Journal of Machine Learning Research 26 (2025) 1-56 Submitted 12/22; Revised 4/25; Published 5/25

Uplift Model Evaluation with Ordinal Dominance Graphs

Brecht Verbeken BRECHT.VERBEKEN@VUB.BE
Department of Business Technology and Operations,
Data Analytics Laboratory
Vrije Universiteit Brussel (VUB)
Pleinlaan 2, 1050 Brussels, Belgium

Marie-Anne Guerry MARIE-ANNE.GUERRY@VUB.BE
Department of Business Technology and Operations,
Data Analytics Laboratory
Vrije Universiteit Brussel (VUB)
Pleinlaan 2, 1050 Brussels, Belgium

Wouter Verbeke WOUTER.VERBEKE@KULEUVEN.BE
Faculty of Economics and Business,
KU Leuven
Naamsestraat 69, Leuven 3000, Belgium

Sam Verboven SAM.VERBOVEN@VUB.BE
Department of Business Technology and Operations,
Vrije Universiteit Brussel (VUB)
Pleinlaan 2, 1050 Brussels, Belgium

Editor: Eric Laber

Abstract

Uplift modelling is a subfield of causal learning that focuses on ranking entities by individual treatment effects. Uplift models are typically evaluated using Qini curves or Qini scores. While intuitive, the theoretical grounding for Qini in the literature is limited, and the mathematical connection to the well-understood Receiver Operating Characteristic (ROC) curve is unclear. In this paper, we introduce pROCini, a novel uplift evaluation metric that improves upon Qini in two important ways. First, it explicitly incorporates more information by taking into account negative outcomes. Second, it leverages this additional information within the Ordinal Dominance Graph framework, which is the basis behind the well known ROC curve, resulting in a mathematically well-behaved metric that facilitates theoretical analysis. We derive confidence bounds for pROCini, exploiting its theoretical properties. Finally, we empirically validate the improved discriminative power of ROCini and pROCini in a simulation study as well as via experiments on real data.

Keywords: Uplift modelling, Qini, ROC, Ordinal Dominance Graphs, pROCini

1. Introduction

Accurate prediction of the causal effects of treatments at the individual entity level is leading to radically improved decision-making in many different fields such as health care (Jaskowski and Jaroszewicz, 2012; Berrevoets et al., 2020; Verboven and Martin, 2022), marketing (Lo, 2002; Gubela et al., 2017), education (Olaya et al., 2020b) and human resources management

(Rombaut and Guerry, 2020). Often, the operational setting is subject to constraints, e.g., budgetary or capacity-wise condtraints, inducing prioritization of treatment assignment. Usually, treatment is of the greatest importance, and thus prioritized, to those individuals for whom the treatment effect is the greatest.

Uplift modelling is a subfield of causal learning that explicitly supports decisions featuring scarcity of treatment capacity through optimizing the causal effect ranking on a target population. This ranking aspect sets it apart from the classic Individual Treatment Effect (ITE) (Shalit et al., 2017) and Conditional Average Treatment Effect (CATE) (Athey et al., 2018) literature which focuses on obtaining well-calibrated point estimates of the causal effect.

Specialized ranking metrics such as the Qini score have been proposed to evaluate uplift models (Radcliffe, 2007; Devriendt et al., 2020; Gutierrez and Gérardy, 2017; Belbahri et al., 2021). Although initial uplift evaluation metrics lack a solid theoretical basis, only recently has research attempted to establish connections to existing theoretical frameworks (Yadlowsky et al., 2024).

Furthermore, although similarity in intuition is claimed, there is no explicit link with the area under the ROC curve (AUROC), a commonly used ranking evaluation metric for assessing classification performance. This lack of mathematical grounding makes it challenging to assess the significance of metric outcomes correctly. For example, many papers have reported the unstable behaviour of uplift models (Olaya et al., 2020a; Diemert et al., 2018; Devriendt et al., 2018). This instability has previously been attributed to the characteristics of the data set, the models, and the evaluation metric. The main roadblock to a deeper understanding of uplift modelling results is the lack of well-understood mathematical evaluation. Introducing such a metric for causal effect ranking that allows for theoretical grounding thus represents a fundamental step for uplift modelling to mature as a field of study, and is the key research objective and contribution of this paper.

In the next section, we review the preliminaries of uplift modelling and its evaluation. Afterwards, we turn to the Qini score and set up a simulation protocol to study the properties of the distribution of the Qini score. In Section 3, we review the connection between the Qini curve and the ROC curve. As a stepping stone towards our main contribution, we introduce the ROCini score. This measure lays the foundation for our primary innovation, the pROCini score, by illustrating how to capture more relevant information within an uplift evaluation metric. Using Ordinal Dominance Graphs, we then present a mathematical foundation to extend the ROCini score to the pROCini score, a mathematically well-behaved metric that allows for theoretical grounding. Furthermore, through its direct connection to the ROC curve, the pROCini curve can be linked to work on the ROC curve of the past fifty years. We demonstrate the superior theoretical properties of the pROCini score by deriving confidence bounds in Section 3.3. In Section 3.4, we propose a simulation study that can be used to compare the performance of various uplift modelling metrics. We report that the ROCini and pROCini scores outperform the existing Qini scores in terms of discriminative power. Finally, in Section 4, we present experiments on real data revealing that in realistic scenarios the choice of the evaluation metric impacts the model selection.

2. Preliminaries and Background

2.1 Uplift modelling

Uplift modelling aims to inform optimal treatment assignment by ranking individuals according to their estimated net benefit from treatment, often using methods related to CATE estimation but optimized for decision-making rather than pure estimation accuracy (Gubela et al., 2020; Fernández and Provost, 2019; De Vos et al., 2024). We focus on uplift modelling with binary treatments $\mathcal{T}_i \in \{0, 1\}$, which is the most common case in the literature, where $\mathcal{T}_i = 1$ indicates that the treatment is applied to individual i (treatment group), whereas $\mathcal{T}_i = 0$ signifies that the individual is not treated (control group). The potential outcomes for each individual can be represented as Y_T and Y_C, where Y_T is the outcome if treated ($\mathcal{T}_i = 1$), and Y_C is the outcome if not treated ($\mathcal{T}_i = 0$). The outcome is binary, and thus both Y_T and Y_C take values in $\{0, 1\}$.

The uplift for an individual i can then be formalized as follows:

$$\tau_i = \Pr(Y_T = 1 \mid \mathbf{X}_i) - \Pr(Y_C = 1 \mid \mathbf{X}_i). \tag{1}$$

However, for any given individual, only one of these potential outcomes is observable—either Y_T or Y_C—which is commonly referred to as the fundamental problem of causal inference (Holland, 1986). The observed outcome for individual i is defined as

$$Y_i = \mathcal{T}_i Y_T + (1 - \mathcal{T}_i) Y_C. \tag{2}$$

In line with the uplift modelling literature, we assume that the treatment is randomly assigned (as in a randomized controlled trial). This implies that *strong ignorability* (Rosenbaum and Rubin, 1983) is satisfied—a combination of the following: (i) ignorability, meaning that the potential outcomes are conditionally independent of treatment assignment given covariates,

$$(Y_T, Y_C) \perp \mathcal{T}_i \mid \mathbf{X}_i, \tag{3}$$

and (ii) positivity, ensuring that all individuals have a nonzero probability of receiving either treatment or control,

$$0 < \Pr(T_i = 1 \mid X_i) < 1 \quad \forall\, X_i. \tag{4}$$

Furthermore, it is assumed that the observed outcome corresponds to the potential outcome under the received treatment and that there is no interference between individuals. Together, these assumptions ensure that causal effects are identifiable.

2.2 Evaluation of uplift models

Owing to the fundamental problem of causal inference, uplift models require specific evaluation metrics. The most common metric is the Qini score, which is obtained from the Qini curve (Radcliffe, 2007). However, various definitions and implementations of both the Qini curve and the closely related uplift curve exist in the literature. These variations include different normalization techniques and methodologies for ranking subjects, with some approaches ranking all subjects together and others ranking control and treatment groups separately (Devriendt et al., 2020). Furthermore, generalizations such as the Adjusted Qini Curve and the Cumulative Gains Qini curve have been proposed (Gutierrez and Gérardy,

2017). For the purposes of this paper, we define the Qini curve using a joint ranking approach that combines both control and treatment groups. Specifically, we consider the following:

The Qini curve is the curve of the function $Q(\varphi)$ defined as

$$Q(\varphi) = \frac{n_T^1(\varphi)}{n_T} - \frac{n_C^1(\varphi)}{n_C}. \tag{5}$$

Where n_T and n_C are the numbers of people in the treatment and control groups, respectively, and where $n_C^1(\varphi)$ and $n_T^1(\varphi)$ correspond to the numbers of people with favourable outcomes in the first φ proportion (ranked from highest to lowest estimated uplift) of subjects in the control and treatment groups, respectively.

The random chance Qini curve is represented by a straight line through (0,0) and

$$\left(1,\ \frac{n_T^1}{n_T} - \frac{n_C^1}{n_C}\right).$$

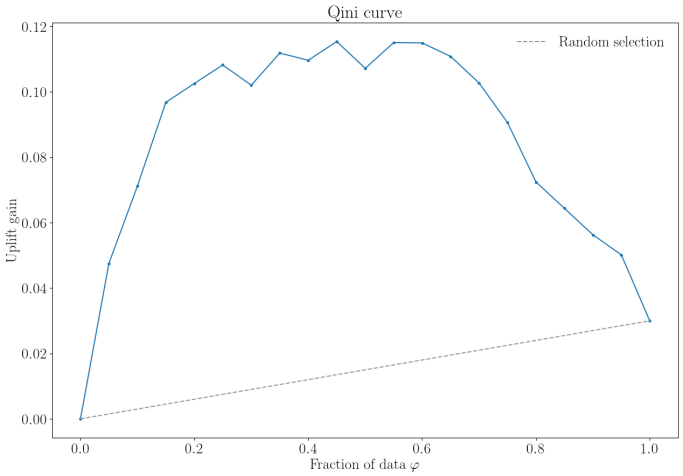


Figure 1: Example of a Qini curve

The Qini score (QS) is then defined as the area between the Qini curve of the model and the random chance Qini curve. A QS can be calculated for the whole population, or a certain percentile, and compresses the uplift ranking performance into a single value. These curves, often referred to as "uplift curves" in a general context, have been extended with various alternative weighting schemes to address group imbalances as proposed in (Gutierrez and Gérardy, 2017). However, in this paper, we focus on the Qini curve and score, as defined in Eq. (5) as they remain the most widely used metrics in uplift modelling. The Qini score can be seen as the baseline metric upon which various extensions are built. Importantly, the alternative weighting schemes developed for the Qini curve can be applied to the new metrics we introduced in Section 3.

In practice, the Qini score and its extensions are used to tune the hyperparameters of uplift models, model selection, decide whether to push models to production, and ultimately to design treatment assignment policies. It is thus of primordial importance that the Qini score is a reliable, well-defined metric. As such, it is worthwhile to study the properties of the Qini score and its distribution. Recently, an alternative approach to scoring uplift models has been developed: the Rank-Weighted Average Treatment Effects (RATE) (Yadlowsky et al., 2024). This method offers broader applicability as it can be used with continuous outcomes in addition to binary outcomes. The Targeting Operator Characteristic (TOC) (Zhao et al., 2013) is an example of a RATE curve, and the area under the TOC (AUTOC) has been proposed as a metric to evaluate uplift models. In the context of uplift modelling, the TOC can be defined as follows:

$$\mathrm{TOC}(\varphi) = \frac{n_T^1(\varphi)}{n_T(\varphi)} - \frac{n_C^1(\varphi)}{n_C(\varphi)} - \frac{n_T^1}{n_T} + \frac{n_C^1}{n_C}. \tag{6}$$

2.3 A first look at the Qini

We first analyse the stochastic behaviour of the Qini score, which reflects its ability to discriminate between treatment responses. We use Algorithm 1 to gain insights into the distribution of the Qini scores. The idea is to fix an underlying data-generating model to examine the distribution of the Qini scores in a toy setup with high aleatoric uncertainty.

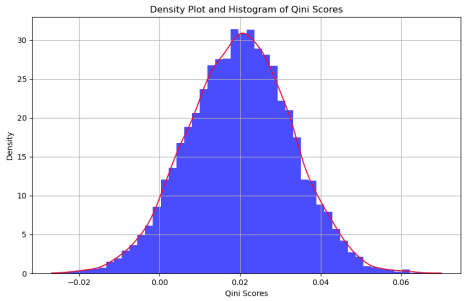
In Algorithm 1 the ground truth uplift ranking represents the perfect model. We then add Gaussian noise to represent the model error. For each individual we sample three probabilities: the probability of a positive outcome conditional on being in the control group, the individual uplift and the Gaussian error. The sum of those three quantities corresponds to the probability of a positive outcome conditional on being in the treatment group. The control group probability (PC) was drawn from a beta distribution. The individual uplift (IU) was drawn from a normal distribution and clamped to ensure that $0 \leq \mathrm{PC} + \mathrm{IU} \leq 1$. The final uplift with noise (IUn) was also drawn from a normal distribution and clamped to satisfy $0 \leq \mathrm{PC} + \mathrm{IU} + \mathrm{IUn} \leq 1$.

The outcomes are sampled with two Bernoulli experiments per individual. A first trial determines the group assignment (control group C or treatment group T), and a second trial determines the binary value of the observed outcome. Finally, the Qini score is calculated. This procedure is repeated $r = 10000$ times. The results are shown in Figure 6.

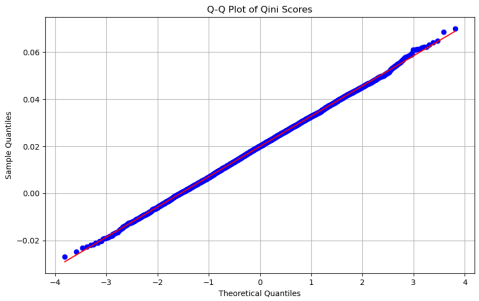
Algorithm 1 Simulation of uplift model scores

Symbols:
  $\mathcal{N}$: normal distribution
  $\mathcal{B}$: beta distribution
  B: Bernoulli distribution
Input:
  r: the number of runs
  N: the total number of individuals in the sample
  α, β: parameters of the beta distribution
  v: the variance of the individual uplifts
  E: a list of variances of the Gaussian error
  S: a list of uplift model metrics
Initialize:
  Draw probabilities of positive outcome in control group $\text{PC} \sim \mathcal{B}(\alpha, \beta)$
  Draw individual uplifts $\text{IU} \sim \mathcal{N}(0, v)$
  Cap IU: $\text{IU} = \max(-\text{PC}, \min(\text{IU}, 1 - \text{PC}))$
  Draw individual uplifts with Gaussian error $\text{IUn} \sim \mathcal{N}(0, \epsilon)$
  Cap IUn: $\text{IUn} = \max(-\text{PC} - \text{IU}, \min(\text{IUn}, 1 - \text{PC} - \text{IU}))$
for $\epsilon \in E$ **do** ▷ fix model error
  for $j = 1 \dots r$ **do**
    for $k = 1 \dots N$ **do**
      draw $\text{Obs} \sim B(0.5)$ ▷ Bernoulli experiment
      if $\text{Obs} = 0$ **then**
        draw $\text{Out} \sim B(\text{PC})$
      else
        draw $\text{Out} \sim B(\text{PC} + \text{IU})$
      end if
    end for
    for score in S **do**
      $\text{score}(\text{Obs}, \text{Out}, \text{IU} + \text{IUn})$ ▷ score data with error
    end for
  end for
end for

The distribution of the Qini scores as depicted in Figure 2 is asymptotically normal. This is validated using a Shapiro–Wilk test (Shapiro and Wilk, 1965), which yields $W = 0.9998$ and $p = 0.55$. This result implies that the normality hypothesis cannot be rejected for $\alpha = 0.05$. Furthermore, it is noteworthy that part of the QS distribution lies beneath zero, corresponding to a worse than random performance. Such negative Qini scores may arise in scenarios where treatment effects are difficult to predict or are only weakly identifiable from covariates, leading to misrankings that perform worse than random.



(a) Density plot of the simulated distribution



(b) Q-Q plot of simulated vs. normal distribution

Figure 2: Results of Algorithm 1 with $r = 10\,000$, $N = 1\,000$, $(\alpha, \beta) = (12, 12)$, $v = 0.1$, $E = \{0.1\}$, $S = \{\text{QS}\}$.

In a second experiment we vary the main parameters $[N, (\alpha, \beta), v, E]$ of Algorithm 1 in order to determine their effect on the distribution of the Qini score. The results in Table 1 highlight that the population size significantly affects the variance of the distribution of the Qini score. As expected, a larger population size leads to a more narrower distribution, whereas the mean remains unchanged. In Figure 3 we observe that a stronger signal, in the sense of a larger v, subsequently leads to a distribution shift to the right, which corresponds to a higher Qini score, whereas adding a larger model error leads to a lower Qini score in each case.

