, all favoring Ord-MAP. The probability of this occurring by chance is $(1/2)^9 = 0.001953^{**}$, which is clearly statistically significant.
- Table 2 is a similar summary of another set of results (Tables B.5–B.6), structured by the link function (in columns) and threshold structure (in rows).
- Table 3 combines summaries from two procedures. The left part refers to Procedure 3 and reports, for each of the three datasets in Table B.7, the number of significant results (out of 5, corresponding to different values of the `nbest` hyperparameter), and how many favor Ord-MAP versus MAP. The right part corresponds to Procedure 4 and summarizes Table B.8: for each summary function used in tuning the random forest model, we record the number of significant outcomes (out of 10, corresponding to the datasets), along with the number of them that favor Ord-MAP and MAP, respectively.

These summaries collectively underscore the consistent advantage of Ord-MAP criterion over MAP across a variety of models and settings, which is also visually supported by Figs. B.8–B.14.

6. Some simulations

To complement both the theoretical results established in the previous sections and the experiments conducted with real-world datasets, we designed a controlled simulation study aimed at analyzing the behavior of the MAP and Ord-MAP decision rules under varying levels of uncertainty. While experimental results already suggested a consistent advantage of Ord-MAP in ordinal settings, these simulations allow us to systematically vary the predictive distributions and the true class position, in order to confirm and better understand the conditions under which each decision rule performs best.

To better understand the empirical performance of the MAP (mode) and Ord-MAP (median) decision criteria under varying uncertainty conditions, we conducted a simulation study using synthetic ordinal

[6] The p-value is considered “exact” because it is computed directly from combinatorial probabilities under the null hypothesis of no difference, without relying on asymptotic approximations. While related to *permutation tests* or *randomization tests*, which evaluate all possible label assignments under the assumption of no effect, our approach does not involve resampling. Instead, it relies on the binomial probability of obtaining the observed number of favorable comparisons by chance, ensuring a precise measure of statistical significance.

Table 3
Summary table of the results: Procedures 3 and 4. First row: number of significant results/in favor of Ord-MAP/in favor of MAP. Second row: exact p-value. Only significant p-values (< 0.10) are reported, and **all** of them indicate better performance with Ord-MAP.

Procedure 3			Procedure 4	
Dataset			Summary function	
(a) WVS	(b) Wine	(c) Hearth	Accuracy	MAE
5/5/0 0.031250*	2/0/2	5/5/0 0.031250*	10/10/0 0.000977***	10/10/0 0.000977***

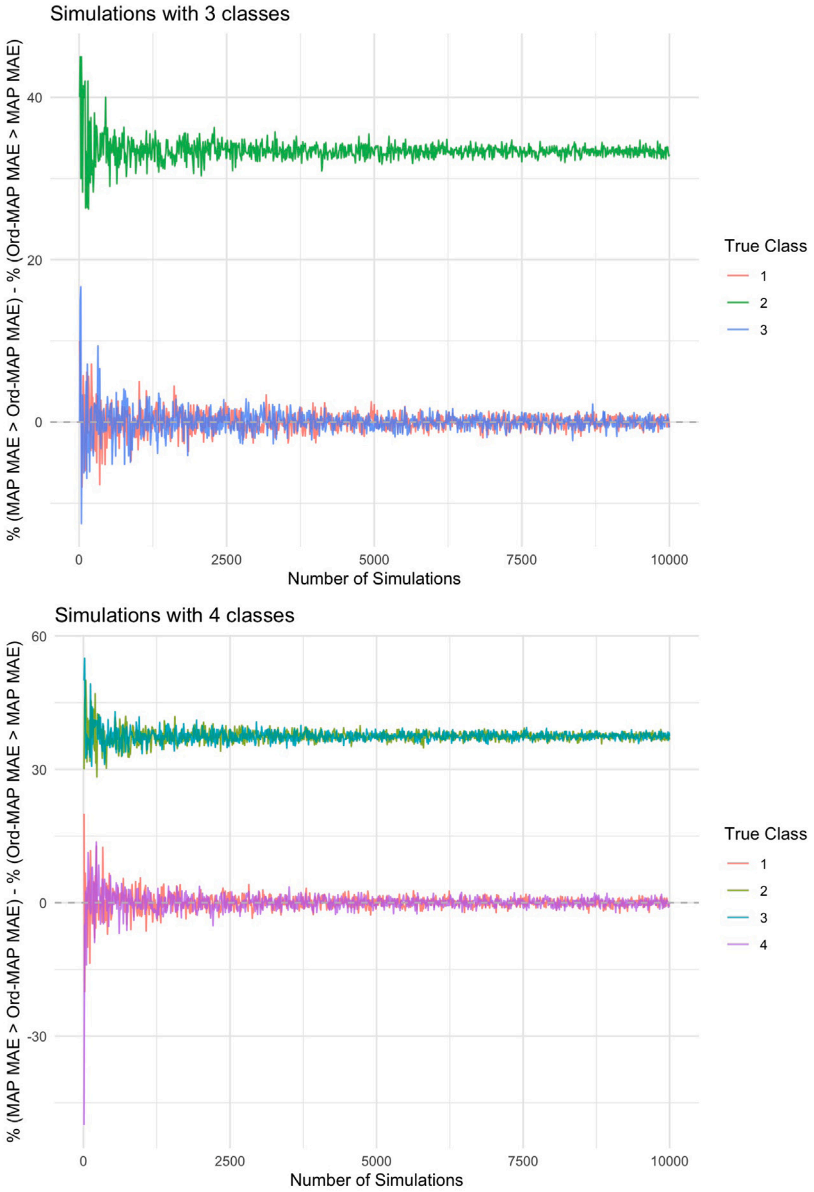


Fig. 3. Simulations with 3 and with 4 ordinal classes. For each true class value and number of simulations, we display the difference between the percentage of times the Ord-MAP decision rule outperforms MAP (in terms of MAE) and the percentage of times the opposite occurs. Positive values indicate better performance of Ord-MAP.

probability distributions with three, four, five and six ordered classes. For each fixed number of classes (from 3 to 6), we generated a set of random discrete probability distributions over the classes. To simulate realistic probabilistic scenarios without ties, each probability distribution was generated by drawing a number equal to the number of classes of independent values from a uniform distribution, normalizing them to sum to 1. To ensure that no two classes receive exactly the same probability — thus avoiding ambiguities in the mode (MAP criterion) — a small random jitter was added to each probability before renormalization. This approach guarantees strictly ordered probability vectors while preserving the variability and stochastic nature of the simulation setup.

Then, we computed the corresponding predictions using both the MAP and Ord-MAP criteria. For a range of simulation sizes from 10 to 10,000, increasing in steps of 10, to assess how robust the observed patterns were as the sample size increased, and for each possible fixed true class, we computed the percentage of cases in which the MAE of

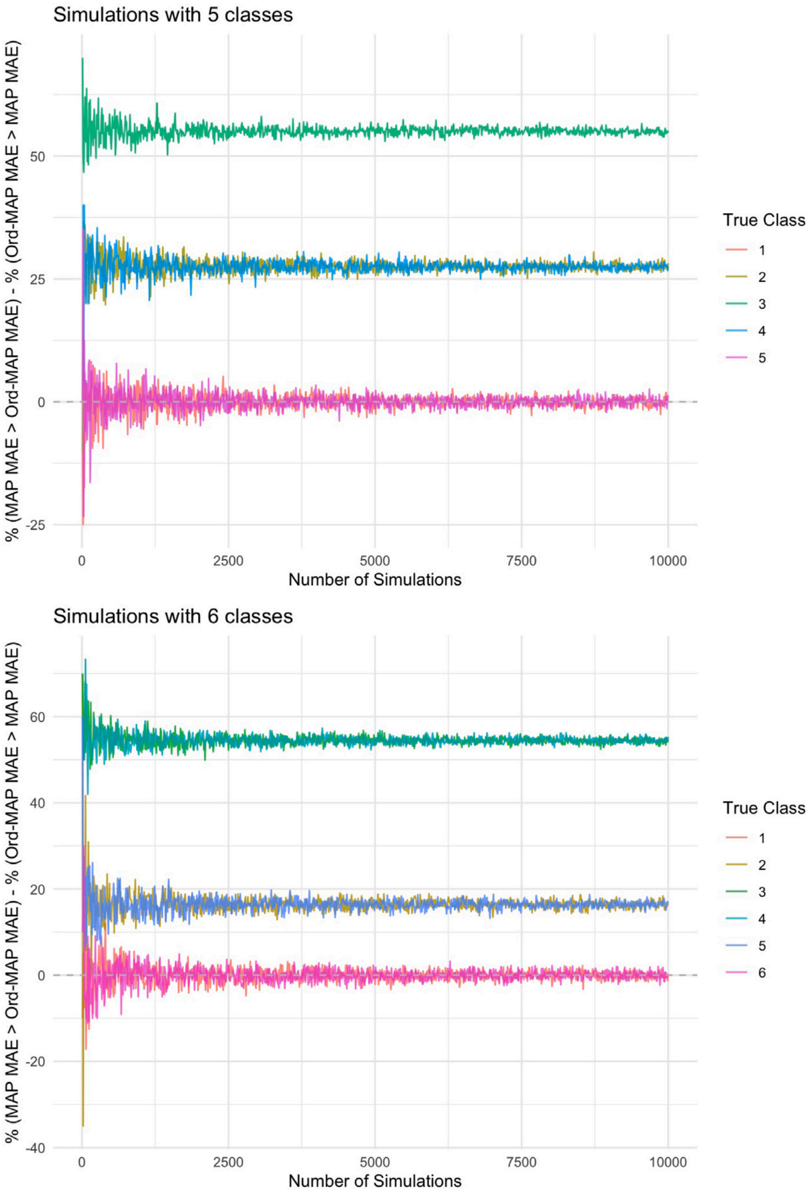


Fig. 4. Simulations with 5 and with 6 classes. As in Fig. 3, we show the advantage of Ord-MAP over MAP across a range of simulation sizes and fixed true class values.

Ord-MAP was lower (then, better) than that of MAP, and vice versa. This allowed us to assess the relative performance of the two criteria as a function of sample size and true class position, as shown in Figs. 3–4 below.

To complement the analysis of the empirical performance of MAP and Ord-MAP, we further explored how the MAE difference between these two criteria varies with the entropy of the simulated probability distribution. Specifically, we fixed each possible true class in turn and simulated 10,000 random probability distributions as described above. For each probability distribution that serves as the input for both the MAP and Ord-MAP decision criteria, we computed the associated Shannon entropy, which reflects its level of uncertainty, and also computed the MAE difference between MAP and Ord-MAP predictions. Grouping the results into entropy bins, we visualized the average MAE difference as a function of entropy, separately for each fixed true class, in Figs. 5–6, where **error bars** are used to represent the uncertainty or variability of the mean differences in MAE (MAP − Ord-MAP) for each entropy bin. The length of the error bars corresponds to the standard error of the mean (SEM), showing the precision of the mean estimate. The shorter the error bars, the more precise the estimate, suggesting that the improvement observed is reliable and can be attributed to the characteristics of the Ord-MAP criterion, while longer bars suggest greater variability in the results and that the results might be more sensitive to random fluctuations or variations in the data.

A clear pattern emerges from the simulations: the Ord-MAP yields lower mean absolute error (MAE) than the standard MAP rule, particularly when the true class lies in the interior of the ordinal scale (i.e., not in the extreme categories). This is aligned with the fact that Ord-MAP, which is sensitive to the order among classes, minimizes expected absolute error (λ-loss), as we established in Theorem 1.

We observe in Figs. 3–4 that Ord-MAP tends to produce predictions that are closer to the true class when it is centrally located, reducing the MAE: when the true class lies near the center of the ordinal scale, the median is more likely to yield predictions closer to the actual value, thereby reducing the mean absolute error. In contrast, when the true class is at one of the boundaries of the scale, both MAP and Ord-MAP often give the same prediction, and their performance gap narrows. Moreover, the advantage of Ord-MAP over MAP increases progressively as the true class moves toward the center of the ordinal scale, in a pattern that is symmetric with respect to both extremes — an expected

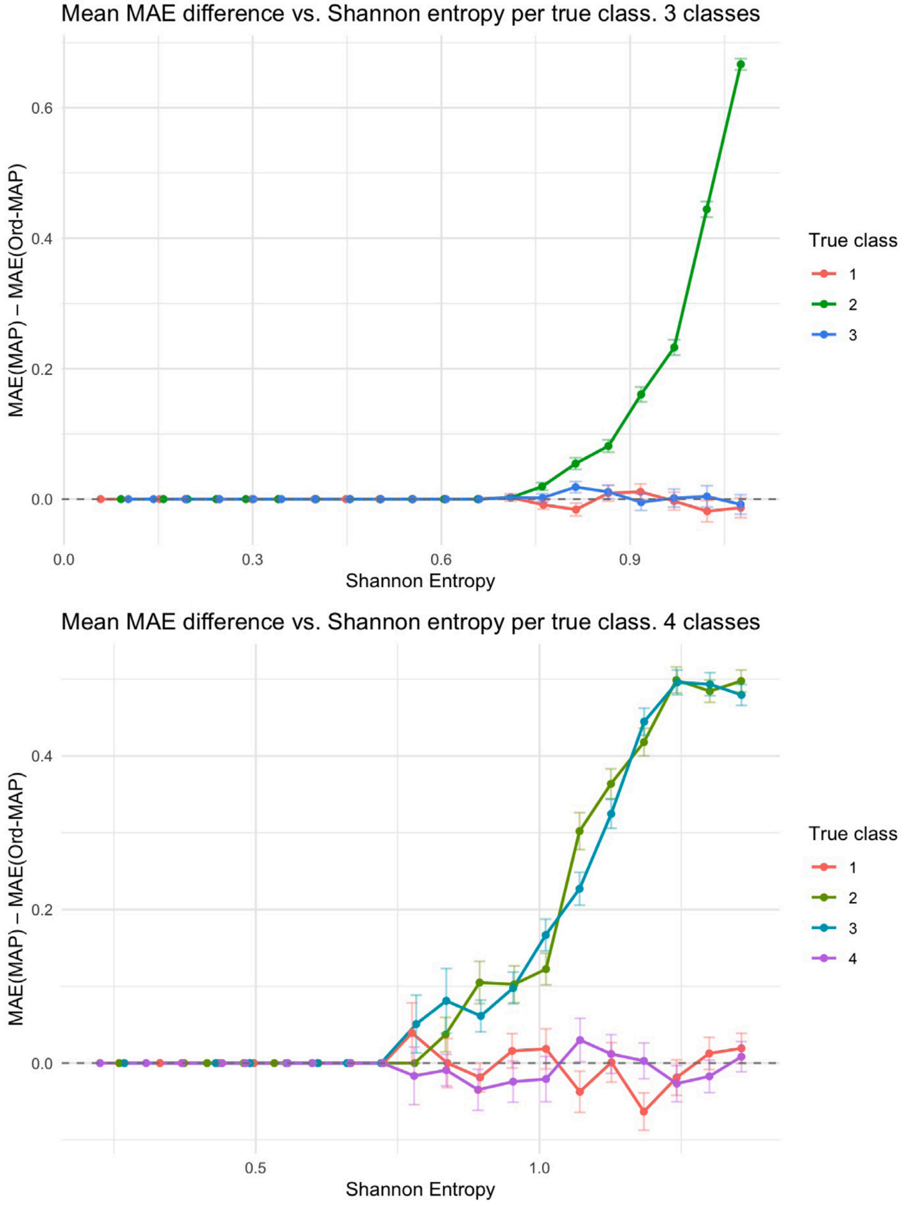


Fig. 5. Average MAE difference between MAP and Ord-MAP as a function of Shannon entropy of the simulated probability distribution with 3 and 4 ordinal classes, for each fixed true class.

behavior that is nevertheless confirmed empirically by the simulations — indicating that the median-based criterion becomes increasingly effective in minimizing prediction error for centrally located categories. These findings support the suitability of Ord-MAP as a principled alternative to MAP in ordinal classification.

On the other hand, as Shannon entropy increased, we observed in Figs. 5–6 that the MAE difference between MAP and Ord-MAP also grew, with Ord-MAP criterion demonstrating progressively superior performance. This trend was especially noticeable when the true class moved toward the center of the ordinal scale, exhibiting symmetric behavior at both extremes. Furthermore, the improvements were more pronounced under conditions of higher entropy. Thus, the relationship between entropy and MAE difference reinforces the robustness of Ord-MAP in the face of increasing uncertainty, providing strong evidence that this median-based criterion is not only theoretically advantageous but also practically effective in exploiting the ordinal structure of the problem, particularly in scenarios with moderate to high uncertainty.

7. Conclusions

This study has introduced the Ord-MAP criterion as a theoretically grounded and practically effective approach to ordinal classification. By explicitly incorporating the ordered nature of the target variable into the decision-making process, Ord-MAP bridges the gap between probabilistic predictions and ordinal relationships. Fig. 7 presents a schematic comparison of the Ord-MAP criterion for ordinal classification and the MAP criterion. This parallel layout facilitates the observation of similarities and differences between the two criteria, attending to different components: class assignment, loss function, associated performance metric, expected loss, score and optimality.

The rigorous theoretical framework underpinning Ord-MAP ensures its robustness, while the experimental results, obtained using diverse datasets and models — including classifiers specifically designed for ordinal data — clearly demonstrate that the Ord-MAP criterion outperforms the traditional MAP criterion in terms of mean absolute error (MAE). Notably, Ord-MAP outperformed MAP not only in generic ordinal classification tasks but also in scenarios with advanced ordinal

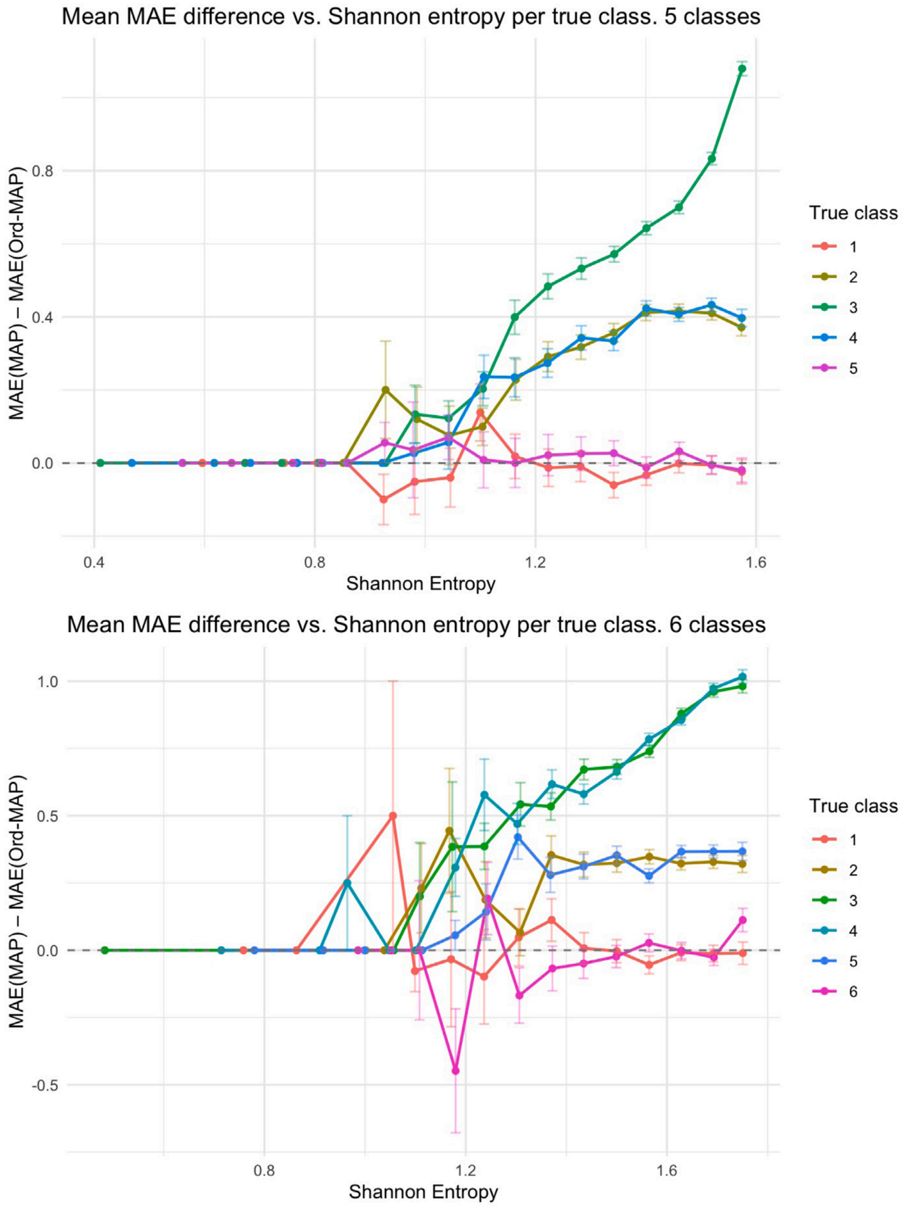
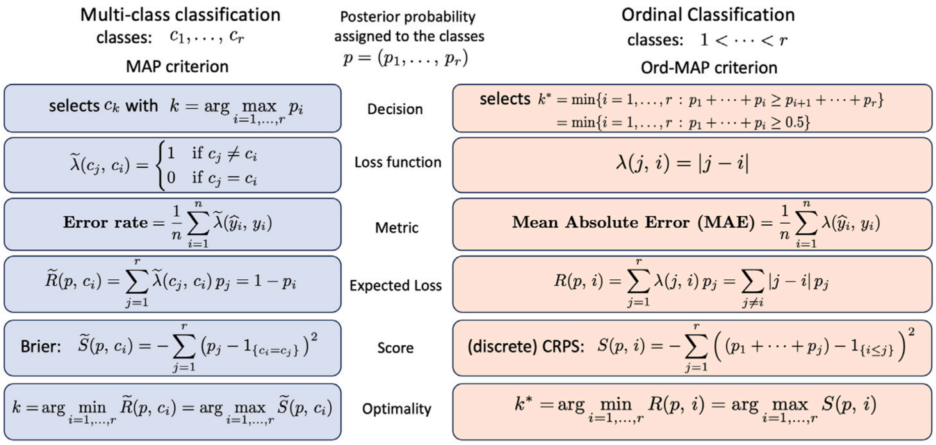