In conclusion, using the Qini score to evaluate uplift models in settings with small population sizes and small effect sizes may lead to misguided results that are not generalizable. This could offer an explanation for the instability of uplift models as evaluated using the Qini score, as reported in the literature (Devriendt et al., 2018).

2.4 Connection of the Qini curve with the ROC

The ROC curve is a graph that illustrates the diagnostic ability of a binary classifier. The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at different threshold levels $t \in [0, 1]$. Given for each individual i the probability p_i of a positive outcome, we can classify for the threshold level t all individuals i with $p_i > t$ as positive and calculate the rates $\text{TPR}(t)$ and $\text{FPR}(t)$ corresponding to this classification. These rates can be considered the coordinates of the point $(\text{FPR}(t), \text{TPR}(t))$.

If we let t vary between 0 and 1 we obtain a graph of a parametric function (with parameter t) in the space $[0, 1] \times [0, 1]$, which is called the ROC curve. In general, this graph is stepwise. However, in cases where there are ties in scores among individuals with different true outcomes, the graph will contain slanted segments.

Conversely, the Qini curve, as defined in Eq. (5), is the graph of a function of one variable (φ). In fact, this function aggregates the information of the two rates $\frac{n_T^1(\varphi)}{n_T}$ and $\frac{n_C^1(\varphi)}{n_C}$.

It is important to emphasize the difference in how the ROC curve and the Qini curve are constructed: one way to draw a parallel between the ROC curve and the Qini curve is

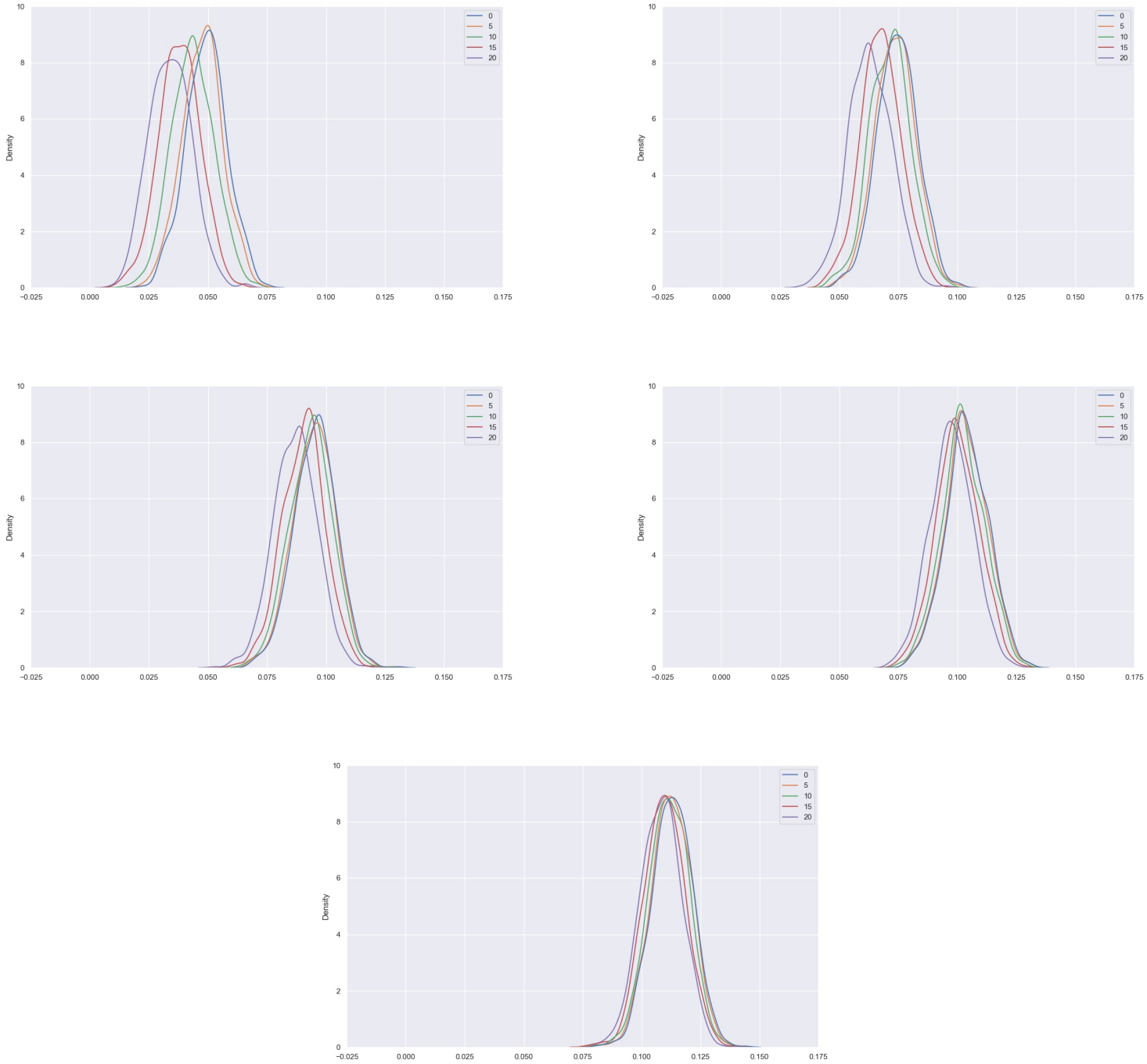
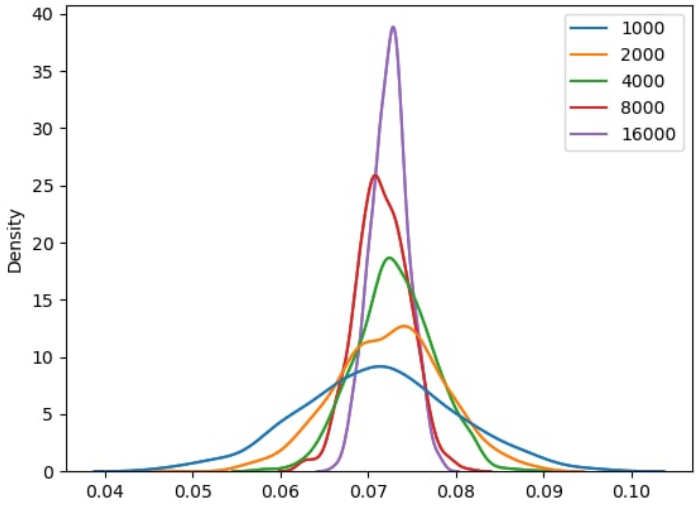


Figure 3: The results of Algorithm 1 with $r = 1000, N = 1000, (\alpha, \beta) = (12, 12)$ and from left to right with $v = 0.20, v = 0.35, v = 0.5, v = 0.65$ and $v = 0.80$ with $E = \{0.05, 0.1, 0.15, 0.2\}$ in each subfigure.



N	mean	variance
1000	0.711	78×10^{-6}
2000	0.724	38×10^{-6}
4000	0.730	20×10^{-6}
8000	0.715	9×10^{-6}
16000	0.724	5×10^{-6}

Table 1: Distribution of the Qini score for various population sizes with $r = 1000, (\alpha, \beta) = (2.5, 2.5), v = 0.35$ and $E = \{0.05\}$.

to identify $\frac{n_T^1(\varphi)}{n_T}$ with the true positive rate and $\frac{n_C^1(\varphi)}{n_C}$ with the false positive rate. This was explored in (Kuusisto et al., 2014) where the similarities between uplift curves and ROC curves were exploited to write the Qini curve as the difference between two curves. This approach could be utilized to derive confidence bounds for the Qini curve and the associated Qini Score.

However, in the ROC setting, the proportion φ of the population that corresponds to a certain point on the curve is only implicit since both the TPR(t)-axis and the FPR(t)-axis are functions of this proportion. While this proportion is not as such visible on the ROC curve, the Qini curve is graphed in a coordinate system where φ is presented on the horizontal axis. Due to this lack of a direct connection, the theoretical work conducted on ROC can not be straightforwardly applied to the Qini curve.

The Qini curve, which was originally developed independently, was later incorporated into a broader analytical framework by (Yadlowsky et al., 2024), who demonstrated that the Qini score could be situated within the family of rank-weighted average treatment effect (RATE) metrics. This general family of measures is designed for evaluating and comparing the effectiveness of treatment prioritization strategies. Notably, this framework also encompasses the (AU)TOC measure (see Eq. (**??**)).

RATE metrics provide a way to summarize the quality of a treatment prioritization rule in ranking units according to their potential outcomes without committing to a specific treatment policy. This formulation enables precise tailoring: by selecting appropriate weight functions, researchers can emphasize different segments of the ranking—such as the top decile—to align with application-specific costs and error tolerances. Moreover, because the RATE framework is rooted in the Neyman–Rubin potential outcomes model, it delivers

direct interpretability in terms of causal treatment effects and supports rigorous statistical inference.

3. The ROCini and pROCini metrics

3.1 ROC-like

In this section, we propose an alternative for the Qini curve that (i) explicitly includes more information, (ii) is more closely tied to the classical ROC curve, allowing improved theoretical grounding, and (iii) behaves mathematically better, as it is bound to the unit square and consistently has $(0, 0)$ as the starting point and $(1, 1)$ as the ending point. To understand our approach, first consider the classical ROC curve. A ROC curve can be viewed as a plot of "good" cases versus "bad" cases. Specifically, the y-axis represents the True Positive Rate (proportion of actual positives correctly identified), which we can think of as "good" cases. The x-axis represents the False Positive Rate (proportion of actual negatives incorrectly identified as positive), which we can consider as "bad" cases.

In uplift modelling with binary treatments, individuals are typically categorized into four key segments to optimize targeted interventions: Lost Causes, Do Not Disturbs, Persuadables, and Sure Things. Lost Causes are individuals who will not respond positively regardless of the treatment, making any effort wasted. Do Not Disturbs are those who might react negatively to the treatment, potentially causing harm or dissatisfaction if targeted. Persuadables are the primary focus, as they are likely to respond positively to the treatment and thus represent the most efficient use of resources. Finally, Sure Things are individuals who will respond positively without any intervention, making targeting them redundant.

When adapting the ROC-concept to uplift modelling, careful consideration is given to defining "good" and "bad" cases within the context. For binary outcomes, the objective is to distinguish between individuals who should be targeted ("Good Targets") and those who should not ("Bad Targets"). "Good Targets" are identified where the outcome is positive with treatment ($Y_T = 1$) and negative without treatment ($Y_C = 0$), which are also known as Persuadables. "Bad Targets" encompass all other cases, including Lost Causes, Sure Things, and Do Not Disturbs. Specifically, some "Bad Targets" can be directly identified from the data: cases with $(Y = 0, \mathcal{T} = 1)$ must be either Do Not Disturbs or Lost Causes, whereas cases with $(Y = 1, \mathcal{T} = 0)$ must be either Do Not Disturbs or Sure Things. To apply ROC methodologies, we define positive instances as $(Y = 1, \mathcal{T} = 1)$ and $(Y = 0, \mathcal{T} = 0)$, which are potentially "Good Targets", and negative instances as $(Y = 1, \mathcal{T} = 0)$ and $(Y = 0, \mathcal{T} = 1)$, which are definitely "Bad Targets".

Considering the treatment group (T) and the control group (C) as two distinct entities, we can now perform an ROC-like analysis as follows:

T : We identify the fraction $\frac{n_T^1(\varphi)}{n_T^1}$ with the TPR and the fraction $\frac{n_T^0(\varphi)}{n_T^0}$ with the FPR.

C : We identify the fraction $\frac{n_C^0(\varphi)}{n_C^0}$ with the TPR and the fraction $\frac{n_C^1(\varphi)}{n_C^1}$ with the FPR.

Suppose that $\Pr(\mathcal{T}=1) = \Pr(T|X^i)$ and that the individuals are ranked by their uplift. For the treatment group, a good ranking corresponds to individuals with a high estimated uplift being relatively more likely to respond positively, i.e., more likely to be in $n_T^1(\varphi)$. Conversely, for the control group, a high estimated uplift corresponds to a relatively lower chance of belonging to $n_C^1(\varphi)$. We can then use the uplift as a ranking criterion to combine those two groups. Keeping the structure of the Qini curve, i.e., plotting an informative variable along the y-axis as a function of the proportion φ on the x-axis, the ROCini curve is obtained:

$$\text{ROCini}(\varphi) = (\frac{n_T^1(\varphi)}{n_T^1} - \frac{n_T^0(\varphi)}{n_T^0}) + (\frac{n_C^0(\varphi)}{n_C^0} - \frac{n_C^1(\varphi)}{n_C^1}). \tag{7}$$

Remark 1 *In the special case where the class proportions are equal, i.e., $n_T^0 = n_T^1 = n_C^0 = n_C^1$, we can use the identities $n_T^0(\phi) = n^T(\phi) - n_T^1(\phi)$ and $n_C^0(\phi) = n^C(\phi) - n_C^1(\phi)$ to rewrite ROCini as:*

$$\text{ROCini}(\phi) = 2(n_T^1(\phi) - n_C^1(\phi)) + (n_C(\phi) - n_T(\phi)),$$

For a randomized experiment, the second term is expected to be small, leading to:

$$\text{ROCini}(\phi) \approx 2\, Qini(\phi).$$

This implies that under these conditions, ROCini is essentially a scaled version of Qini. However, this holds only in this balanced scenario. In general, when class proportions differ, the additional terms in ROCini capture the structural imbalances between the treatment and control groups.

Analogous to the Qini curve, the ROCini curve can be used to obtain a ROCini score by calculating the area underneath the ROCini curve. The ROCini curve inherently includes more (explicit) information through additionally incorporating $n_T^0(\varphi)$ and $n_C^0(\varphi)$. Although a classical ROC function ranges between 0 and 1, the ROCini function ranges from 0 to 1 and then back to 0. Note that the area underneath the ROCini curve lies between -1 and 1. The ROCini score (ROCiniS) is then defined as the area under this curve.

3.2 Connecting the ROCini to the ROC using Ordinal Dominance Graphs: the pROCini

Like those of Qini, the axes of ROCini differ significantly from those of the traditional ROC. Ordinal Dominance Graphs (ODG) (Darlington, 1973) can be used to compare the probability density functions of two random variables, X and Y. Using Ordinal Dominance Graphs, we can cast the ROCini, similar to the ROC, in the form of two parametric functions of the proportion of the population, along the axes. In this way we cannot only visualise all the information we want to use in the unit square, but more importantly we can build on the extensive literature that was developed for the ROC curve and use it in the setting of uplift modelling.

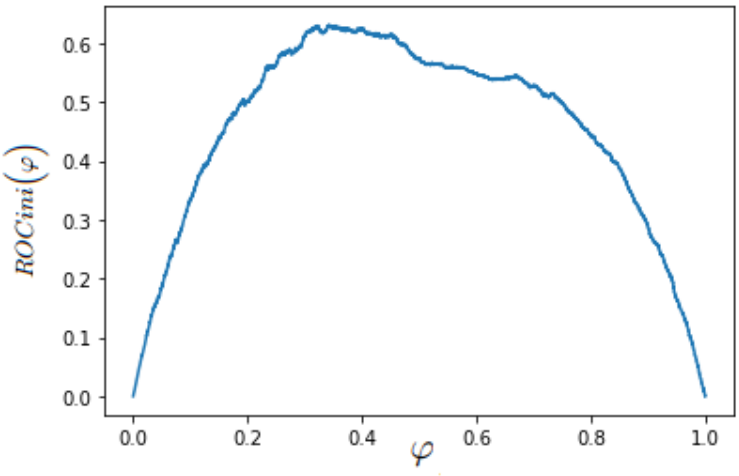


Figure 4: Example of a ROCini curve

Definition 2 *(Bamber, 1975) For an arbitrary $t \in [0,1]$, we define $G(t)$ as follows:*

$$G(t) = (X(t), Y(t)) = (\Pr(X \leq t), \Pr(Y \leq t)). \tag{8}$$

The Ordinal Dominance Graph $ODG(X,Y)$ is then defined as the curve consisting of all points $G(t)$.

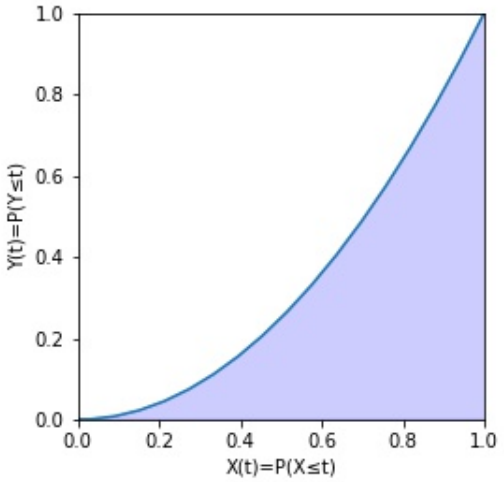


Figure 5: Example of an ODG where the AUC corresponds to the blue region

Note that the ROC curve can be viewed in this setting:

Theorem 3 *(Bamber, 1975) The ROC curve is a rotated ODG with* $\Pr(X \geq t) = FPR(t)$ *and* $\Pr(Y \geq t) = TPR(t)$

The Qini score (Radcliffe and Surry, 2011) was inspired by the AUROC, the Area Under the ROC curve, which corresponds to the area above an ODG. Similarly, a score can be derived by taking the area above any ODG. Specifically, the area above the $ODG(X, Y)$, denoted $A(X, Y)$, measures the extent to which the distribution of X lies underneath the distribution of Y. In the case of the ROC, this interpretation coincides with the interpretation of the AUROC as the probability that a random positive instance ranks above a random negative instance.