Fig. 6. Average MAE difference between MAP and Ord-MAP as a function of Shannon entropy of the simulated probability distribution with 5 and 6 ordinal classes, for each fixed true class.



Multi-class classification classes: $c_1, \ldots, c_r$ MAP criterion	Posterior probability assigned to the classes $p = (p_1, \ldots, p_r)$	Ordinal Classification classes: $1 < \cdots < r$ Ord-MAP criterion
selects c_k with $k = \arg\max_{i=1,\ldots,r} p_i$	Decision	selects $k^* = \min\{i = 1, \ldots, r : p_1 + \cdots + p_i \geq p_{i+1} + \cdots + p_r\}$ $= \min\{i = 1, \ldots, r : p_1 + \cdots + p_i \geq 0.5\}$
$\widetilde{\lambda}(c_j, c_i) = \begin{cases} 1 & \text{if } c_j \neq c_i \\ 0 & \text{if } c_j = c_i \end{cases}$	Loss function	$\lambda(j, i) = \lvert j - i \rvert$
$\textbf{Error rate} = \frac{1}{n} \sum_{i=1}^{n} \widetilde{\lambda}(\widehat{y}_i, y_i)$	Metric	$\textbf{Mean Absolute Error (MAE)} = \frac{1}{n} \sum_{i=1}^{n} \lambda(\widehat{y}_i, y_i)$
$\widetilde{R}(p, c_i) = \sum_{j=1}^{r} \widetilde{\lambda}(c_j, c_i)\, p_j = 1 - p_i$	Expected Loss	$R(p, i) = \sum_{j=1}^{r} \lambda(j, i)\, p_j = \sum_{j \neq i} \lvert j - i \rvert\, p_j$
Brier: $\widetilde{S}(p, c_i) = -\sum_{j=1}^{r} \left(p_j - 1_{\{c_i = c_j\}}\right)^2$	Score	(discrete) CRPS: $S(p, i) = -\sum_{j=1}^{r} \Big((p_1 + \cdots + p_j) - 1_{\{i \leq j\}} \Big)^2$
$k = \arg\min_{i=1,\ldots,r} \widetilde{R}(p, c_i) = \arg\max_{i=1,\ldots,r} \widetilde{S}(p, c_i)$	Optimality	$k^* = \arg\min_{i=1,\ldots,r} R(p, i) = \arg\max_{i=1,\ldots,r} S(p, i)$

Fig. 7. Comparison between the MAP criterion for nominal classification and the Ord-MAP criterion for ordinal classification.

classifiers. For instance, significant reductions in MAE were observed in models employing cumulative link structures, ordinal forests, and random forests tuned with accuracy or MAE. In addition to these empirical results, controlled simulations further corroborate that Ord-MAP consistently improves prediction across the ordinal scale when using MAE as evaluation metric — especially for centrally located classes, with symmetric improvements observed as the true class moves away from the extremes, and for probability distributions with moderate to high entropy. These findings underscore the suitability of Ord-MAP for capturing the intrinsic order of the target variable, marking it as a valuable tool for both theoretical exploration and practical application in ordinal classification, especially in scenarios with moderate to high uncertainty.

The most significant theoretical contribution of this work lies in establishing that the MAP criterion, which selects the class with the highest assigned probability (the **mode**), is appropriate for nominal classification but suboptimal for ordinal data. For ordinal classification, we demonstrate that the proper extension of the MAP criterion is to select the first class such that the cumulative probability, considering the order of the classes, exceeds 0.5, that is, the **median**. This insight fundamentally reframes how probabilistic methods should approach ordinal classification, bridging the gap between theory and application. Although some previous works have already used the notion of a median, in some sense, in the setting of ordinal classification (see [34] and references therein), they do not explicitly formulate a criterion like Ord-MAP. In contrast, the present work introduces a median-based decision rule that is both classifier-agnostic and grounded in a formal optimality result. As such, it provides a simple, clear, and elegant solution to an open question in the field.

Future research could explore the integration of the Ord-MAP criterion into machine learning models for more complex settings, such as cost-sensitive, interval scale or imbalanced ordinal datasets. Additionally, studying its computational efficiency in large-scale problems would further clarify its practical utility. These efforts would help refine the Ord-MAP criterion and explore its potential to redefine state-of-the-art practices in ordinal classification, benefiting a wide range of real-world applications.

Funding

The author is supported by Ministerio de Ciencia e Innovación, Spain, Gobierno de España, project ref. PID2021-123733NB-I00.

Code availability

The R scripts used to implement the experimental phase detailed in Section 5 and simulations in Section 6 can be found at https://github.com/RosDelgado/Ord_MAP

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The author would like to thank the editors and anonymous reviewers for their insightful comments and constructive suggestions, which have led to a clearer and more refined presentation of the manuscript, as well as a stronger and more thorough experimental section. The simulation study presented in Section 6 was carried out in response to a reviewer's suggestion, and it has proven to be a valuable addition since this controlled analysis of synthetic probability distributions has contributed to a deeper understanding of the empirical behavior of the proposed criterion and has further reinforced the practical relevance of the theoretical findings.

Appendix A. The properness of the brier score $\widetilde{S}$ and the discretized continuous ranked probability score (CRPS) S

Theorem 2. *The Brier score defined by* (3) *and the discretized CRPS defined by* (5) *are* **proper** *relative to* $\mathcal{P}_r$*, where* $\mathcal{P}_r$ *is specified in* (4).

Proof. To prove this result, we use Theorem 1 [24] with adapted notations. This result guarantees that a score $\breve{S}$ is proper is and only if there exists a convex, real-valued function $\breve{G}$ on $\mathcal{P}_r$, such that for all $k \in \{1, \dots, r\}$ and $p \in \mathcal{P}_r$, the following holds:

$$\breve{S}(p,\, k) = \breve{G}(p) - < \breve{G}'(p),\, p > + \breve{G}'_k(p), \tag{A.1}$$

where $\breve{G}'(p)$ denotes the subgradient of $\breve{G}$ at point p. We will apply this result with $\breve{S}$ set to $\widetilde{S}$ (the Brier score) and to S (the discretized CRPS). In each case, we will determine the corresponding function $\breve{G}$, which we will denote as $\widetilde{G}$ for $\widetilde{S}$ and G for S.

Note that by (A.1),

$$\begin{aligned}
\breve{S}(p,\, p) &= \sum_{k=1}^{r} \breve{S}(p,\, k)\, p_k = \sum_{k=1}^{r} \big(\breve{G}(p) - < \breve{G}'(p),\, p > + \breve{G}'_k(p) \big)\, p_k \\
&= \big(\breve{G}(p) - < \breve{G}'(p),\, p > \big) \sum_{k=1}^{r} p_k + \sum_{k=1}^{r} \breve{G}'_k(p)\, p_k \\
&= \big(\breve{G}(p) - < \breve{G}'(p),\, p > \big) + < \breve{G}'(p),\, p >= \breve{G}(p).
\end{aligned}$$

Therefore, our candidate is function $\breve{G} : \mathcal{P}_r \longrightarrow \mathbb{R}$ defined by

$$\breve{G}(p) = \breve{S}(p,\, p) = \sum_{k=1}^{r} \breve{S}(p,\, k)\, p_k. \quad \text{Then,}$$

$$\widetilde{G}(p) = \sum_{k=1}^{r} \widetilde{S}(p,\, k)\, p_k = -\sum_{k=1}^{r} \left(\sum_{j=1}^{r} \Big(p_j - 1_{\{c_k = c_j\}} \Big)^2 \right) p_k, \tag{A.2}$$

$$G(p) = \sum_{k=1}^{r} S(p,\, k)\, p_k = -\sum_{k=1}^{r} \left(\sum_{j=1}^{r} \Big((p_1 + \cdots + p_j) - 1_{\{k \le j\}} \Big)^2 \right) p_k. \tag{A.3}$$

The rest of the proof is divided into three steps: obtaining a more practical expression for $\breve{G}$, verifying that (A.1) holds, and showing that $\breve{G}$ is convex. Then, a direct application of Theorem 1 [24] completes the proof.