Lemma 4 *(Bamber, 1975) For every ODG(X,Y) we have the following:*

$$A(X, Y) = \Pr(X \leq Y)$$

Proof

$$\begin{aligned} A(X,Y) &= \int_0^1 X(t)\, Y'(t)\, dt \\ &= \int_0^1 \Pr(X \leq t)\, d\Pr(Y \leq t) \\ &= \int_0^1 \Pr(X \leq t)\, \mathrm{PDF}_Y(t)\, dt \\ &= \Pr(X \leq Y). \end{aligned}$$

where $PDF_Y(t)$ represents the probability density function of Y. ■

To obtain an even better performance metric, we propose applying the ordinal dominance graph framework to the previously constructed ROCini. This step allows us to create a more general and flexible metric. We start by redefining the True Positive Rate (TPR) and False Positive Rate (FPR) as weighted averages of the TPR and FPR in the treatment and control groups, respectively:

$$\mathrm{TPR}(\varphi) = w_p \frac{n_T^1(\varphi)}{n_T^1} + (1 - w_p) \frac{n_C^0(\varphi)}{n_C^0}\,, \tag{9}$$

$$\mathrm{FPR}(\varphi) = w_n \frac{n_T^0(\varphi)}{n_T^0} + (1 - w_n) \frac{n_C^1(\varphi)}{n_C^1}\,. \tag{10}$$

where $w_p \in [0, 1]$ is the weight of the treatment group for the TPR and where $w_n \in [0, 1]$ is the weight of the treatment group for the FPR. From this general formulation, we can derive two specific instances:

The pROCini (probabilistic ROCini), which we propose, sets $w_p = w_n = \frac{1}{2}$, resulting in:

$$\mathrm{pROCini}(\varphi) = \left(\frac{\frac{n_T^0(\varphi)}{n_T^0} + \frac{n_C^1(\varphi)}{n_C^1}}{2}\,,\ \frac{\frac{n_T^1(\varphi)}{n_T^1} + \frac{n_C^0(\varphi)}{n_C^0}}{2} \right). \tag{11}$$

where $\varphi \in [0, 1]$ corresponds to the φ percentage of the highest ranked individuals. We divide both $X(\varphi)$ and $Y(\varphi)$ by two to transform the graph into the unit square in a uniform

fashion, as in Definition 2. The pROCini score (pROCiniS) is then defined as the area under this curve.

The CROC, as defined by Verbeke et al. (2020), sets w_p and w_n to correspond to the fractions of treated individuals among positive and negative examples, respectively. Specifically, $w_p = \frac{n_T^1}{n_T^1+n_C^0}$ and $w_n = \frac{n_T^0}{n_T^0+n_C^1}$, which coincides with our pROCini curve in the case where $n_T^1 = n_C^0$ and $n_T^0 = n_C^1$, resulting in:

$$\text{CROC}(\varphi) = \Big(\frac{n_T^0(\varphi) + n_C^1(\varphi)}{n_T^0 + n_C^1}\ ,\ \frac{n_T^1(\varphi) + n_C^0(\varphi)}{n_T^1 + n_C^0}\Big). \tag{12}$$

3.3 Inherited properties from the ODG framework

3.3.1 Interpretations and general properties

Embedding metrics within the Ordinal Dominance Graph (ODG) framework offers several advantages. Notably, by applying Lemma 4, we gain an intuitive interpretation of the pROCiniS (or its generalizations). It can be understood as the probability that a randomly selected "Good Target" is ranked higher by the model than a randomly chosen "Bad Target". Importantly, however, these "targets" are idealized concepts, as they are defined as weighted averages (see Equations (9,10)). This interpretation provides a clear and meaningful way to assess the performance of uplift models in discriminating between potentially persuadable individuals and those who are likely not to be influenced by the treatment.

The ODG framework also provides a natural criterion for distinguishing between "Good Targets" and "Bad Targets", which leads to a candidate cut-off point. This approach is analogous to the use of Youden's J statistic in traditional ROC analysis (Peirce, 1884). In the context of uplift modelling, we can define a similar statistic as:

$$J(\varphi) = \text{TPR}(\varphi)\ -\ \text{FPR}(\varphi). \tag{13}$$

where $\text{TPR}(\varphi)$ and $\text{FPR}(\varphi)$ are as defined in Equations (9, 10). The optimal cut-off point corresponds to the maximum value of $J(\varphi)$ over all possible thresholds φ. This has an intuitive geometric interpretation: it represents the maximum vertical distance between the ODG curve and the diagonal (or chance) line (Schisterman et al., 2005).

From a practical standpoint, evaluation typically focuses on specific portions of the curve, such as the Qini score at 10% of the population. This practice acknowledges that in many real-world applications, interventions are often limited to a subset of the population owing to resource constraints. The ODG framework naturally accommodates this approach, as demonstrated by (Dodd and Pepe, 2003). In the same spirit, it is worth noting that different uplift models may excel at predicting uplift for different segments of the population distribution. This phenomenon is analogous to how different ROC curves may outperform others in specific regions of the plot.

In summary, embedding metrics within the ODG framework offers clear methodological advantages: metrics such as pROCini provide greater computational efficiency and robustness owing to reliance on empirical proportions rather than nuisance parameter estimation, thereby reducing sensitivity to model misspecification. Additionally, the ODG framework

inherently supports cost-sensitive weighting schemes ((Shao et al., 2023)), facilitates natural covariate stratification for exploring treatment heterogeneity ((Dodd and Pepe, 2003; Sukhatme and Beam, 1994)), and ensures rigorous statistical inference through established confidence interval methods. Furthermore, the framework naturally facilitates stratification by covariates—enabling evaluation within subpopulations to reveal treatment heterogeneity (Dodd and Pepe, 2003; Sukhatme and Beam, 1994). By explicitly incorporating both positive and negative cases, the ODG-based approach provides a balanced assessment of discriminatory power with established statistical methods for constructing confidence intervals that support rigorous model comparisons.

3.3.2 Confidence bounds for model evaluation scores

In practice, it is necessary to be able to discriminate between AUROC values when comparing the performance of different models. For this reason, much effort has been expended in previous work to develop confidence bounds for AUROC values (Hilgers, 1991; Cortes and Mohri, 2004). These techniques are suitable for the more general setting of ODG as well and most of the ideas behind them were already presented in (Bamber, 1975).
In some uplift modelling applications, such as marketing (Baier and Stöcker, 2022), the population size is often large. In such cases, due to the central limit theorem, it is often reasonable to assume that the X and Y variables follow a normal distribution. If X and Y are independent as well, we obtain

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2) \quad \text{and} \quad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2).$$

In previous work (Somoza and Mossman, 1991), it was established that this leads to:

$$A(X,Y) = \phi\Big(\frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}}\Big).$$

Moreover, the quantity $\frac{\mu_Y - \mu_X}{\sqrt{\sigma_Y^2 + \sigma_X^2}}$ is closely related to d_a, the index of discrimination between normal distributions, which was introduced in (Simpson and Fitter, 1973):

$$d_a = \frac{\mu_Y - \mu_X}{\sqrt{\frac{\sigma_X^2 + \sigma_Y^2}{2}}}.$$

Furthermore, as noted in (Hanley et al., 1983), for continuous X and Y we have that

$$A(X,Y) = \frac{U}{N_X \times N_Y}. \tag{14}$$

where U is the Mann-Whitney statistic and where N_X and N_Y are the numbers of observations in the X and Y groups respectively. Note that it might be possible that observations in X (or Y) have to be weighted differently (as is the case for the pROCini).This can be resolved by setting $N_X = 2 \times \min\{N_{X_1}, N_{X_2}\}$ and $N_Y = 2 \times \min\{N_{Y_1}, N_{Y_2}\}$, where N_{X_i} and N_{Y_i} correspond to the number of people of type i in X and Y respectively . The existing estimates (Mason and Graham, 2002; Hanley et al., 1983; Cortes and Mohri, 2004; Macskassy and Provost, 2004) for the variance of $A(X,Y)$ can be used in this setting.

Most notably:

Van Dantzig (Van Dantzig, 1951):

$$s^2_{\max} = \frac{A(1-A)}{N_L} \quad \text{with} \quad N_L = \min\{N_X, N_Y\}, \tag{15}$$

Hanley & McNeil (Hanley et al., 1983):

$$s^2_A = \frac{A(1-A) + (N_X - 1)(Q_1 - A^2) + (N_Y - 1)(Q_2 - A^2)}{N_X N_Y} \tag{16}$$

$$\text{with } Q_1 = \frac{A}{2-A} \text{ and } Q_2 = \frac{2A^2}{1+A}.$$

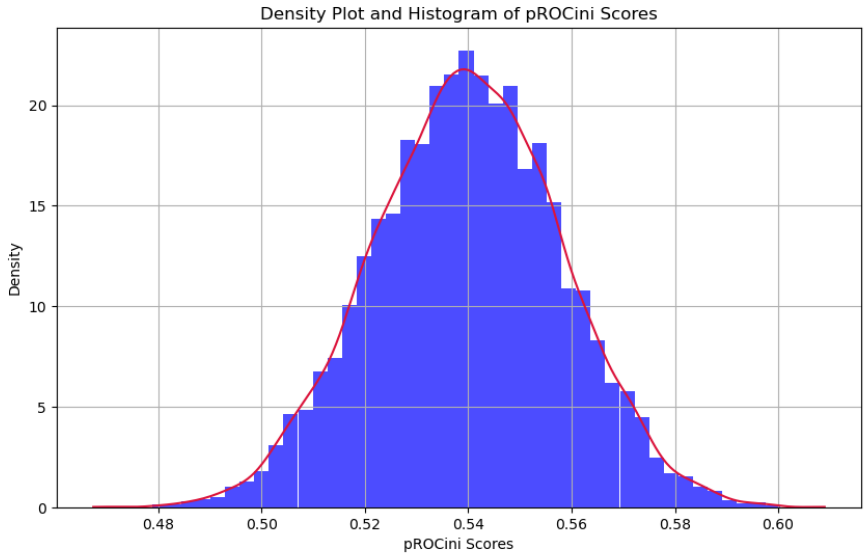
Where Eq. (15) leads to less tight estimates than Eq. (16). Finally, confidence intervals can be constructed with those variances. Under normality assumptions one can, following (Sen, 1967), deduce the following confidence interval:

$$\left[A - s_A \times z_{\frac{\alpha}{2}} \, , \, A + s_A \times z_{\frac{\alpha}{2}}\right], \tag{17}$$

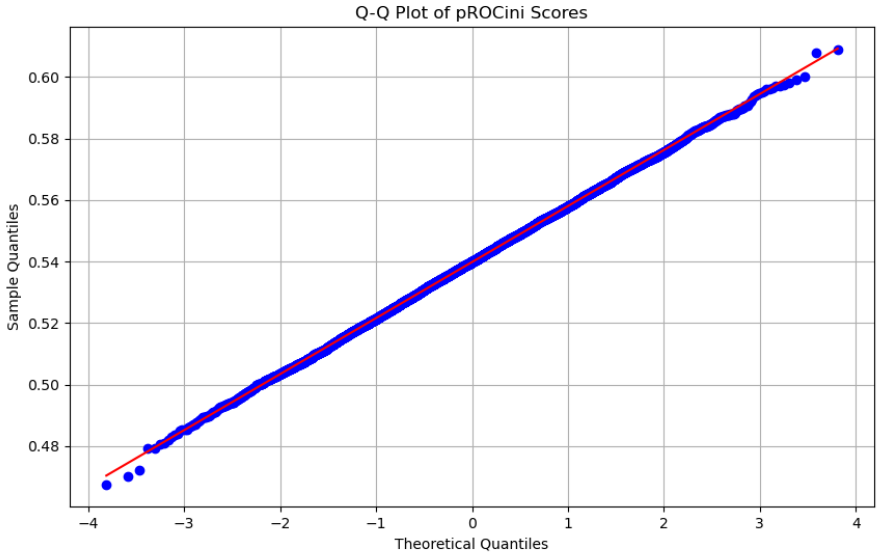
where $z_{\frac{\alpha}{2}}$ corresponds to the appropriate z-score for a $100(1-\alpha)\%$ confidence interval. This can be used to determine whether the difference between pROCini scores is statistically significant by checking whether one falls within the confidence interval of the other.

3.4 A comparison of the discriminative power of uplift scores

Before delving into an extensive simulation study, we first replicate the initial experiment conducted with the Qini score for our new pROCini metric. We perform a Shapiro–Wilk test (Shapiro and Wilk, 1965), obtaining $W = 0.9999$ and $p = 0.76$. This result implies that the normality hypothesis cannot be rejected for $\alpha = 0.05$.



(a) Density plot of the simulated distribution



(b) Q–Q plot of simulated vs. normal distribution

Figure 6: Results of Algorithm 1 with $r = 10\,000$, $N = 1\,000$, $(\alpha, \beta) = (12, 12)$, $v = 0.1$, $E = \{0.1\}$, $S = \{\text{QS}\}$.

Notably, although some pROCini scores fall below 0.5 (which corresponds to a random model), the proportion is substantially lower than the number of Qini scores falling below 0. Specifically, in our experiment, 591 (5.91%) of the Qini scores were below 0, whereas only

131 (1.31%) of the pROCini scores were below 0.5. This marked reduction in below-random performance is promising and may indicate superior performance of the pROCini metric.

To further study the ability of our new metrics to discern between different uplift models and to compare them to ROCini and pROCini, a more extensive simulation study is conducted. The performance in discerning the ground truth ranking (best model) with a noisy ranking (poor model) is compared for QS10 (the Qini score at 10%), TOCS, ROCiniS, pROCiniS and CROCS. We adapt Algorithm 1 to compare different metrics $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$. Furthermore, we set $E = \{0, 0.025, 0.05, 0.075, 0.1\}$ to check the extent to which the metrics can discriminate between the model with and without error. This is applied for models S_v with different signal strength i.e., $v \in \{0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4\}$. In this simulation study we vary the underlying distribution of positive outcomes in the control group, i.e., $(\alpha, \beta) \in \{(0.5, 0.5), (5, 15), (5, 25)\ , (15, 15), (25, 25), (25, 5), (15, 5)\}$ to obtain data about all shapes of beta distributions as presented in Figure 7. Furthermore the individual treatment effects as well as the errors are always considered to be normally distributed. This choice of (α, β) parameters yields varying proportions of binary treatment effect values equal to 0 or 1 across settings. An overview of the resulting distributions is provided in Table 4.