Step 1: We now prove that

$$\widetilde{G}(p) = -1 + \sum_{j=1}^{r} p_j^2, \tag{A.4}$$

$$G(p) = -\sum_{k=1}^{r} (r - k + 1)\, p_k + \sum_{j=1}^{r} (p_1 + \cdots + p_j)^2. \tag{A.5}$$

Indeed, by (A.2),

$$\begin{aligned}
\widetilde{G}(p) &= -\sum_{k=1}^{r} \left(\sum_{j=1}^{r} \Big(p_j - 1_{\{c_k = c_j\}} \Big)^2 \right) p_k = -\sum_{k=1}^{r} \left(\sum_{j \ne k} p_j^2 + \big(p_k - 1 \big)^2 \right) p_k \\
&= -\sum_{k=1}^{r} \left(\sum_{j \ne k} p_j^2 + \big(p_k^2 - 2\, p_k + 1 \big) \right) p_k \\
&= -\sum_{k=1}^{r} \left(\sum_{j=1}^{r} p_j^2 + 1 - 2\, p_k \right) p_k \\
&= -\sum_{j=1}^{r} p_j^2 \sum_{k=1}^{r} p_k - \sum_{k=1}^{r} p_k + 2 \sum_{k=1}^{r} p_k^2 = -1 + \sum_{j=1}^{r} p_j^2,
\end{aligned}$$

and by (A.3),

$$\begin{aligned}
G(p) &= -\sum_{k=1}^{r} \left(\sum_{j=1}^{r} \Big((p_1 + \cdots + p_j) - 1_{\{k \le j\}} \Big)^2 \right) p_k \\
&= -\sum_{k=1}^{r} \left(\sum_{j=1}^{k-1} \Big((p_1 + \cdots + p_j) - 0 \Big)^2 + \sum_{j=k}^{r} \Big((p_1 + \cdots + p_j) - 1 \Big)^2 \right) p_k
\end{aligned}$$

$$= -\sum_{k=1}^{r}\Big(\sum_{j=1}^{k-1}(p_1+\cdots+p_j)^2+\sum_{j=k}^{r}\Big((p_1+\cdots+p_j)^2$$
$$+\,1-2(p_1+\cdots+p_j)\Big)\Big)\,p_k$$
$$= -\sum_{k=1}^{r}\left(\sum_{j=1}^{r}(p_1+\cdots+p_j)^2+(r-k+1)-2\sum_{j=k}^{r}(p_1+\cdots+p_j)\right)p_k$$
$$= -\sum_{j=1}^{r}(p_1+\cdots+p_j)^2-\sum_{k=1}^{r}(r-k+1)\,p_k+2\sum_{k=1}^{r}\Big(\sum_{j=k}^{r}(p_1+\cdots+p_j)\Big)\,p_k$$
$$= -\sum_{k=1}^{r}(r-k+1)\,p_k+\sum_{j=1}^{r}(p_1+\cdots+p_j)^2,$$

where in the last equality we have used that

$$\sum_{k=1}^{r}\Big(\sum_{j=k}^{r}(p_1+\cdots+p_j)\Big)\,p_k=\sum_{j=1}^{r}(p_1+\cdots+p_j)^2, \tag{A.6}$$

which is immediate to check since

$$\sum_{k=1}^{r}\Big(\sum_{j=k}^{r}(p_1+\cdots+p_j)\Big)\,p_k=\sum_{j=1}^{r}\Big(\sum_{k=1}^{j}(p_1+\cdots+p_j)\,p_k\Big)$$
$$=\sum_{j=1}^{r}\Big(\sum_{k=1}^{j}p_k\Big)\,(p_1+\cdots+p_j)$$
$$=\sum_{j=1}^{r}(p_1+\cdots+p_j)^2.$$

Note that, as usual, when a sum does not make sense, such as $\sum_{j=1}^{k-1}$ for $k=1$, it simply disappears from the expression.

Step 2: We now verify that (A.1) holds. For this, we use the expression (A.4) obtained in the previous step, and from it, we observe that for any k, $\widetilde{G}'_k(p)=2\,p_k$ and then,

$$<\widetilde{G}'(p),\,p>=\sum_{k=1}^{r}(2\,p_k)\,p_k=2\sum_{k=1}^{r}p_k^2 \quad \text{and by (A.4),}$$

$$\widetilde{G}(p)-<\widetilde{G}'(p),\,p>+\widetilde{G}'_k(p)=-1+\sum_{j=1}^{r}p_j^2-2\sum_{j=1}^{r}p_j^2+2\,p_k$$
$$=-1-\sum_{j=1}^{r}p_j^2+2\,p_k=-1-\sum_{j\neq k}^{r}p_j^2-p_k^2+2\,p_k=-\sum_{j\neq k}^{r}p_j^2-(p_k-1)^2$$
$$=-\sum_{j=1}^{r}\Big(p_j-1_{\{c_k=c_j\}}\Big)^2=\widetilde{S}(p,\,k).$$

Analogously, by (A.5), $G'_k(p)=-(r-k+1)+2\sum_{j=k}^{r}(p_1+\cdots+p_j)$ and then,

$$<G'(p),\,p>=-\sum_{k=1}^{r}\Big((r-k+1)-2\sum_{j=k}^{r}(p_1+\cdots+p_j)\Big)\,p_k$$
$$=-\sum_{k=1}^{r}(r-k+1)\,p_k+2\sum_{j=1}^{r}(p_1+\cdots+p_j)^2$$

using (A.6). Therefore, by (A.5),

$$G(p)-<G'(p),\,p>+G'_k(p)=-\sum_{k=1}^{r}(r-k+1)\,p_k+\sum_{j=1}^{r}(p_1+\cdots+p_j)^2$$
$$+\left(\sum_{k=1}^{r}(r-k+1)p_k-2\sum_{j=1}^{r}(p_1+\cdots+p_j)^2\right)-(r-k+1)$$
$$+\,2\sum_{j=k}^{r}(p_1+\cdots+p_j)$$
$$=-\sum_{j=1}^{r}(p_1+\cdots+p_j)^2-(r-k+1)+2\sum_{j=k}^{r}(p_1+\cdots+p_j)$$
$$=-\sum_{j=1}^{k-1}(p_1+\cdots+p_j)^2-\sum_{j=k}^{r}\Big((p_1+\cdots+p_j)^2-2\,(p_1+\cdots+p_j)+1\Big)$$
$$=-\sum_{j=1}^{k-1}(p_1+\cdots+p_j)^2-\sum_{j=k}^{r}\Big((p_1+\cdots+p_j)-1\Big)^2=S(p,\,k).$$

Step 3: Finally, we prove that function $\breve{G}$ is convex, for both $\breve{G}=\widetilde{G}$ and $\breve{G}=G$. Let $\alpha\in(0,\,1)$ and $p,\,q\in\mathcal{P}_r$, then, by (A.4),

$$\widetilde{G}((1-\alpha)\,p+\alpha\,q)=-1+\sum_{j=1}^{r}\Big((1-\alpha)\,p_j+\alpha\,q_j\Big)^2$$
$$=-1+(1-\alpha)^2\sum_{j=1}^{r}p_j^2+\alpha^2\sum_{j=1}^{r}q_j^2+2\,\alpha\,(1-\alpha)\sum_{j=1}^{r}p_j\,q_j.$$

On the other hand,

$$(1-\alpha)\,\widetilde{G}(p)+\alpha\,\widetilde{G}(q)=(1-\alpha)\Big(-1+\sum_{j=1}^{r}p_j^2\Big)+\alpha\Big(-1+\sum_{j=1}^{r}q_j^2\Big)$$
$$=(1-\alpha)\sum_{j=1}^{r}p_j^2+\alpha\sum_{j=1}^{r}q_j^2-1.$$

By comparing these expressions it is immediate to check that $\widetilde{G}((1-\alpha)\,p+\alpha\,q)\le(1-\alpha)\,\widetilde{G}(p)+\alpha\,\widetilde{G}(q)$ (that is, that $\widetilde{G}$ is a convex function). Indeed, if we denote $\sum_{j=1}^{r}p_j^2$ and $\sum_{j=1}^{r}q_j^2$ by A and B, respectively, we must prove that

$$(1-\alpha)^2\,A+\alpha^2\,B+2\,\alpha\,(1-\alpha)\sum_{j=1}^{r}p_j\,q_j\le(1-\alpha)\,A+\alpha\,B,$$

which is equivalent to $A+B-2\sum_{j=1}^{r}p_j\,q_j\ge 0$ and this holds trivially since

$$A+B-2\sum_{j=1}^{r}p_j\,q_j=\sum_{j=1}^{r}(p_j-q_j)^2.$$

Analogously, by (A.5),

$$G((1-\alpha)\,p+\alpha\,q)=-\sum_{k=1}^{r}(r-k+1)\,\big((1-\alpha)\,p_k+\alpha\,q_k\big)$$
$$+\sum_{j=1}^{r}\Big(\big((1-\alpha)\,p_1+\alpha\,q_1\big)+\cdots+\big((1-\alpha)\,p_j+\alpha\,q_j\big)\Big)^2$$
$$=-(1-\alpha)\sum_{k=1}^{r}(r-k+1)\,p_k-\alpha\sum_{k=1}^{r}(r-k+1)\,q_k$$
$$+\sum_{j=1}^{r}\Big((1-\alpha)\,(p_1+\cdots+p_j)+\alpha\,(q_1+\cdots+q_j)\Big)^2$$
$$=-\,(1-\alpha)\sum_{k=1}^{r}(r-k+1)p_k-\alpha\sum_{k=1}^{r}(r-k+1)q_k$$
$$+\,(1-\alpha)^2\sum_{j=1}^{r}(p_1+\cdots+p_j)^2$$
$$+\,\alpha^2\sum_{j=1}^{r}(q_1+\cdots+q_j)^2+2\,\alpha\,(1-\alpha)\sum_{j=1}^{r}(p_1+\cdots+p_j)\,(q_1+\cdots+q_j).$$

On the other hand,

$$(1-\alpha)G(p)+\alpha G(q)=-\,(1-\alpha)\sum_{k=1}^{r}(r-k+1)p_k$$
$$+\,(1-\alpha)\sum_{j=1}^{r}(p_1+\cdots+p_j)^2$$
$$-\,\alpha\sum_{k=1}^{r}(r-k+1)\,q_k+\alpha\sum_{j=1}^{r}(q_1+\cdots+q_j)^2.$$

The convexity of G is proven by verifying that $G((1-\alpha)\,p+\alpha\,q)\le(1-\alpha)\,G(p)+\alpha\,G(q)$. This is true if $\sum_{j=1}^{r}a_j^2+\sum_{j=1}^{r}b_j^2-2\sum_{j=1}^{r}a_j\,b_j\ge 0$, where

Table B.4
Procedure 1: p-values comparing MAE for MAP and Ord-MAP criteria across datasets. ✗ denotes that model fitting failed to converge. Only significant p-values (< 0.10) are reported, and **all** of them indicate better performance with Ord-MAP. In parentheses, the number of classes.