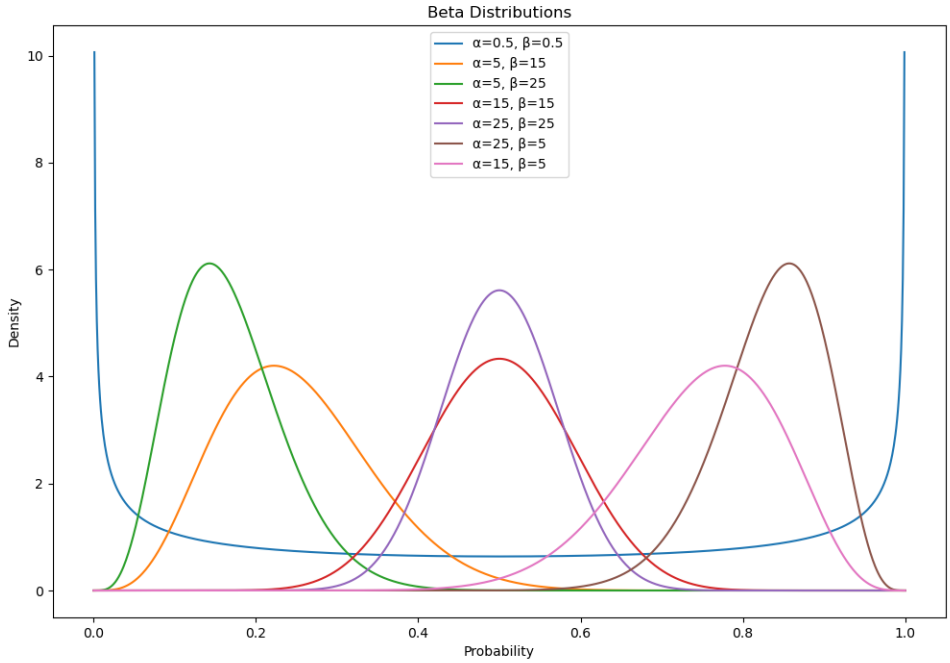
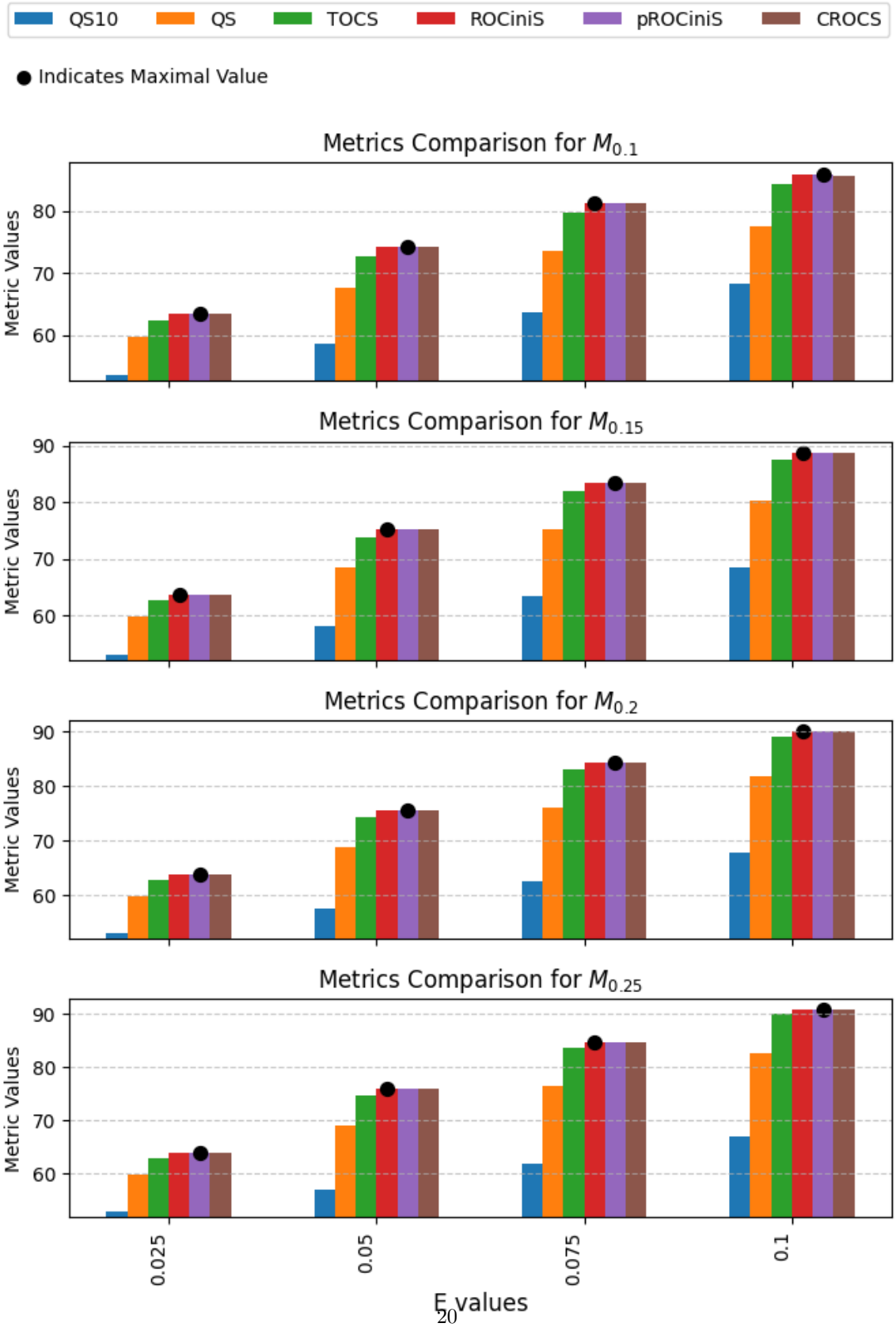


Figure 7: Various beta distributions used in the simulation

This procedure is repeated $r = 1000000$ times. For each metric, the scores of the perfect model are compared with the scores of the models with errors. We report the percentage of the runs in which the highest score is assigned to the perfect model. Finally, the results

are compared using the proportion Z-Test (Zou et al., 2003). Our null hypothesis is that the QS is a better metric, i.e. that it has a higher proportion. We indicated the highest proportion for each row in bold. The results presented in Figures 8 to 14 and Tables 5 to 11 in the Appendix demonstrate that our three proposed metrics ROCiniS, pROCiniS, and CROCS frequently yield statistically significant improvements on QS and TOCS. Overall, pROCiniS exhibits the best performance. However, an exception is observed in Figure 10, where $(\alpha, \beta) = (5, 25)$, representing a scenario in which the baseline probability (the probability of a positive outcome in the control group) is skewed to the right and generally low. We hypothesize that this may be attributed to the fact that such conditions render positive outcomes in both the treatment and control groups more rare and, consequently, more significant. This phenomenon is not as well captured by ROCiniS, pROCiniS, and CROCS, but plays a more prominent role in TOCS and QS, particularly when the signal strength v is greater.

1a.png


QS10
QS
TOCS
ROCiniS
pROCiniS
CROCS
Indicates Maximal Value
Metrics Comparison for $M_{0.1}$
Metrics Comparison for $M_{0.15}$
Metrics Comparison for $M_{0.2}$
Metrics Comparison for $M_{0.25}$
Metric Values
60
70
80
90
0.025
0.05
0.075
0.1
E values

1b.png

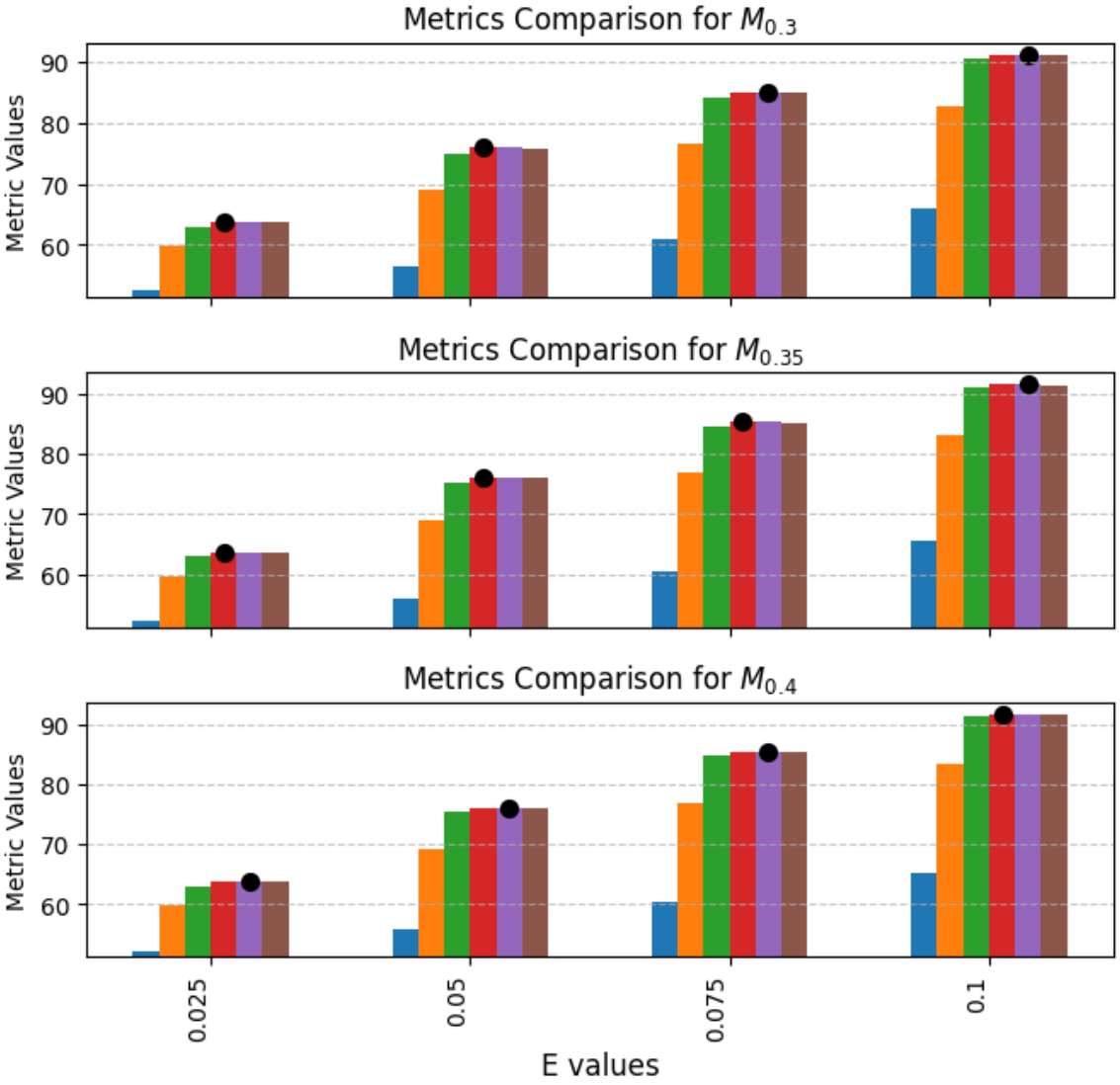
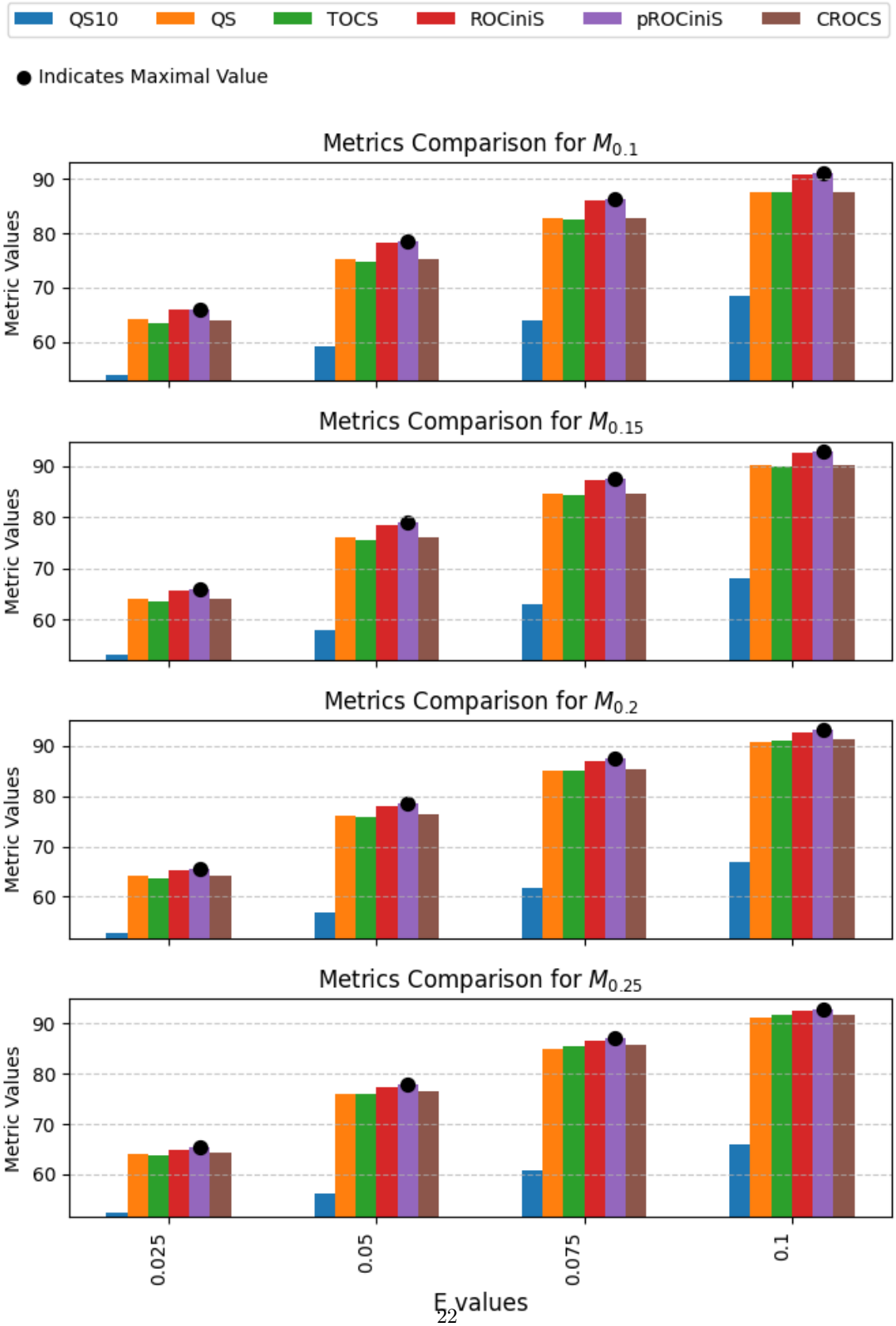


Figure 8: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (0.5, 0.5)$.

2a.png


QS10
QS
TOCS
ROCiniS
pROCiniS
CROCS
Indicates Maximal Value
Metrics Comparison for $M_{0.1}$
Metrics Comparison for $M_{0.15}$
Metrics Comparison for $M_{0.2}$
Metrics Comparison for $M_{0.25}$
Metric Values
90
80
70
60
0.025
0.05
0.075
0.1
E values

2b.png

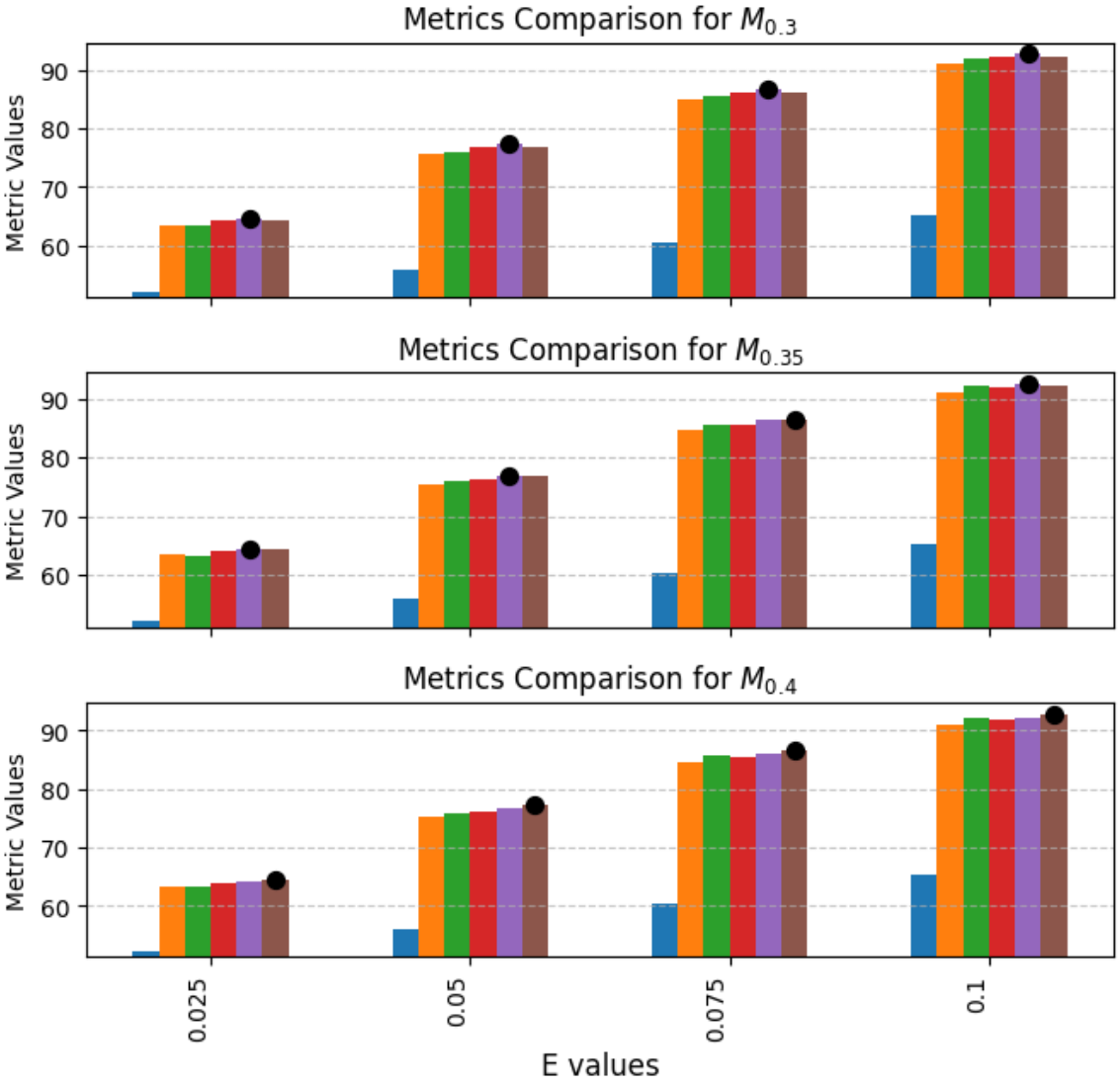
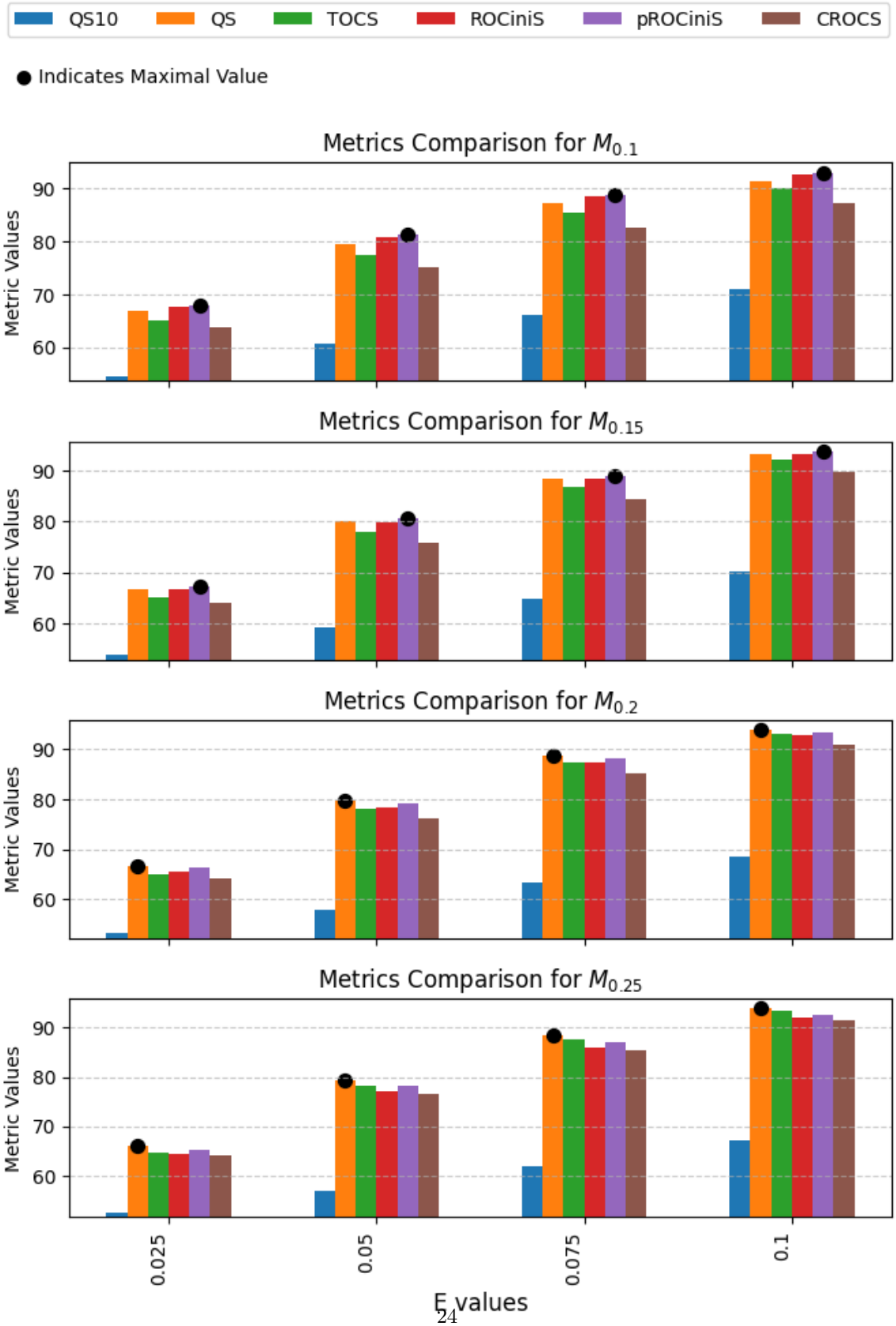


Figure 9: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (5, 15)$.

3a.png


QS10
QS
TOCS
ROCiniS
pROCiniS
CROCS
Indicates Maximal Value
Metrics Comparison for $M_{0.1}$
Metrics Comparison for $M_{0.15}$
Metrics Comparison for $M_{0.2}$
Metrics Comparison for $M_{0.25}$
Metric Values
90
80
70
60
0.025
0.05
0.075
0.1
E values

3b.png

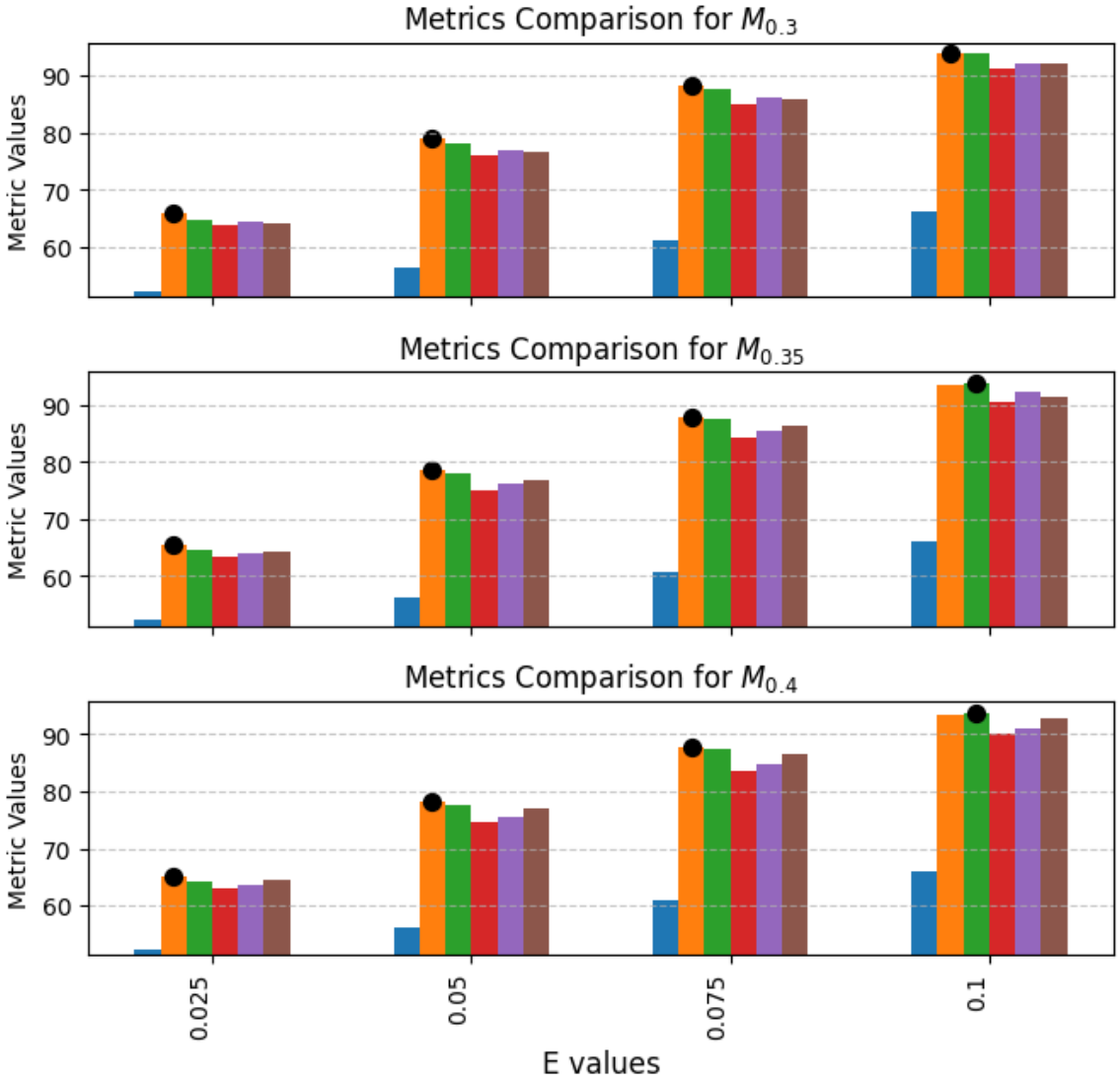
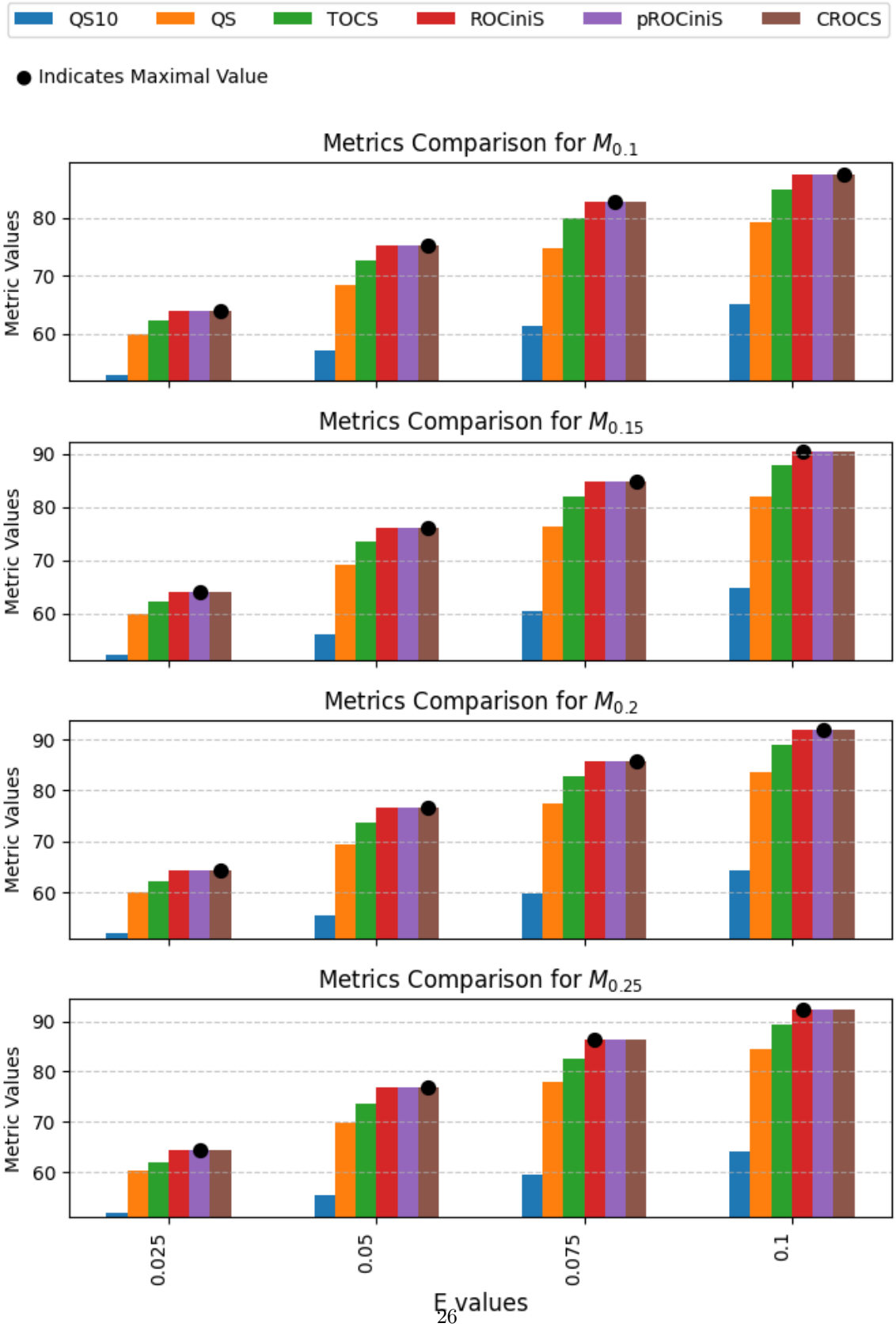


Figure 10: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (5, 25)$.

4a.png


QS10
QS
TOCS
ROCiniS
pROCiniS
CROCS
Indicates Maximal Value
Metrics Comparison for $M_{0.1}$
Metrics Comparison for $M_{0.15}$
Metrics Comparison for $M_{0.2}$
Metrics Comparison for $M_{0.25}$
Metric Values
60
70
80
90
0.025
0.05
0.075
0.1
E values

4b.png

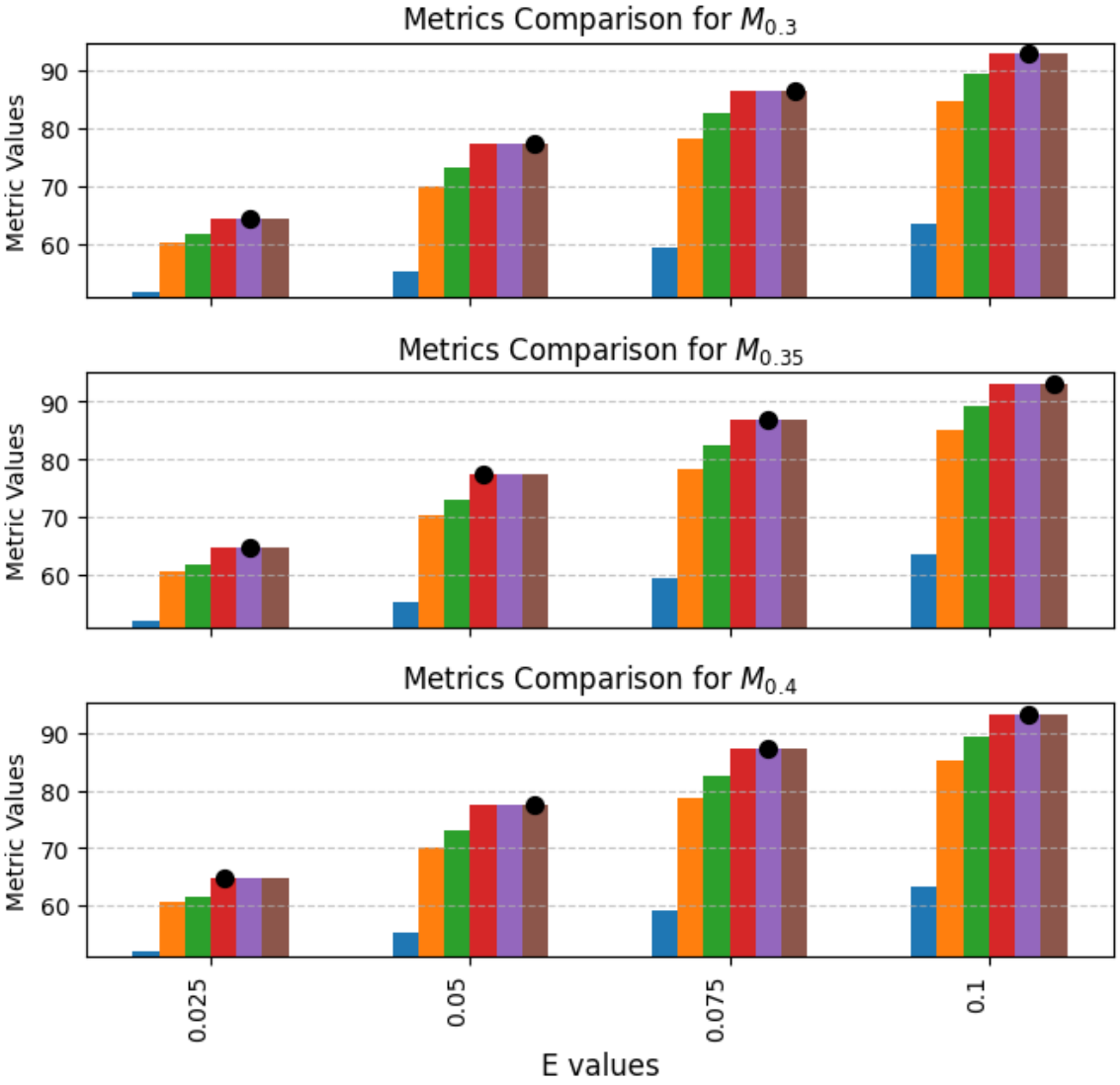
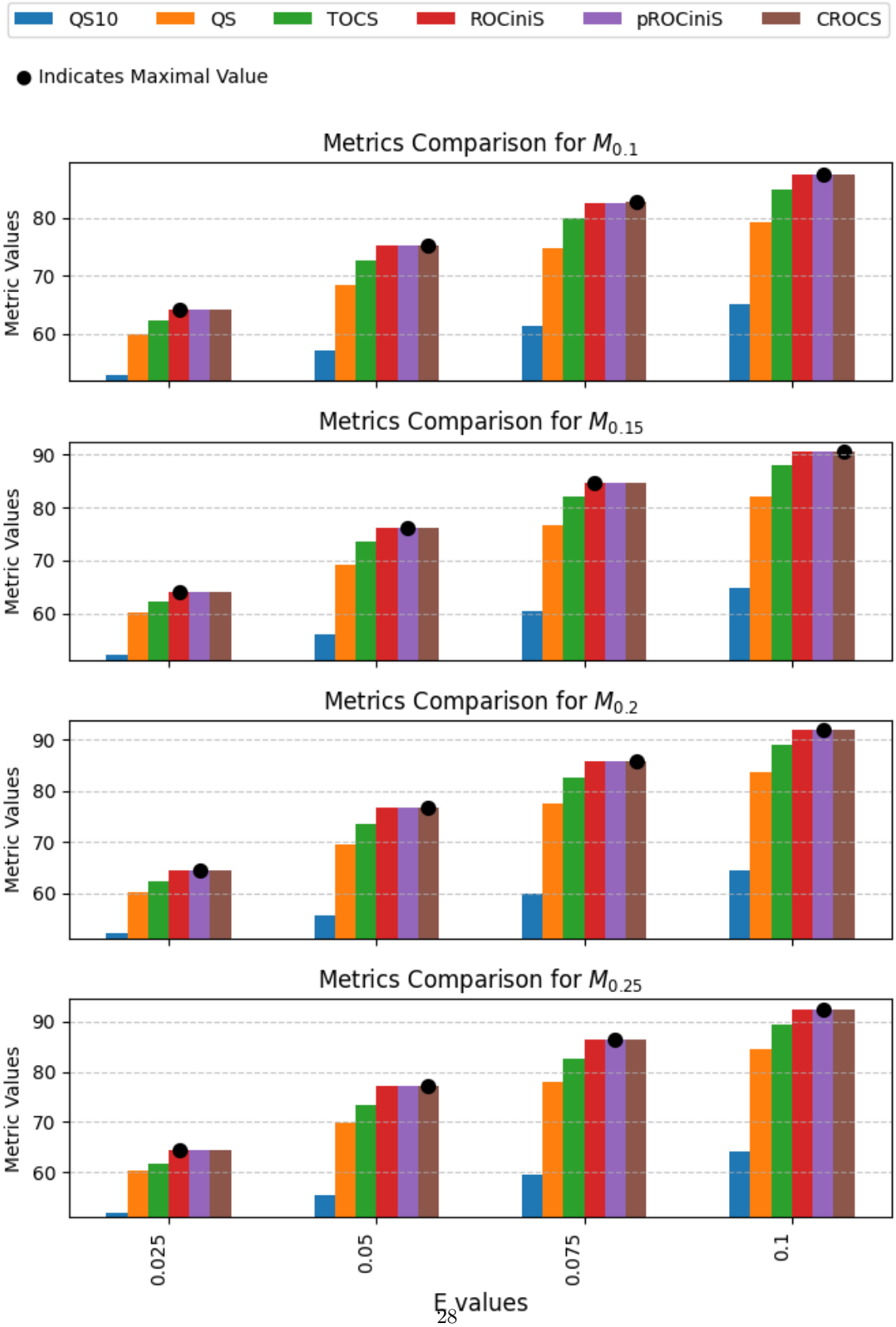