		Link function				
Dataset		logistic	probit	loglog	cloglog	cauchit
(a) WVS	(3 c)	0.024414*	0.032227*		0.052734$^{\bullet}$	
(b) Wine	(5 c)		0.050174$^{\bullet}$	0.006426**	0.029529*	
(c) Hearth	(5 c)	0.004883**	0.010353*	0.061767$^{\bullet}$	0.004545**	
(d) Parkin.	V5 (6 c)	0.000977***	✗	✗	✗	0.000977***
	V5 (5 c)	0.000977***	0.000977***	✗	✗	0.000977***
	V5 (4 c)	0.032227*	✗	✗	✗	0.004883**
	V6 (6 c)	0.000977***	0.024414*	0.032227*	✗	0.006836**
	V6 (5 c)	0.000977***	0.000977***	✗	0.000977***	0.000977***
	V6 (4 c)	0.000977***	0.000977***	0.000977***	0.000977***	0.002930**
(e) CES11	(4 c)	0.000977***	0.000977***	0.001953**	0.000977***	0.000977***

Table B.5
Procedure 2 (I): p-values comparing MAE for MAP and Ord-MAP criteria across datasets. ✗ denotes that model fitting failed to converge. Only significant p-values (< 0.10) are reported, and **all** of them indicate better performance with Ord-MAP. In parentheses, the number of classes.

		Link function				
(a) WVS (3c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.024414*	0.032227*		0.052734$^{\bullet}$	
Thresh.	Symmetric	0.024414*	0.032227*		0.052734$^{\bullet}$	
struct.	Symmetric2		0.096680$^{\bullet}$	0.002930**		0.006836**
	Equidistant	0.024414*	0.032227*		0.052734$^{\bullet}$	
(b) Wine (5c)		logistic	probit	loglog	cloglog	cauchit
	Flexible		0.050174$^{\bullet}$	0.006426**	0.029529*	
Thresh.	Symmetric			0.002945**	0.090725$^{\bullet}$	
struct.	Symmetric2					0.050174$^{\bullet}$
	Equidistant			0.022005*		
(c) Hearth (5c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.004883**	0.010353*	0.061767$^{\bullet}$	0.004545**	0.091447$^{\bullet}$
Thresh.	Symmetric	0.017305*	0.009491**	0.009766**	0.002945**	
struct.	Symmetric2	0.004576**	0.009796**	0.018555*	0.001953**	0.006836**
	Equidistant	0.007133**	0.021913*	0.005388**	0.002930**	
(d) Parkinson V5 (6c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.000977***	✗	✗	✗	0.000977***
Thresh.	Symmetric	0.001953**	✗	✗	✗	0.001953**
struct.	Symmetric2	0.001953**	✗	✗	✗	0.009766**
	Equidistant	0.000977***	✗	✗	✗	0.000977***
(d) Parkinson V5 (5c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.000977***	✗	✗	✗	0.029012*
Thresh.	Symmetric	0.000977***	✗	✗	✗	0.032227*
struct.	Symmetric2	0.000977***	✗	✗	✗	0.022005*
	Equidistant	0.000977***	✗	✗	✗	0.000977***

$a_j = p_1 + \cdots + p_j$ and $b_j = q_1 + \cdots + q_j$, which is trivially true since for any $j = 1, \ldots, r$, it holds that $a_j^2 + b_j^2 - 2\, a_j\, b_j = (a_j - b_j)^2 \geq 0$. □

Appendix B. Tables and figures for the experimental phase (Section 5)

See Tables B.4–B.8 and Figs. B.8–B.13.

Table B.6
Procedure 2 (II): p-values comparing MAE for MAP and Ord-MAP criteria across datasets. ✗ denotes that model fitting failed to converge. Only significant p-values (< 0.10) are reported, and **all** of them indicate better performance with Ord-MAP. In parentheses, the number of classes.

		Link function				
(d) Parkinson V5 (4c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.061767$^{\bullet}$	✗	✗	✗	0.000977***
Thresh.	Symmetric	0.006836**	✗	✗	✗	0.000977***
struct.	Symmetric2	0.009766**	✗	✗	✗	0.000977***
	Equidistant	0.006836**	✗	✗	✗	0.000977***
(d) Parkinson V6 (6c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.000977***	✗	✗	✗	0.000977***
Thresh.	Symmetric	0.000977***	✗	✗	✗	0.000977***
struct.	Symmetric2	0.000977***	✗	✗	✗	0.000977***
	Equidistant	0.000977***	✗	✗	✗	0.000977***
(d) Parkinson V6 (5c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.000977***	✗	✗	✗	0.000977***
Thresh.	Symmetric	0.000977***	✗	✗	✗	0.080078$^{\bullet}$
struct.	Symmetric2	0.000977***	✗	✗	✗	0.080078$^{\bullet}$
	Equidistant	0.000977***	✗	✗	0.000977***	0.000977***
(d) Parkinson V6 (4c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	0.000977***	✗	✗	✗	0.000977***
Thresh.	Symmetric	0.000977***	✗	✗	✗	
struct.	Symmetric2	0.001953**	✗	✗	✗	
	Equidistant	0.009766**	✗	✗	0.001953**	
(e) CES11 (4c)		logistic	probit	loglog	cloglog	cauchit
	Flexible	✗	✗	✗	✗	✗
Thresh.	Symmetric	✗	✗	✗	✗	✗
struct.	Symmetric2	✗	✗	✗	✗	✗
	Equidistant	✗	✗	✗	✗	✗

Table B.7
Procedure 3: p-values comparing MAE for MAP and Ord-MAP criteria across datasets for the different values of the hyperparameter `nbest`. The default value, which is 10, and its corresponding p-value, are highlighted in boldface. Only significant p-values (< 0.10) are reported. In red, p-values in favor of MAP criterion. The rest are in favor of Ord-MAP. Results for Parkinson and CES11 datasets are not included since the probability distributions assigned by the classifier are so heavily skewed toward the true class that both MAP and Ord-MAP yielded identical predictions, always correct, making it impossible to observe differences. In parentheses, the number of classes.

Dataset		`nbest` hyperparameter				
(a) WVS	(3 c)	8	9	**10**	11	12
		0.000977***	0.000977***	**0.001953****	0.000977***	0.000977***
(b) Wine	(5 c)	8	9	**10**	11	12
			0.090725$^{\bullet}$			0.050174$^{\bullet}$
(c) Hearth	(5 c)	8	9	**10**	11	12
		0.010431*	0.0654297$^{\bullet}$	**0.005362****	0.009766**	0.009766**

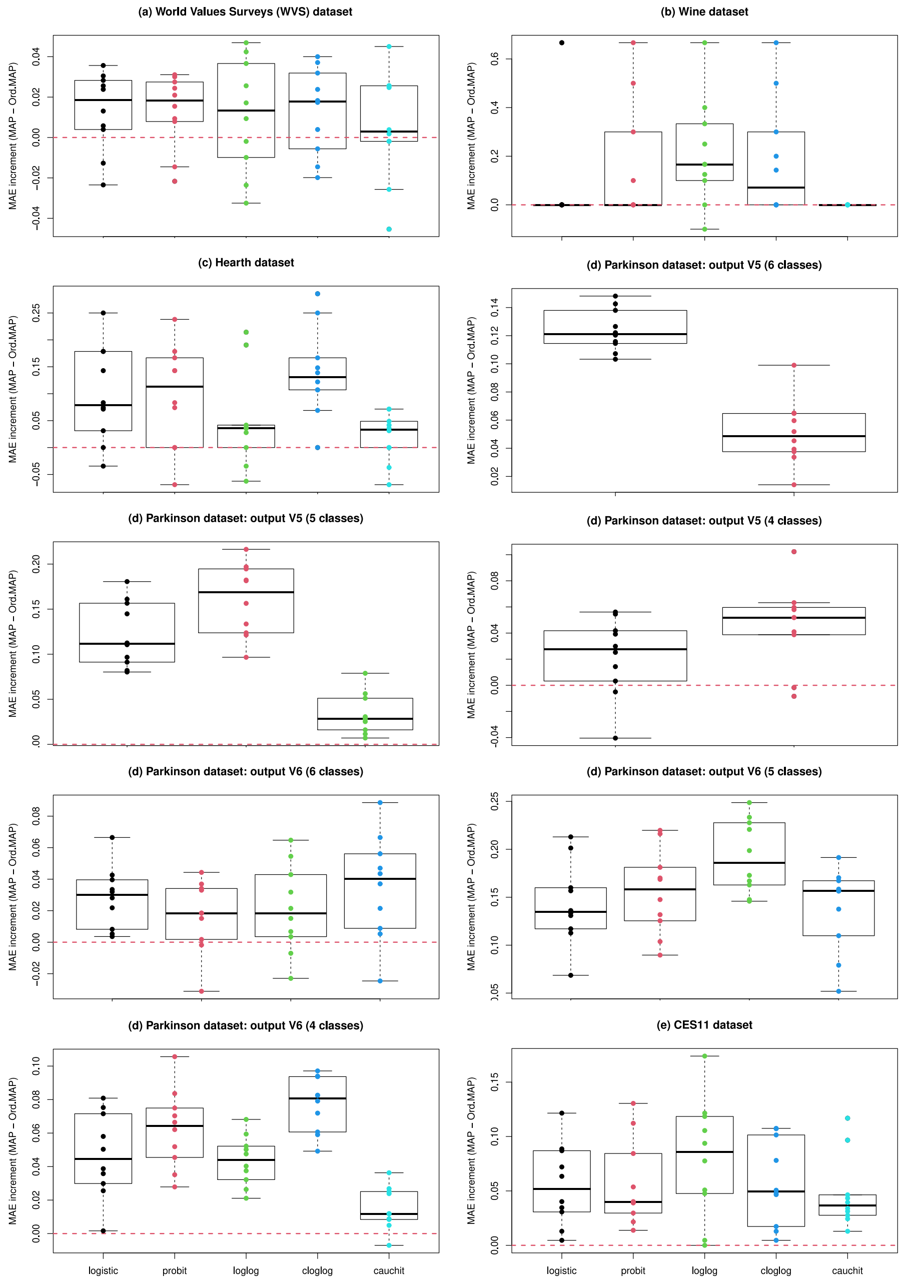
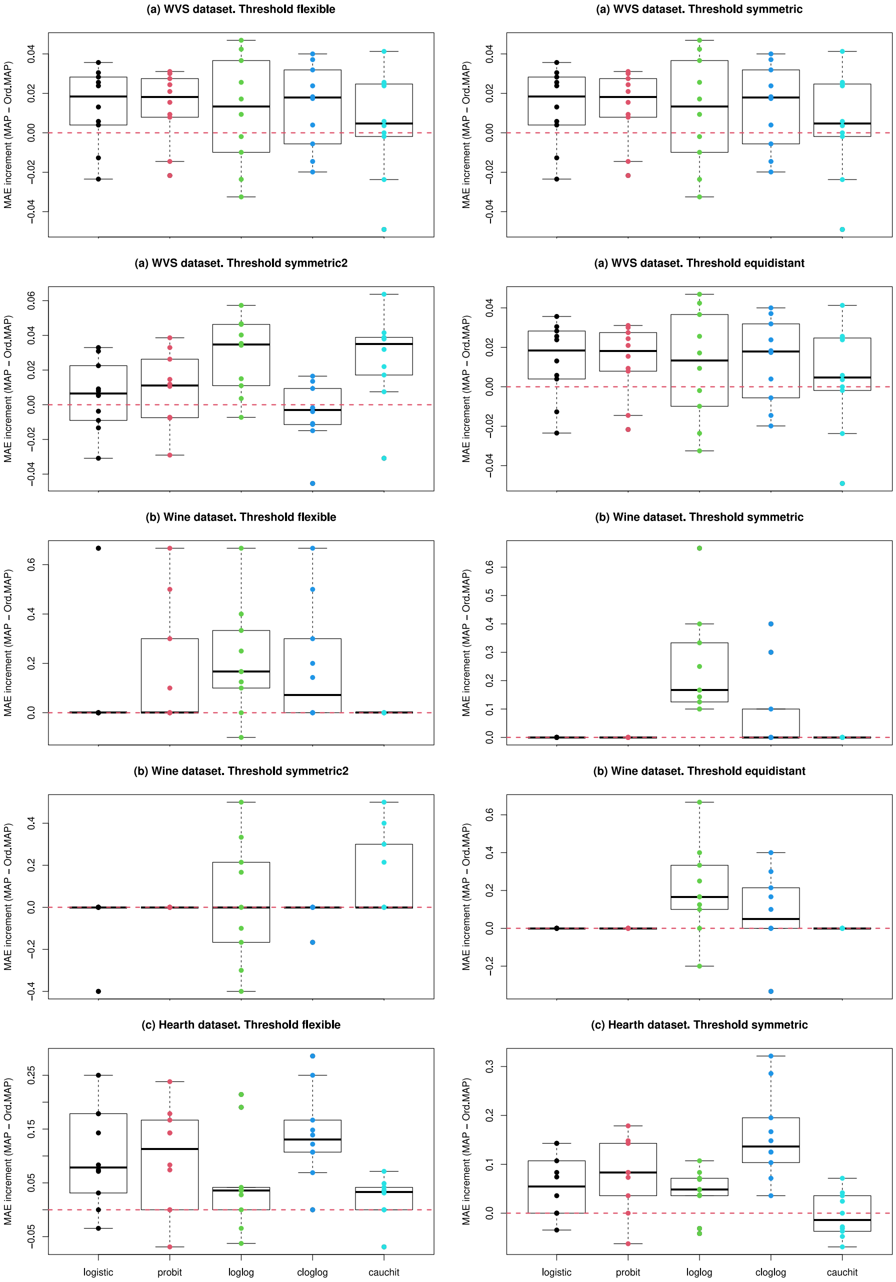