Figure 11: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (15, 15)$.

5a.png


QS10
QS
TOCS
ROCiniS
pROCiniS
CROCS
Indicates Maximal Value
Metrics Comparison for $M_{0.1}$
Metrics Comparison for $M_{0.15}$
Metrics Comparison for $M_{0.2}$
Metrics Comparison for $M_{0.25}$
Metric Values
60
70
80
90
0.025
0.05
0.075
0.1
E values

5b.png

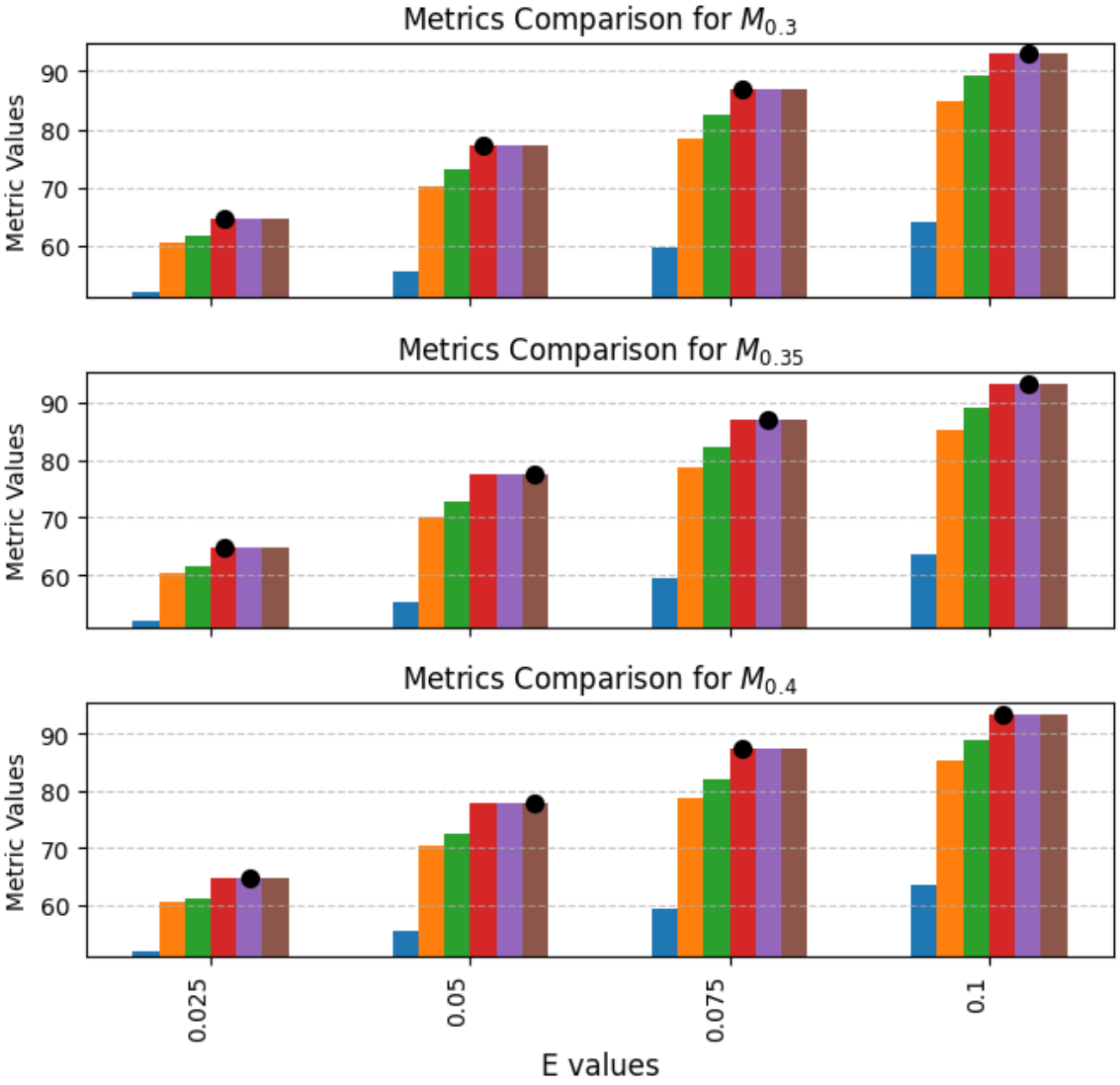
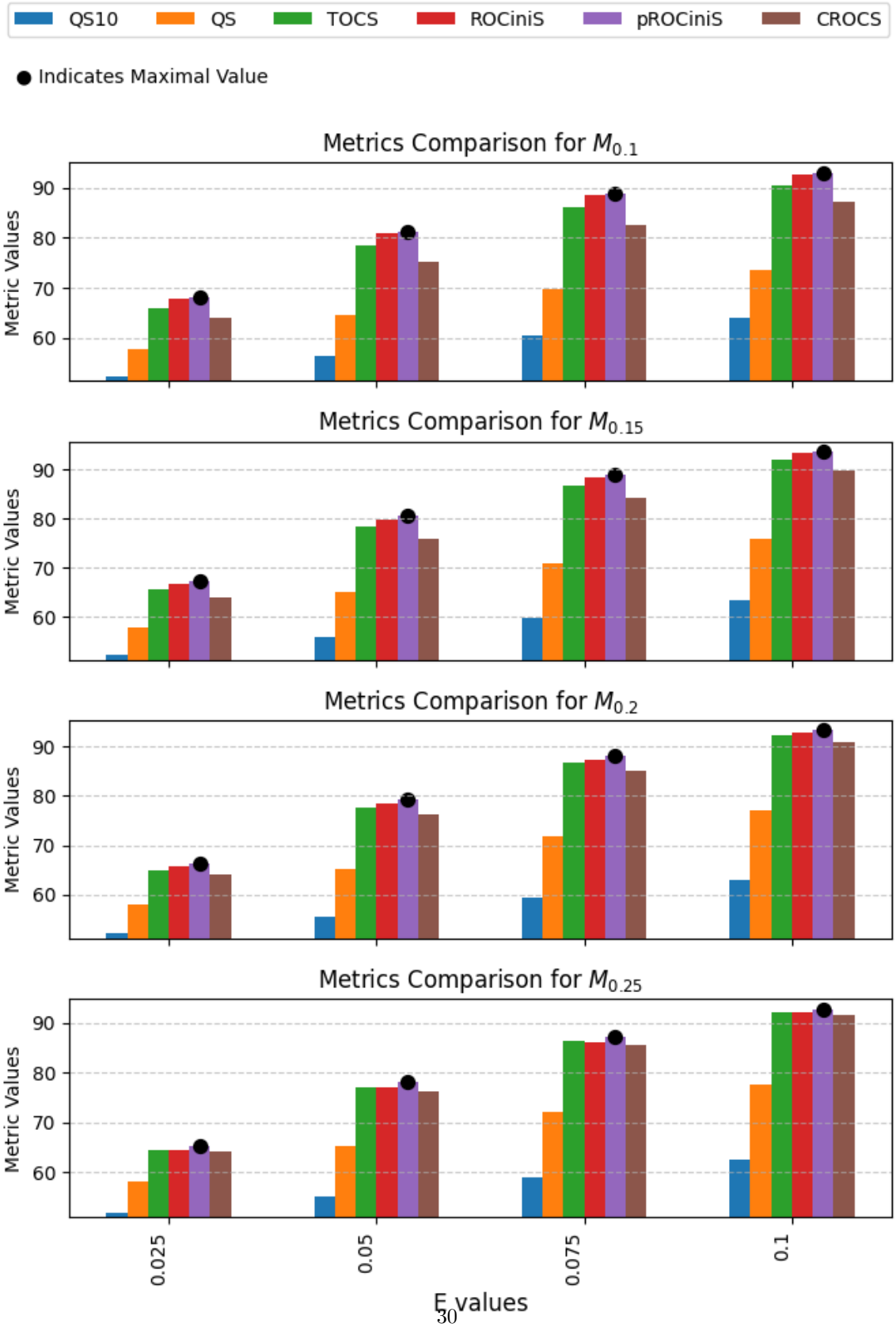


Figure 12: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (25, 25)$.

6a.png


QS10
QS
TOCS
ROCiniS
pROCiniS
CROCS
Indicates Maximal Value
Metrics Comparison for $M_{0.1}$
Metrics Comparison for $M_{0.15}$
Metrics Comparison for $M_{0.2}$
Metrics Comparison for $M_{0.25}$
Metric Values
90
80
70
60
0.025
0.05
0.075
0.1
E values

6b.png

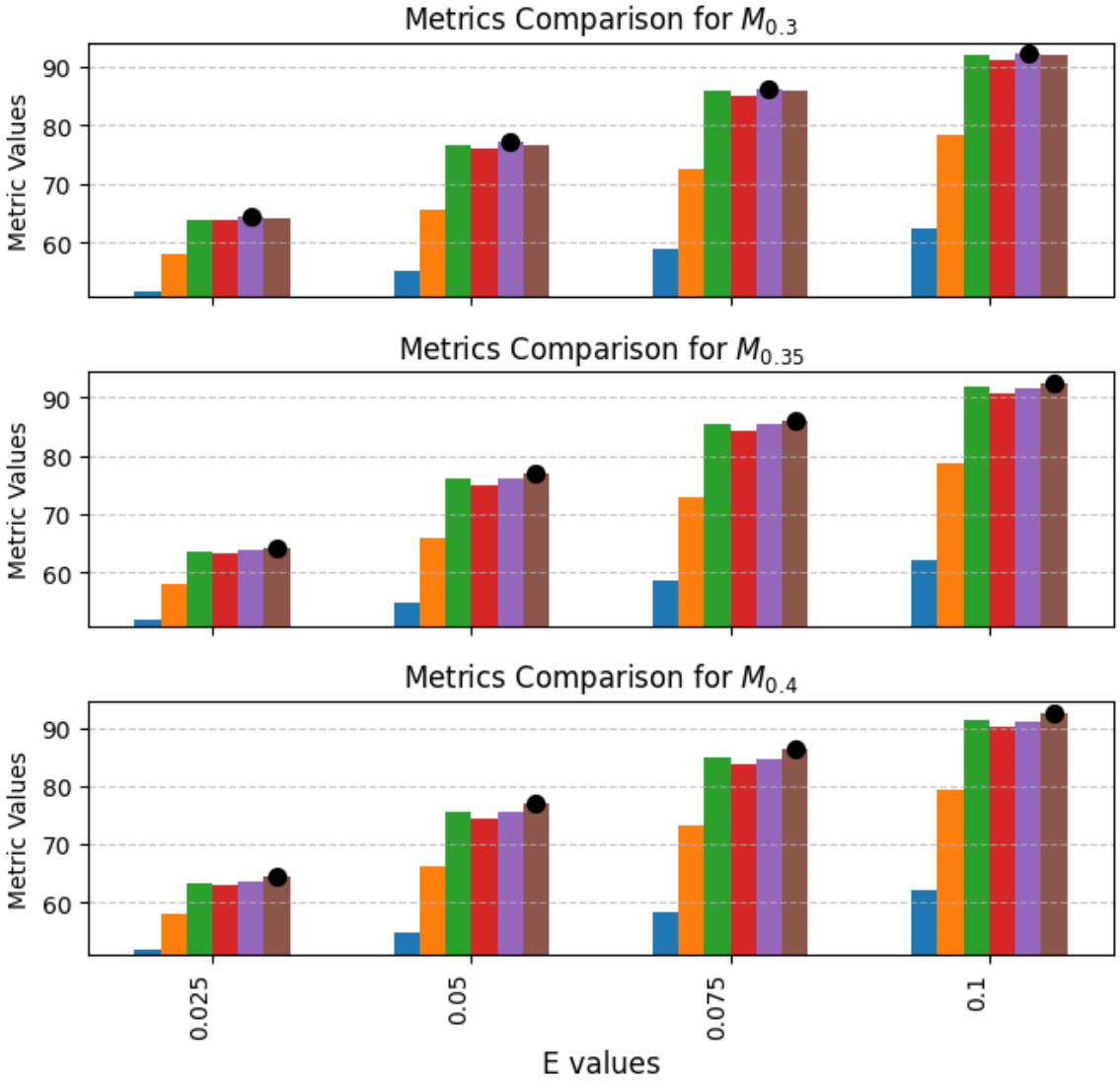
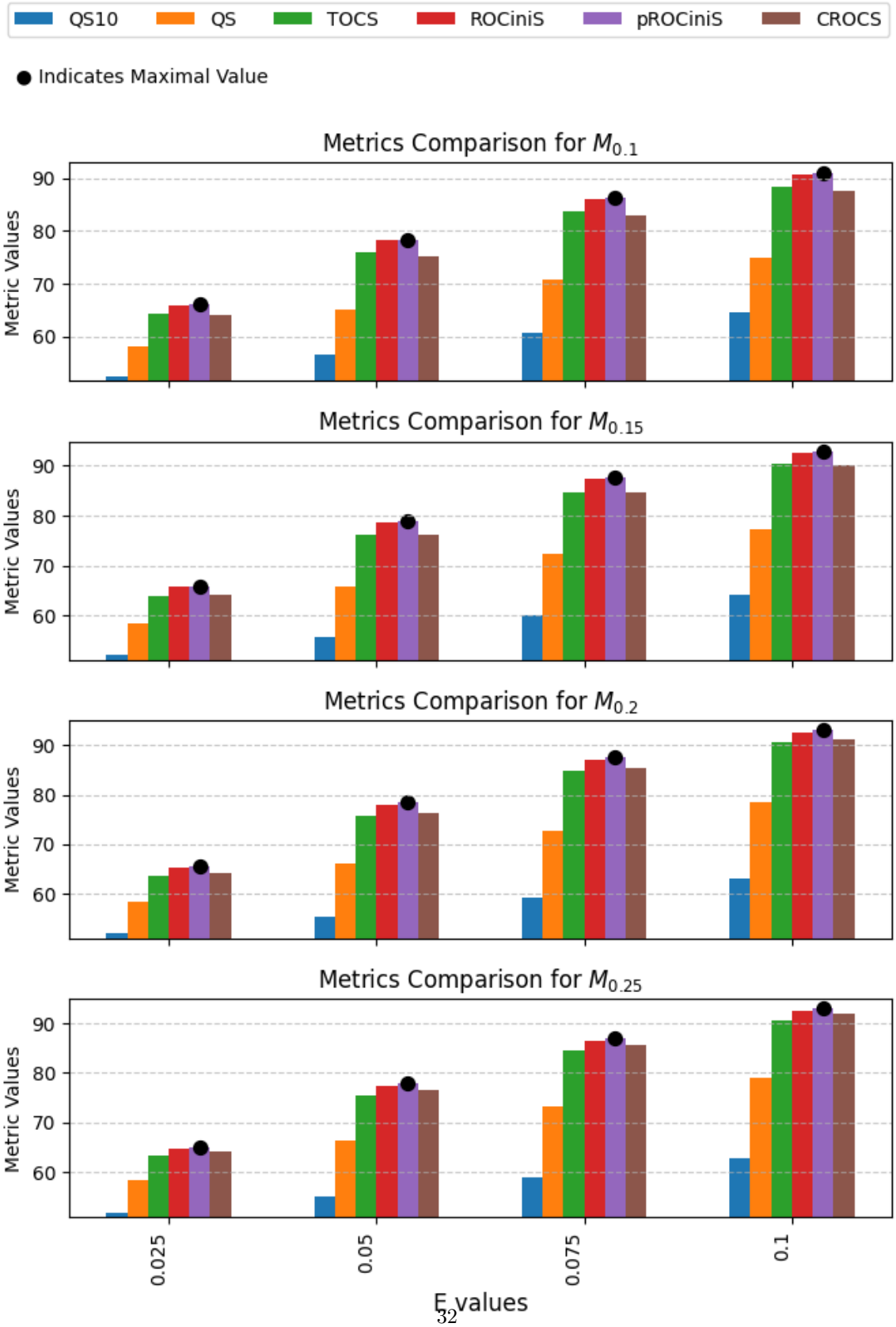


Figure 13: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (25, 5)$.

7a.png


QS10
QS
TOCS
ROCiniS
pROCiniS
CROCS
● Indicates Maximal Value
Metrics Comparison for $M_{0.1}$
Metrics Comparison for $M_{0.15}$
Metrics Comparison for $M_{0.2}$
Metrics Comparison for $M_{0.25}$
Metric Values
90
80
70
60
0.025
0.05
0.075
0.1
E values

7b.png

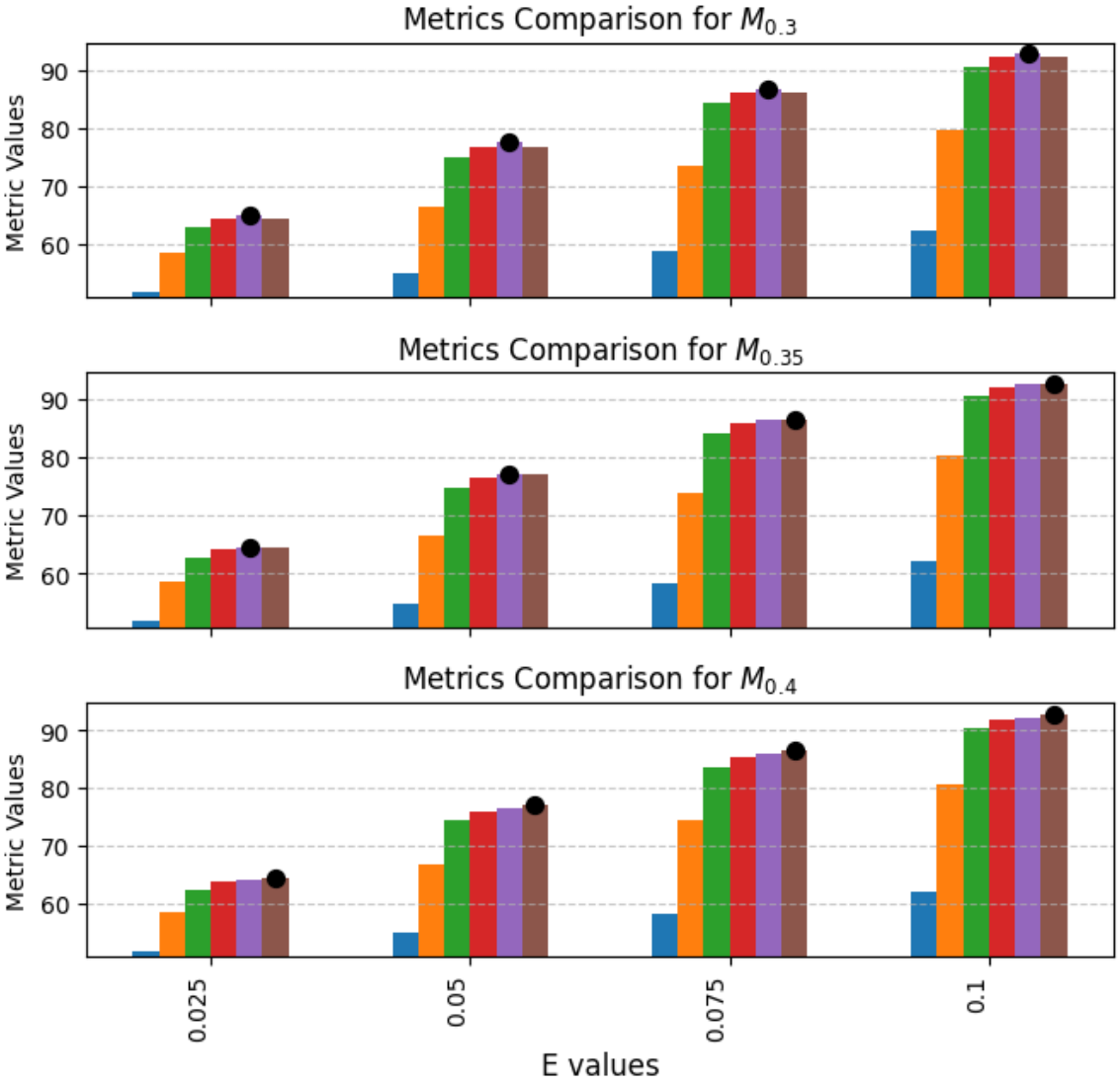


Figure 14: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (15, 5)$.

4. Does uplift metric choice matter for real data?

4.1 Empirical Evaluation: Real Data

In this subsection, we present the results on three commonly used uplift modelling benchmark data sets: the `Hillstrom` (Hillstrom, 2008), `Criteo` (Diemert, Eustache et al., 2018), and `Information` (Writer and Others, 2021) data sets. For each data set, we evaluate eight commonly used uplift models and strategy combinations using the `sklift` package. We consider four strategies: (i) the S-Learner (Künzel et al., 2019), (ii) the class transformation approach (Jaskowski and Jaroszewicz, 2012), (iii) the CATE-generating transformation of the outcome (Athey and Imbens, 2015) and (iv) the two-model approach (Betlei et al., 2018). Each of these strategies is run with both Logistic Regression and XGBoost. Finally, we report the ranks of the test set metric result for QS10, TOCS, ROCiniS, pROCiniS and CROCS.

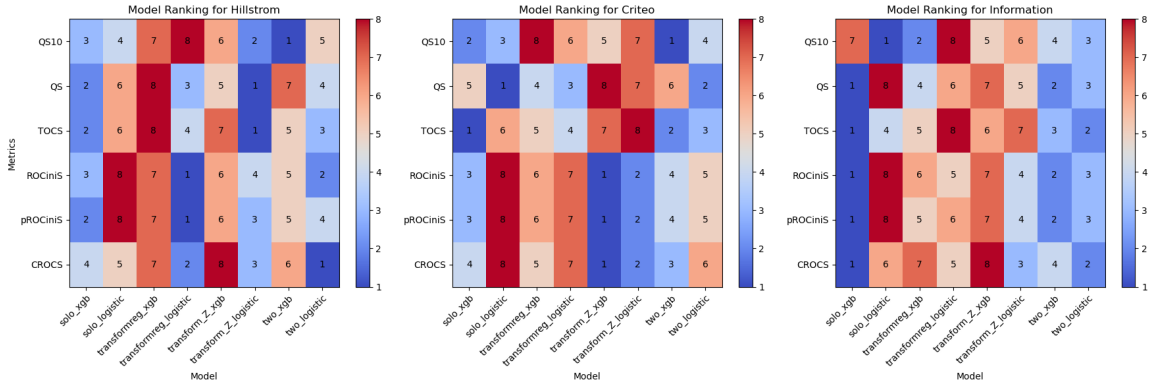


Figure 15: Heatmaps of the eight modelling strategies ranked by test set performance for the six metrics. Prefix *solo* refers to the S-learner strategy, *transformreg* to the class transformation approach, *transform_Z* to the CATE-generating transformation approach, and *two* to the two-model approach. The suffixes *xgb* and *logistic* indicate whether XGBoost or Logistic Regression was used, respectively.

Assuming the practitioner selects the final uplift model solely based on its generalization performance, Figure 8 demonstrates that the chosen uplift metric significantly impacts the final model selection. For the `Hillstrom` and `Criteo` data sets, four different strategies achieve the top rank across six metrics. In contrast, for the `Information` data set, the model selection is considerably more consistent across metrics, with only QS10 producing a different (*solo_logistic*) optimal model compared to the others (*solo_xgb*). Although no normative conclusions can be drawn from this experiment due to lack of access to the ground truth, it is clear that in realistic scenarios the choice of metric significantly impacts the model selection, and thus the effectiveness of the machine learning system.

4.2 Semi-Synthetic Evaluation

Having established that metric choice significantly affects model selection, and that our proposed metrics outperform existing ones in our simulated experiments, we now evaluate whether these advantages carry over in a semisynthetic setting based on real-world covariates

and treatment assignments. Specifically, we augment the original Hillstrom data set by generating synthetic outcomes via a logistic function, given by

$$p_i = \frac{1}{1 + \exp\left[-\left(\beta_0 + X_i^\top \beta + \beta_t T_i + \epsilon_i\right)\right]},$$

where X_i represents the original (standardized) features for observation i, T_i represents the treatment indicator, β_t represents the average treatment effect parameter, and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ introduces random variation. From this specification, the true individual treatment effect for each instance i is explicitly defined as the difference between probabilities under treatment and control scenarios, namely:

$$\text{ITE}_i = p_i(T_i = 1) - p_i(T_i = 0).$$

In line with the literature, the original covariates and treatment assignments are preserved to reflect a realistic multivariate structure, while simulating outcomes via a logistic function with parameters $\beta_0 = 0.0$, $\beta = \mathbf{1}$, $\beta_t = 0.5$, and $\sigma = 0.1$ (Marchese et al., 2025; Hill, 2011; Alaa and Van Der Schaar, 2017). This setup yields nonlinear treatment response behaviour and allows full control over the treatment effect strength and noise. The known ground truth ITEs enable direct and interpretable evaluations of model and metric performances under realistic but controlled conditions.

We trained four uplift models based on standard S-learner and T-learner strategies Künzel et al. (2019); Curth and Van der Schaar (2021), each implemented with two widely used base learners: Logistic Regression and XGBoost. The S-learners model the outcome using a single model with treatment included as an additional feature, whereas the T-learners fit separate models to treated and control groups, estimating the uplift as the difference in the predicted probabilities. For each learner, uplift predictions were obtained by computing the difference in the predicted probabilities under the treatment and control scenarios.

We applied these models to a population of 1,000 observations drawn from the Hillstrom data set. Using the known ground truth ITEs from our simulation setup, we evaluated each model's ability to recover the true treatment effects. This was done using several comparison metrics: mean squared error (MSE), Spearman's rank correlation, Kendall's τ, a custom weighted Kendall distance, Earth Mover's Distance (EMD), and mean rank distance (MRD) between estimated and true ITEs. These metrics provide complementary views of performance in both absolute error and ranking alignment.

The weighted Kendall distance penalizes pairwise ranking disagreements more heavily when the underlying true ITEs differ substantially. Formally, we define it as

$$\text{WeightedKendall}(x, y) = 1 - \frac{\sum_{i<j} w_{ij}\, \mathbb{I}[\text{sign}(x_i - x_j) \neq \text{sign}(y_i - y_j)]}{\sum_{i<j} w_{ij}}, \quad w_{ij} = |y_i - y_j|.$$

Model	MSE	Spearman's ρ	Kendall's τ	WeightedKendall	EMD	MRD
S_LR	0.0030	0.9902	0.9214	0.9960	0.0377	8.7200
T_LR	0.0098	0.6102	0.4586	0.8067	0.0541	56.3800
S_XGB	0.0253	0.4630	0.3583	0.7292	0.0766	63.7600
T_XGB	0.0608	0.2252	0.1574	0.6075	0.1251	85.1900

Table 2: Evaluation of the four uplift models based on their proximity to the ground truth ITEs across multiple metrics.

Table 2 provides a quantitative assessment of how closely each of the four trained models approximates the ground truth ITEs in the semisynthetic setting. Across all metrics—both value-based (MSE, EMD) and rank-based (Spearman, Kendall, Mean Rank Distance)—the S-Learner with Logistic Regression (S_LR) clearly outperforms the others. T_LR ranks second, followed by S_XGB, with T_XGB performing worst. These results allow us to construct an empirical ranking of model performance on this data set, which will serve as a reference for evaluating the behaviour of different uplift metrics.
Given the known ground truth ITEs and the performance of the trained models (Table 2), one would expect that a good uplift metric ranks the models in the following order: `true_ITE` > S_LR > T_LR > S_XGB > T_XGB. To evaluate this, we use a slight adaptation of Algorithm 1 to simulate 1000000 runs and check if the metrics $S = \{$QS10, TOCS, ROCiniS, pROCiniS, CROCS$\}$ provide the same ordering of the models.

For each simulation run, we record whether the expected rank order holds across all adjacent pairs. For example, whether the metric ranks `true_ITE` above S_LR, S_LR above T_LR, etc. After 1,000,000 runs of the algorithm, the proportion of times each inequality is satisfied is shown below.

Metric	`true_ITE` > S_LR	S_LR > T_LR	T_LR > S_XGB	S_XGB > T_XGB
QS10	50.1830	50.4570	52.2646	50.8006
QS	50.9785	58.8713	54.2210	55.4796
TOCS	50.7734	57.5253	54.6594	56.7970
ROCiniS	51.3898	62.3618	55.4541	57.6529
pROCiniS	51.4539	62.5246	55.4111	57.6352
CROCS	51.4615	62.6598	55.2145	57.7181

Table 3: Proportion of simulation runs (out of 1,000,000) where each metric correctly ranks adjacent model pairs according to ground truth quality.

Using a two-proportion Z-test at a 99% confidence level (i.e., $\alpha = 0.01$), we find that pROCiniS significantly outperforms QS10, QS, and TOCS in correctly ranking the models.These findings are consistent with our synthetic experiments, in which similar conclusions were reached. Moreover, the relatively low discriminatory power observed is to be expected given the small differences between the ITE estimates and the significant overlap in their score distributions when sampling. This subtle separation among the models underscores that even modest improvements in metric performance can lead to markedly different model selections in practice.

5. Conclusions

In this article, an in-depth analysis of uplift modelling evaluation was presented. First, the distributional properties of the classic Qini score were examined in a simulation experiment. The simulation results show how the treatment effect size and population size are key factors in the stability of uplift modelling evaluation. Second, we introduced ROCiniS and pROCiniS, two new metrics with more attractive mathematical properties. Additionally, simulations show that ROCiniS and pROCiniS significantly outperform current metrics in discerning between good and bad uplift models. This conclusion is further supported by a semisynthetic experiment based on real-world covariates and treatment assignments, where pROCiniS and ROCiniS again demonstrated superior ability to recover the correct model ranking. These results confirm that the advantages of the proposed metrics extend beyond controlled simulations and hold in more realistic, data-driven settings. Finally, the close relationship between the pROCini curve and the ROC curve is demonstrated, and confidence bounds for the pROCiniS are derived using theory regarding ODGs.

We see multiple avenues for future work. From a practical perspective, further investigation of the relationship between data set characteristics and the stability of uplift evaluation will help practitioners make more informed decisions about uplift For practitioners, we suggest prioritizing the use of ROCiniS and pROCiniS, as they generally demonstrate superior performance compared with other metrics, with the notable exception being when the probability of a positive outcome in the control group is low, as was illustrated in Table 4. To further refine metric selection, practitioners can conduct semisynthetic simulations to determine the most suitable metric for their specific problem. If this approach is not feasible, the use of multiple metrics to assess sensitivity, as shown in Figure 8, is advisable. From an academic perspective, the ROCini and pROCini not only integrate seamlessly with existing extensions of the Qini curve (e.g., group weighting as in (Gutierrez and Gérardy, 2017)) but also provide markedly improved model discrimination, offering a more sensitive tool for evaluating subtle treatment effect differences. Research directions investigating such extensions are thus also compatible with these new metrics. Building on the ODG theory, a compelling direction for future research is exploring the application of concepts akin to the ROC convex hull (ROCCH) (Provost and Fawcett, 2001; Fawcett and Niculescu-Mizil, 2007) in uplift modelling. This approach could facilitate the creation of composite models that harness the strengths of various uplift models across different population segments, potentially enhancing overall performance.