Fig. B.8. Procedure 1: boxplots of MAE increments between MAP and Ord-MAP depending on the link function, across datasets.



(a) WVS dataset. Threshold flexible

(a) WVS dataset. Threshold symmetric

(a) WVS dataset. Threshold symmetric2

(a) WVS dataset. Threshold equidistant

(b) Wine dataset. Threshold flexible

(b) Wine dataset. Threshold symmetric

(b) Wine dataset. Threshold symmetric2

(b) Wine dataset. Threshold equidistant

(c) Hearth dataset. Threshold flexible

(c) Hearth dataset. Threshold symmetric

Fig. B.9. Procedure 2 (I): boxplots of MAE increments between MAP and Ord-MAP depending on the link function and the threshold structure, across datasets.

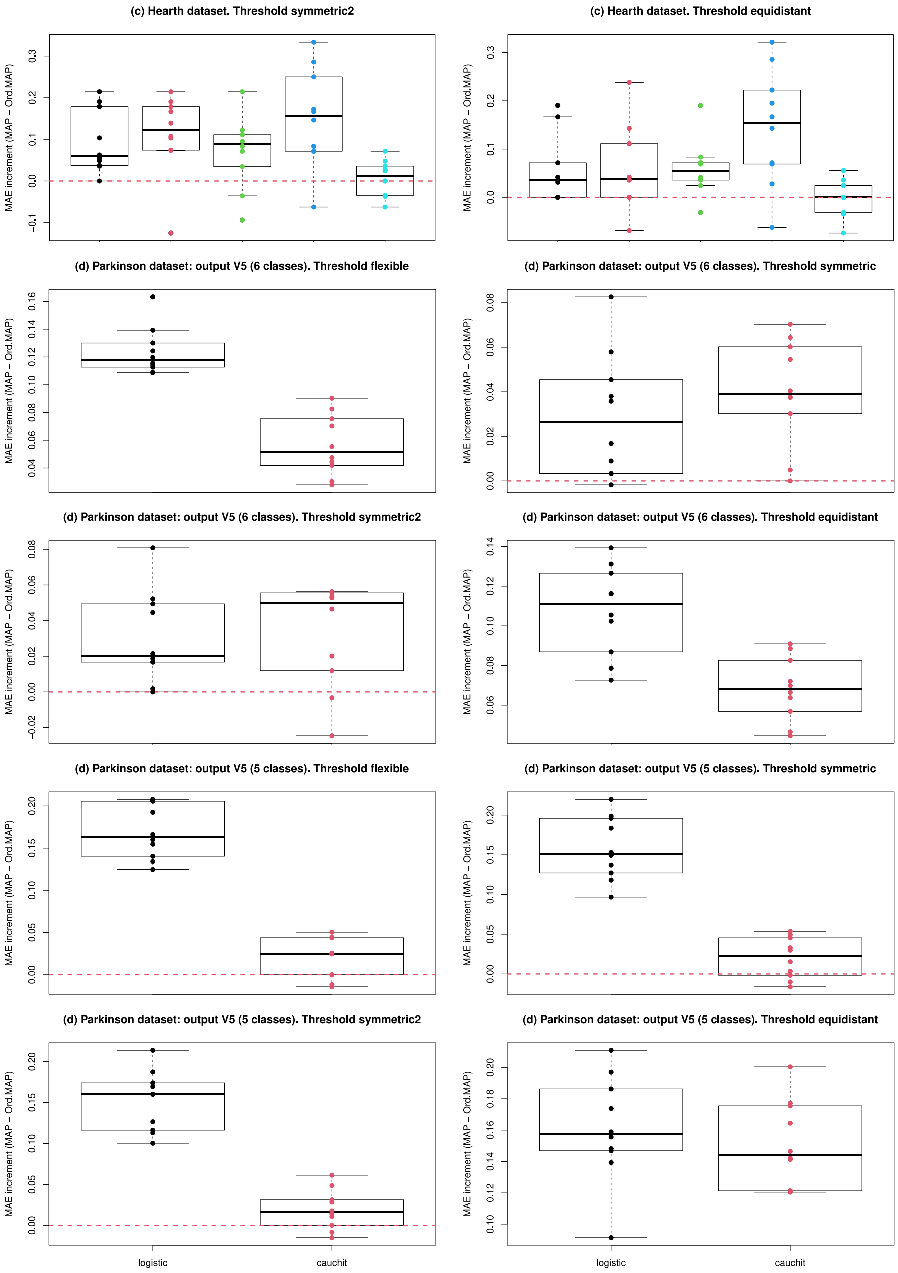
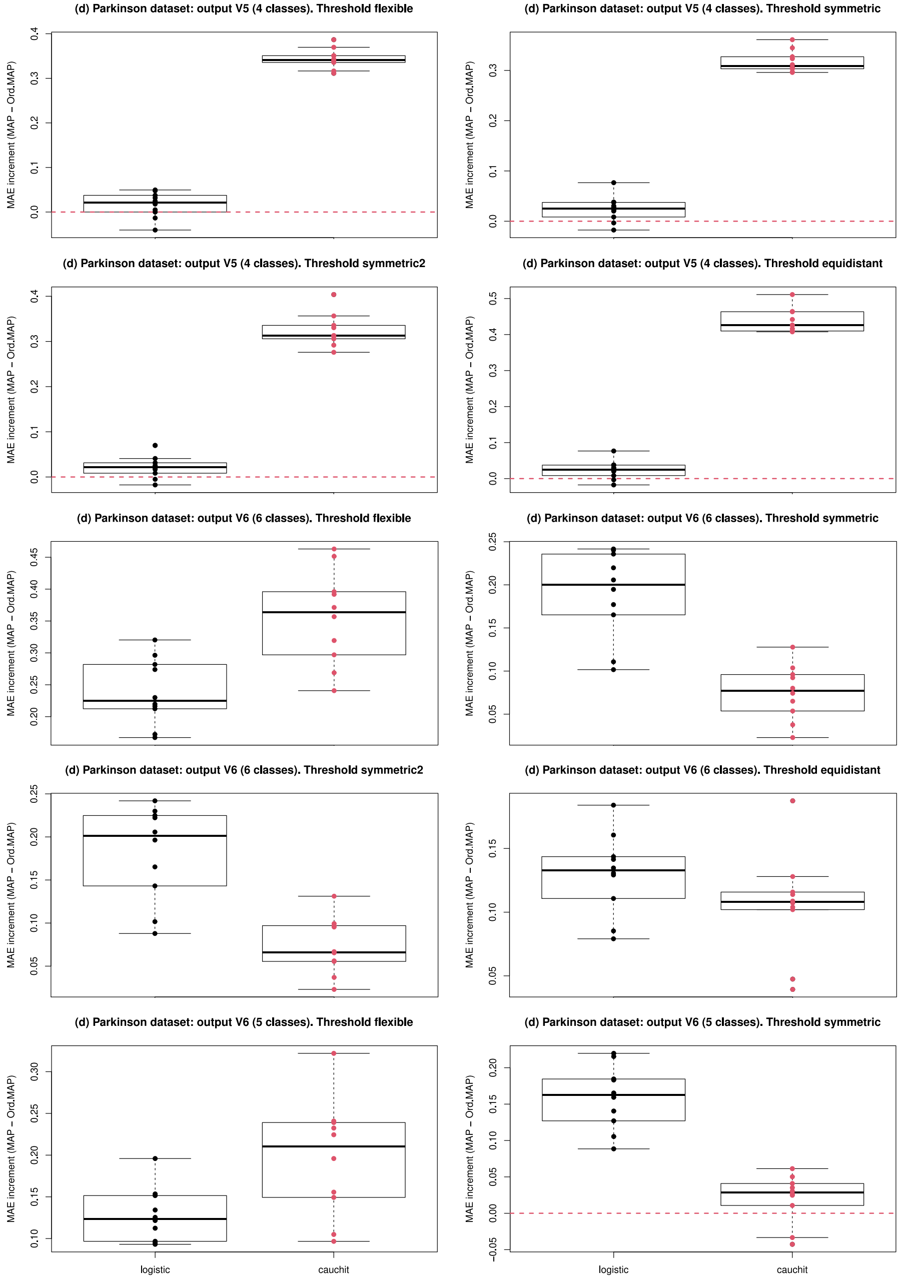


Fig. B.10. Procedure 2 (II): boxplots of MAE increments between MAP and Ord-MAP depending on the link function and the threshold structure, across datasets.


(d) Parkinson dataset: output V5 (4 classes). Threshold flexible

(d) Parkinson dataset: output V5 (4 classes). Threshold symmetric

(d) Parkinson dataset: output V5 (4 classes). Threshold symmetric2

(d) Parkinson dataset: output V5 (4 classes). Threshold equidistant

(d) Parkinson dataset: output V6 (6 classes). Threshold flexible

(d) Parkinson dataset: output V6 (6 classes). Threshold symmetric

(d) Parkinson dataset: output V6 (6 classes). Threshold symmetric2

(d) Parkinson dataset: output V6 (6 classes). Threshold equidistant

(d) Parkinson dataset: output V6 (5 classes). Threshold flexible

(d) Parkinson dataset: output V6 (5 classes). Threshold symmetric

Fig. B.11. Procedure 2 (III): boxplots of MAE increments between MAP and Ord-MAP depending on the link function and the threshold structure, across datasets.

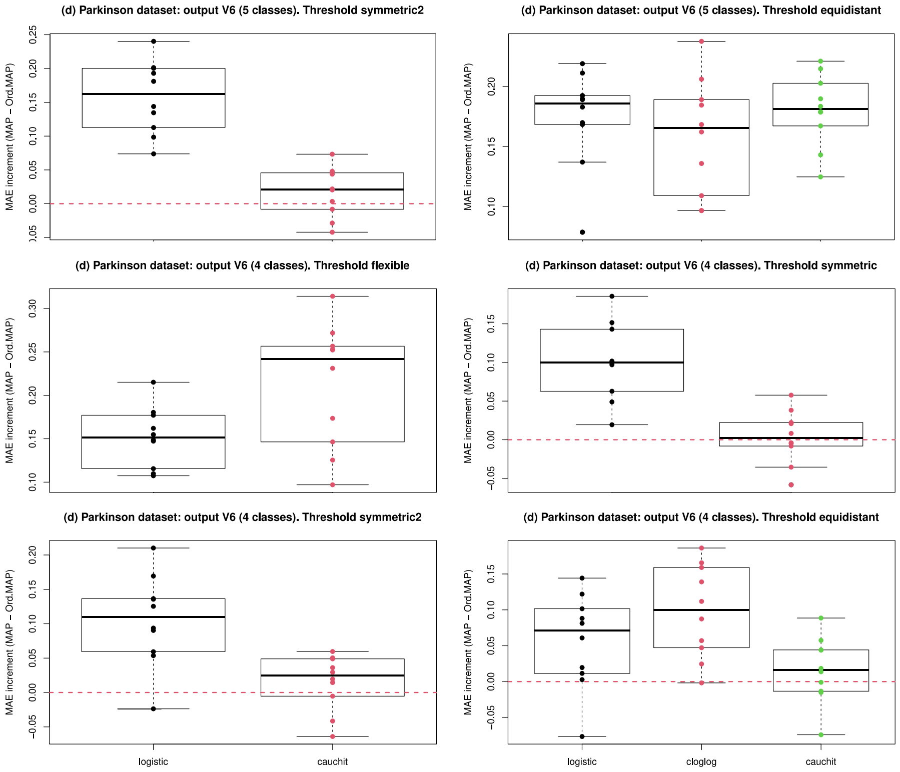


Fig. B.12. Procedure 2 (IV): boxplots of MAE increments between MAP and Ord-MAP depending on the link function and the threshold structure, across datasets.

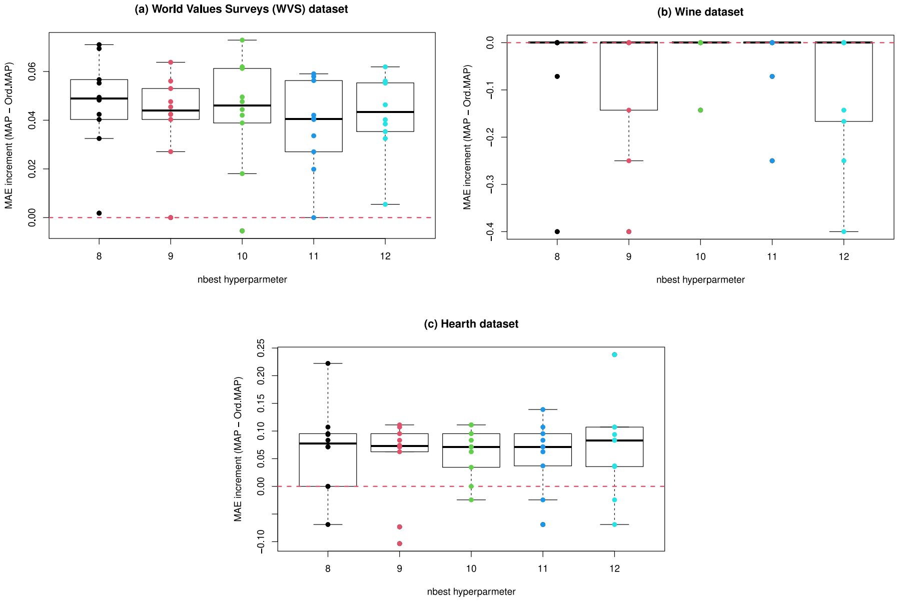


Fig. B.13. Procedure 3: boxplots of MAE increments between MAP and Ord-MAP depending on the nbest hyperparameter, for the WVS, Wine and Hearth datasets.

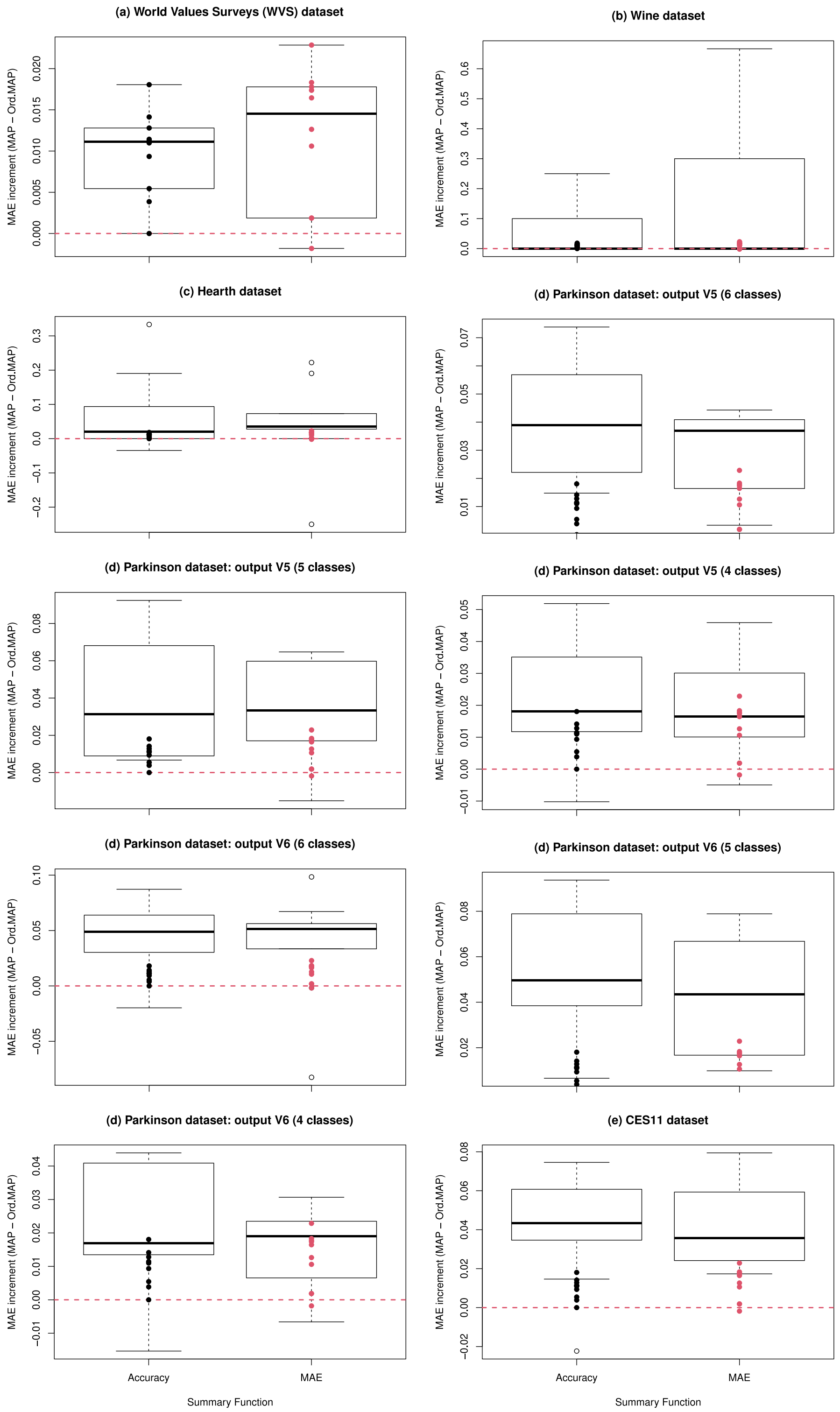


Fig. B.14. Procedure 4: boxplots of MAE increments between MAP and Ord-MAP depending on the summary function used for tuning the random forest model, across datasets.

Table B.8
Procedure 4: p-values comparing MAE for MAP and Ord-MAP criteria across datasets for the two summary functions. Only significant p-values (< 0.10) are reported, and **all** of them indicate better performance with Ord-MAP. In parentheses, the number of classes.

		Summary function	
Dataset		Accuracy	MAE
(a) WVS	(3 c)	0.004576**	0.001953**
(b) Wine	(5 c)	0.048756*	0.090725$^{\bullet}$
(c) Hearth	(5 c)	0.037963*	0.041992*
(d) Parkinson	V5 (6 c)	0.000977***	0.000977***
	V5 (5 c)	0.000977***	0.004883**
	V5 (4 c)	0.004883**	0.004883**
	V6 (6 c)	0.002930**	0.041992*
	V6 (5 c)	0.000977***	0.000977***
	V6 (4 c)	0.006836**	0.004883**
(e) CES11	(4 c)	0.002930**	0.000977***

Data availability

I have shared the source (R package or URL) of the used datasets.

References

[1] R. Likert, A technique for the measurement of attitudes, Arch. Psychol. 22 (140) (1932) 5–55.

[2] B. Pang, L. Lee, Opinion mining and sentiment analysis, Found. Trends Inf. Retr. 2 (1–2) (2008) 1–135, http://dx.doi.org/10.1561/1500000011.

[3] G. Binotto, R. Delgado, Adapting performance metrics for ordinal classification to interval scale: Length matters, Mach. Learn. 114 (41) (2025) http://dx.doi.org/10.1007/s10994-024-06654-4.

[4] J.L. García-Lapresta, R. González del Pozo, D. Pérez-Román, Metrizable ordinal proximity measures and their aggregation, Inform. Sci. 448–449 (2018) 149–163, http://dx.doi.org/10.1016/j.ins.2018.03.034.

[5] J.L. García-Lapresta, D. Pérez-Román, Aggregating opinions in non-uniform ordered qualitative scales, Appl. Soft Comput. 67 (2018) 652–657, http://dx.doi.org/10.1016/j.asoc.2017.05.064.

[6] R. Delgado, F. Fernández-Peláez, N. Pallarés, et al., Predictive risk models for COVID-19 patients using the multi-thresholding meta-algorithm, Sci. Rep. 14 (2024) 28453, http://dx.doi.org/10.1038/s41598-024-77386-7.

[7] G. Singer, A. Ratnovsky, S. Naftali, Classification of severity of trachea stenosis from EEG signals using ordinal decision-tree based algorithms and ensemble-based ordinal and non-ordinal algorithms, Expert Syst. Appl. 173 (2021) 114707, http://dx.doi.org/10.1016/j.eswa.2021.114707.

[8] L. Gaudette, N. Japkowicz, Evaluation methods for ordinal classification, Lecture Notes in Comput. Sci. 5549 (2009) 207–210, http://dx.doi.org/10.1007/978-3-642-01818-3_25.

[9] G. Singer, M. Marudi, Ordinal decision-tree-based ensemble approaches: The case of controlling the daily local growth rate of the COVID-19 epidemic, Entropy 22 (8) (2020) 871, http://dx.doi.org/10.3390/e22080871.

[10] R. Haba, G. Singer, S. Naftali, M.R. Kramer, A. Ratnovsky, A remote and personalised novel approach for monitoring asthma severity levels from EEG signals utilizing classification algorithms, Expert Syst. Appl. 223 (2023) 119799, http://dx.doi.org/10.1016/j.eswa.2023.119799.

[11] P.A. Gutiérrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, C. Hervas-Martinez, Ordinal regression methods: Survey and experimental study, IEEE Trans. Knowl. Data Eng. 28 (1) (2015) 127–146, http://dx.doi.org/10.1109/TKDE.2015.2457911.

[12] H-T. Lin, L. Li, Reduction from cost-sensitive ordinal ranking to weighted binary classification, Neural Comput. 24 (5) (2012) 1329–1367, http://dx.doi.org/10.1162/NECO_a_00265.

[13] J.S. Cardoso, R. Sousa, Classification models with global constraints for ordinal data, in: 2010 Ninth International Conference on Machine Learning and Applications, Washington, DC, USA, 2010, pp. 71–77, http://dx.doi.org/10.1109/ICMLA.2010.18.

[14] G. Singer, M. Golan, R. Shiff, D. Kleper, Evaluating the effectiveness of accommodations given to students with learning impairments: Ordinal and interpretable machine-learning-based methodology, IEEE Trans. Learn. Technol. 15 (6) (2022) 736–746, http://dx.doi.org/10.1109/TLT.2022.3214537.

[15] M. Lázaro, A.R. Figueiras-Vidal, Neural network for ordinal classification of imbalanced data by minimizing a Bayesian cost, Pattern Recognit. 137 (2023) 109303, http://dx.doi.org/10.1016/j.patcog.2023.109303.

[16] M. Marudi, I. Ben-Gal, G. Singer, A decision tree-based method for ordinal classification problems, IISE Trans. 56 (9) (2022) 960–974, http://dx.doi.org/10.1080/24725854.2022.2081745.

[17] A.M. Gómez-Orellana, D. Guijo-Rubio, P.A. Gutiérrez, C. Hervás-Martínez, V.M. Vargas, ORFEO: Ordinal classifier and regressor fusion for estimating an ordinal categorical target, Eng. Appl. Artif. Intell. 133 (Part E) (2024) 108462, http://dx.doi.org/10.1016/j.engappai.2024.108462.

[18] A.M. Gómez-Orellana, V.M. Vargas, P.A. Gutiérrez, J. Pérez-Aracil, S. Salcedo-Sanz, C. Hervás-Martínez, D. Guijo-Rubio, Energy flux prediction using an ordinal soft labelling strategy, in: J.M. Ferrández Vicente, M. Val Calvo, H. Adeli (Eds.), Bioinspired Systems for Translational Applications: From Robotics to Social Engineering, IWINAC 2024, in: Lecture Notes in Computer Science, vol. 14675, Springer, Cham, 2024, http://dx.doi.org/10.1007/978-3-031-61137-7_26.

[19] J.C. Gámez-Granados, A. Esteban, F.J. Rodriguez-Lozano, A. Zafra, An algorithm based on fuzzy ordinal classification to predict students' academic performance, Appl. Intell. 53 (2023) 27537–27559, http://dx.doi.org/10.1007/s10489-023-04810-2.

[20] C. Peláez-Rodríguez, J. Pérez-Aracil, C.M. Marina, L. Prieto-Godino, C. Casanova-Mateo, P.A. Gutiérrez, S. Salcedo-Sanz, A general explicable forecasting framework for weather events based on ordinal classification and inductive rules combined with fuzzy logic, Knowl.-Based Syst. 291 (2024) 111556, http://dx.doi.org/10.1016/j.knosys.2024.111556.

[21] L. Rabkin, I. Cohen, G. Singer, Resource allocation in ordinal classification problems: A prescriptive framework utilizing machine learning and mathematical programming, Eng. Appl. Artif. Intell. 132 (2024) 107914, http://dx.doi.org/10.1016/j.engappai.2024.107914.

[22] Savage L.J., Elicitation of personal probabilities and expectations, J. Amer. Statist. Assoc. 66 (336) (1971) 783–801, http://dx.doi.org/10.1080/01621459.1971.10482346.

[23] G.W. Brier, Verification of forecasts expressed in terms of probability, Mon. Weather Rev. 78 (1) (1950) 1–3, http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

[24] T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction and estimation, J. Amer. Statist. Assoc. 102 (47) (2007) http://dx.doi.org/10.1198/016214506000001437.

[25] A. Robert, L.A.C. Chapman, R. Grah, R. Niehus, F. Sandmann, B. Prasse, S. Funk, A.J. Kucharski, Predicting subnational incidence of COVID-19 cases and deaths in EU countries, BMC Infect. Dis. 24 (2024) 204, http://dx.doi.org/10.1186/s12879-024-08986-x.

[26] J.A. Vrugt, Distribution-based model evaluation and diagnostics: Elicitability, propriety, and scoring rules for hydrograph functionals, Water Resour. Res. 60 (6) (2024) e2023WR036710, http://dx.doi.org/10.1029/2023WR036710.

[27] M. Zamo, P. Naveau, Estimation of the continuous ranked probability score with limited information and applications to ensemble weather forecasts, Math. Geosci. 50 (2018) 209–234, http://dx.doi.org/10.1007/s11004-017-9709-7.

[28] J.R. Cano, S. García, Training set selection for monotonic ordinal classification, Data Knowl. Eng. 112 (2017) 94–105, http://dx.doi.org/10.1016/j.datak.2017.10.003.

[29] X. Liu, F. Fan, L. Kong, Z. Diao, W. Xie, J. Lu, J. You, Unimodal regularized neuron stick-breaking for ordinal classification, Neurocomputing 388 (2020) 34–44, http://dx.doi.org/10.1016/j.neucom.2020.01.025.

[30] V.M. Vargas, P.A. Gutiérrez, C. Hervás-Martínez, Cumulative link models for deep ordinal classification, Neurocomputing 401 (2020) 48–58, http://dx.doi.org/10.1016/j.neucom.2020.03.034.

[31] W. Waegeman, B. De Baets, L. Boullart, ROC analysis in ordinal regression learning, Pattern Recognit. Lett. 29 (2008) 1–9, http://dx.doi.org/10.1016/j.patrec.2007.07.019.

[32] J.S. Cardoso, R. Sousa, Measuring the performance of ordinal classification, Int. J. Pattern Recognit. Artif. Intell. 25 (8) (2011) 1173–1195, http://dx.doi.org/10.1142/S0218001411009093.

[33] R. Hornung, Ordinal forests, J. Classification 37 (2020) 4–17, http://dx.doi.org/10.1007/s00357-018-9302-x.

[34] M. Tang, R. Pérez-Fernández, B. De Baets, Fusing absolute and relative information for augmenting the method of nearest neighbors for ordinal classification, Inf. Fusion 56 (2020) 128–140, http://dx.doi.org/10.1016/j.inffus.2019.10.011.