Finally, we believe that this strand of research, and more specifically the study of confidence bounds, will lead to additional insight into the suitability and robustness of uplift models and their evaluation, warding practitioners from misguided confidence based on small data sets with small treatment effects. Moreover, the derived confidence bounds of the pROCiniS are a natural starting point for theoretically analysing the sometimes contradicting results concerning the metrics as well.

Finally, we assert that the ongoing investigation into uplift model evaluation—and specifically, the exploration of confidence bounds—will further elucidate the reliability and efficacy of these models, protecting practitioners from erroneous conclusions in settings with limited data and minimal treatment effects. The cumulative benefits of the ODG-based metrics such as, superior performance, interpretability, a general framework that can be readily

adapted to cost-sensitive settings with applicable weighting strategies, and the inherent ability to facilitate evaluation within subpopulations—position them as leading contenders for evaluating uplift models. These advantages underpin a more nuanced and accurate model selection process, particularly under conditions of subtle treatment effects and significant outcome distribution overlap, thereby reducing the risk of misinterpretation inherent in traditional metrics. Additionally, the confidence bounds established for pROCiniS provide a robust foundation for theoretical analyses of the occasionally discordant outcomes observed with alternative metrics.

6. Appendix

	Value = 0		Value = 1	
(α, β)	PC+IU	PC+IU+IUn	PC+IU	PC+IU+IUn
(0.5, 0.5)	8-19	5-19	8-19	5-19
(5, 15)	2-28	2-28	0-4	0-4
(5, 25)	7-35	5-35	0-3	0-3
(15, 15)	0-12	0-12	0-12	0-12
(25, 25)	0-11	0-11	0-12	0-11
(25, 5)	0-3	0-2	7-35	5-35
(15, 5)	0-4	0-4	2-28	2-28

Table 4: Minima–maxima ranges in percentages for PC+IU and PC+IU+IUn by (α, β).

The reported values are percentage ranges for $PC + IU$ with or without error being equal to 0 or 1, across increasing signal strengths $v \in 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4$. As v increases, the upper bounds of these percentages rise approximately linearly, reflecting increasing signal clarity. No consistent relationship was observed between the proportion of 0s and 1s and the comparative performance of the metrics under evaluation.

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.1}$						
0.025	53.647	59.7349	62.4481	63.5632	**63.5652**	63.5636
	-87 (≈ 1)		39 ($< 10^{-300}$)	56 ($< 10^{-600}$)	56 ($< 10^{-600}$)	56 ($< 10^{-600}$)
0.05	58.7382	67.7027	72.6621	74.2425	**74.2453**	74.2212
	-131 (≈ 1)		77 ($< 10^{-1200}$)	102 ($< 10^{-2200}$)	102 ($< 10^{-2200}$)	102 ($< 10^{-2200}$)
0.075	63.7874	73.5278	79.6864	**81.2553**	81.2516	81.2229
	-148 (≈ 1)		103 ($< 10^{-2300}$)	131 ($< 10^{-3700}$)	131 ($< 10^{-3700}$)	130 ($< 10^{-3600}$)
0.1	68.2355	77.4588	84.2856	85.7392	**85.7415**	85.6254
	-147 (≈ 1)		123 ($< 10^{-3200}$)	151 ($< 10^{-4900}$)	151 ($< 10^{-4900}$)	149 ($< 10^{-4800}$)
$M_{0.15}$						
0.025	53.2369	59.8295	62.6771	**63.7439**	63.7375	63.7411
	-94 (≈ 1)		41 ($< 10^{-300}$)	57 ($< 10^{-700}$)	57 ($< 10^{-700}$)	57 ($< 10^{-700}$)
0.05	58.1256	68.4894	73.7544	**75.2632**	75.2625	75.2382
	-152 (≈ 1)		82 ($< 10^{-1400}$)	107 ($< 10^{-2400}$)	107 ($< 10^{-2400}$)	106 ($< 10^{-2400}$)
0.075	63.3601	75.1842	81.9048	83.36	**83.362**	83.3246
	-181 (≈ 1)		116 ($< 10^{-2900}$)	143 ($< 10^{-4400}$)	143 ($< 10^{-4400}$)	142 ($< 10^{-4300}$)
0.1	68.5267	80.3508	87.4481	**88.7365**	88.7333	88.6571
	-192 (≈ 1)		137 ($< 10^{-4000}$)	164 ($< 10^{-5800}$)	164 ($< 10^{-5800}$)	162 ($< 10^{-5700}$)
$M_{0.2}$						
0.025	52.9882	59.8008	62.7826	63.7454	**63.7486**	63.7276
	-97 (≈ 1)		43 ($< 10^{-400}$)	57 ($< 10^{-700}$)	57 ($< 10^{-700}$)	57 ($< 10^{-700}$)
0.05	57.4912	68.7379	74.2505	75.4966	**75.5023**	75.4904
	-165 (≈ 1)		86 ($< 10^{-1600}$)	107 ($< 10^{-2400}$)	107 ($< 10^{-2400}$)	106 ($< 10^{-2400}$)
0.075	62.6068	76.0804	83.0418	84.2544	**84.2618**	84.2013
	-207 (≈ 1)		122 ($< 10^{-3200}$)	145 ($< 10^{-4500}$)	145 ($< 10^{-4500}$)	144 ($< 10^{-4500}$)
0.1	67.771	81.6718	88.9778	**90.0404**	90.0338	89.9684
	-226 (≈ 1)		146 ($< 10^{-4600}$)	170 ($< 10^{-6200}$)	170 ($< 10^{-6200}$)	168 ($< 10^{-6100}$)
$M_{0.25}$						
0.025	52.7339	59.7888	62.9522	63.8049	**63.8122**	63.8003
	-101 (≈ 1)		46 ($< 10^{-400}$)	58 ($< 10^{-700}$)	59 ($< 10^{-700}$)	58 ($< 10^{-700}$)
0.05	56.9696	68.9146	74.7297	**75.8157**	75.8136	75.8102
	-175 (≈ 1)		91 ($< 10^{-1800}$)	109 ($< 10^{-2500}$)	109 ($< 10^{-2500}$)	109 ($< 10^{-2500}$)
0.075	61.7889	76.3541	83.6949	**84.7058**	84.6959	84.6831
	-223 (≈ 1)		130 ($< 10^{-3600}$)	149 ($< 10^{-4800}$)	149 ($< 10^{-4800}$)	149 ($< 10^{-4800}$)
0.1	66.9088	82.4464	89.9539	90.7454	**90.7478**	90.6878
	-253 (≈ 1)		154 ($< 10^{-5100}$)	172 ($< 10^{-6400}$)	172 ($< 10^{-6400}$)	171 ($< 10^{-6300}$)

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.3}$						
0.025	52.4578	59.7808	62.9474	**63.882**	63.8796	63.8601
	-104 (≈ 1)		46 ($< 10^{-400}$)	60 ($< 10^{-700}$)	60 ($< 10^{-700}$)	59 ($< 10^{-700}$)
0.05	56.433	68.9922	74.9369	**75.9053**	75.9001	75.8718
	-184 (≈ 1)		94 ($< 10^{-1900}$)	109 ($< 10^{-2600}$)	109 ($< 10^{-2500}$)	109 ($< 10^{-2500}$)
0.075	61.0471	76.641	84.1398	84.9662	**84.9669**	84.9395
	-238 (≈ 1)		134 ($< 10^{-3800}$)	149 ($< 10^{-4800}$)	149 ($< 10^{-4800}$)	149 ($< 10^{-4800}$)
0.1	66.1136	82.8527	90.5552	91.1575	**91.1605**	91.1097
	-272 (≈ 1)		160 ($< 10^{-5500}$)	175 ($< 10^{-6600}$)	175 ($< 10^{-6600}$)	174 ($< 10^{-6500}$)
$M_{0.35}$						
0.025	52.3138	59.8153	62.9519	**63.7653**	63.7544	63.7489
	-107 (≈ 1)		46 ($< 10^{-400}$)	57 ($< 10^{-700}$)	57 ($< 10^{-700}$)	57 ($< 10^{-700}$)
0.05	56.1442	68.9729	75.2133	**76.0185**	76.0117	75.9776
	-187 (≈ 1)		98 ($< 10^{-2100}$)	112 ($< 10^{-2700}$)	111 ($< 10^{-2700}$)	111 ($< 10^{-2600}$)
0.075	60.6303	76.885	84.5985	**85.2346**	85.2329	85.1986
	-248 (≈ 1)		138 ($< 10^{-4100}$)	151 ($< 10^{-4900}$)	151 ($< 10^{-4900}$)	150 ($< 10^{-4800}$)
0.1	65.5189	83.1919	91.0932	91.4854	**91.4859**	91.4294
	-286 (≈ 1)		167 ($< 10^{-6000}$)	176 ($< 10^{-6700}$)	176 ($< 10^{-6700}$)	175 ($< 10^{-6600}$)
$M_{0.4}$						
0.025	52.1985	59.8509	63.0049	63.8321	**63.8327**	63.8155
	-109 (≈ 1)		46 ($< 10^{-400}$)	58 ($< 10^{-700}$)	58 ($< 10^{-700}$)	58 ($< 10^{-700}$)
0.05	55.9193	69.0759	75.3562	76.1277	**76.1312**	76.1257
	-192 (≈ 1)		99 ($< 10^{-2100}$)	112 ($< 10^{-2700}$)	112 ($< 10^{-2700}$)	112 ($< 10^{-2700}$)
0.075	60.4625	76.9621	84.8808	85.3634	**85.3649**	85.3589
	-252 (≈ 1)		143 ($< 10^{-4400}$)	152 ($< 10^{-5000}$)	152 ($< 10^{-5000}$)	152 ($< 10^{-5000}$)
0.1	65.1462	83.3762	91.3428	**91.6794**	91.678	91.6396
	-295 (≈ 1)		170 ($< 10^{-6200}$)	178 ($< 10^{-6800}$)	178 ($< 10^{-6800}$)	177 ($< 10^{-6700}$)

Table 5: Performance table of $S = \{\text{QS10, TOCS, ROCiniS, pROCiniS, CROCS}\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (0.5, 0.5)$.

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.1}$						
0.025	53.8211	64.0946	63.3763	65.978	**66.07**	64.0183
	-148 (≈ 1)		-11 (≈ 1)	28 ($< 10^{-172}$)	29 ($< 10^{-189}$)	-1 (0.87)
0.05	59.09	75.1947	74.633	78.2571	**78.4598**	75.2533
	-242 (≈ 1)		-9 (≈ 1)	51 ($< 10^{-573}$)	55 ($< 10^{-653}$)	1 (0.169)
0.075	64.0875	82.8445	82.5058	86.1065	**86.338**	82.8352
	-300 (≈ 1)		-6 (≈ 1)	64 ($< 10^{-884}$)	68 ($< 10^{-1019}$)	0 (0.569)
0.1	68.4052	87.6221	87.5104	90.7495	**90.965**	87.654
	-328 (≈ 1)		-2 (0.992)	71 ($< 10^{-1104}$)	76 ($< 10^{-1272}$)	1 (0.247)
$M_{0.15}$						
0.025	53.1371	64.167	63.4768	65.7732	**65.9405**	64.1592
	-158 (≈ 1)		-10 (≈ 1)	24 ($< 10^{-125}$)	26 ($< 10^{-153}$)	0 (0.546)
0.05	58.0214	75.9331	75.3728	78.4999	**78.8467**	76.0185
	-269 (≈ 1)		-9 (≈ 1)	43 ($< 10^{-409}$)	49 ($< 10^{-529}$)	1 (0.0788)
0.075	63.0364	84.5113	84.1623	87.2332	**87.5807**	84.5844
	-345 (≈ 1)		-7 (≈ 1)	55 ($< 10^{-666}$)	63 ($< 10^{-855}$)	1 (0.0763)
0.1	67.9334	90.0471	89.9581	92.5017	**92.7778**	90.1667
	-384 (≈ 1)		-2 (0.982)	62 ($< 10^{-824}$)	69 ($< 10^{-1034}$)	3 (0.00231)
$M_{0.2}$						
0.025	52.6934	64.0987	63.5498	65.3009	**65.6097**	64.2038
	-164 (≈ 1)		-8 (≈ 1)	18 ($< 10^{-71}$)	22 ($< 10^{-111}$)	2 (0.0606)
0.05	56.8931	76.0443	75.699	77.9732	**78.4748**	76.2991
	-287 (≈ 1)		-6 (≈ 1)	32 ($< 10^{-231}$)	41 ($< 10^{-368}$)	4 ($< 10^{-5}$)
0.075	61.8672	84.9745	85.0302	87.0533	**87.5548**	85.3273
	-370 (≈ 1)		1 (0.135)	42 ($< 10^{-393}$)	53 ($< 10^{-613}$)	7 ($< 10^{-12}$)
0.1	66.7861	90.7789	91.1332	92.7268	**93.0924**	91.2954
	-415 (≈ 1)		9 ($< 10^{-18}$)	50 ($< 10^{-547}$)	60 ($< 10^{-786}$)	13 ($< 10^{-38}$)
$M_{0.25}$						
0.025	52.4278	63.9736	63.6484	64.8785	**65.2428**	64.3327
	-166 (≈ 1)		-5 (≈ 1)	13 ($< 10^{-41}$)	19 ($< 10^{-79}$)	5 ($< 10^{-8}$)
0.05	56.2106	75.9096	75.9982	77.35	**77.9664**	76.6207
	-294 (≈ 1)		1 (0.0713)	24 ($< 10^{-128}$)	35 ($< 10^{-261}$)	12 ($< 10^{-32}$)
0.075	60.8725	84.9628	85.4816	86.5377	**87.1228**	85.6866
	-383 (≈ 1)		10 ($< 10^{-25}$)	32 ($< 10^{-223}$)	44 ($< 10^{-424}$)	14 ($< 10^{-47}$)
0.1	65.8882	91.0741	91.8048	92.4699	**92.9266**	91.8537
	-433 (≈ 1)		18 ($< 10^{-76}$)	36 ($< 10^{-283}$)	48 ($< 10^{-509}$)	20 ($< 10^{-87}$)

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.3}$						
0.025	52.1724	63.5554	63.3864	64.334	**64.703**	64.3908
	-163 (≈ 1)		-2 (0.993)	11 ($< 10^{-31}$)	17 ($< 10^{-64}$)	12 ($< 10^{-35}$)
0.05	55.9349	75.6705	76.0407	76.7898	**77.4185**	76.7592
	-294 (≈ 1)		6 ($< 10^{-10}$)	19 ($< 10^{-77}$)	29 ($< 10^{-187}$)	18 ($< 10^{-73}$)
0.075	60.4248	84.8501	85.6196	86.029	**86.6544**	86.0673
	-387 (≈ 1)		15 ($< 10^{-53}$)	24 ($< 10^{-124}$)	37 ($< 10^{-292}$)	24 ($< 10^{-132}$)
0.1	65.314	91.0453	92.0911	92.2174	**92.7176**	92.2753
	-441 (≈ 1)		27 ($< 10^{-156}$)	30 ($< 10^{-197}$)	43 ($< 10^{-410}$)	31 ($< 10^{-217}$)
$M_{0.35}$						
0.025	52.0425	63.5655	63.3146	64.0807	**64.4421**	64.4276
	-165 (≈ 1)		-4 (≈ 1)	8 ($< 10^{-14}$)	13 ($< 10^{-38}$)	13 ($< 10^{-37}$)
0.05	55.8783	75.5282	75.9912	76.3649	**76.9725**	76.9045
	-293 (≈ 1)		8 ($< 10^{-14}$)	14 ($< 10^{-44}$)	24 ($< 10^{-127}$)	23 ($< 10^{-116}$)
0.075	60.3367	84.721	85.6847	85.7007	86.3426	**86.419**
	-386 (≈ 1)		19 ($< 10^{-82}$)	20 ($< 10^{-85}$)	33 ($< 10^{-233}$)	34 ($< 10^{-256}$)
0.1	65.1425	90.9883	92.1735	91.9258	92.5377	**92.3998**
	-442 (≈ 1)		30 ($< 10^{-200}$)	24 ($< 10^{-124}$)	36 ($< 10^{-286}$)	40 ($< 10^{-347}$)
$M_{0.4}$						
0.025	52.2513	63.4259	63.2017	63.9212	64.2594	**64.5427**
	-160 (≈ 1)		-3 (≈ 1)	7 ($< 10^{-13}$)	12 ($< 10^{-35}$)	16 ($< 10^{-61}$)
0.05	55.9044	75.2794	75.8179	76.18	76.7506	**77.206**
	-288 (≈ 1)		9 ($< 10^{-19}$)	15 ($< 10^{-50}$)	24 ($< 10^{-131}$)	32 ($< 10^{-225}$)
0.075	60.3659	84.5171	85.6542	85.4506	86.0144	**86.6178**
	-382 (≈ 1)		23 ($< 10^{-113}$)	18 ($< 10^{-76}$)	30 ($< 10^{-196}$)	42 ($< 10^{-390}$)
0.1	65.2528	90.8962	92.2667	91.8219	92.2737	**92.8155**
	-438 (≈ 1)		35 ($< 10^{-267}$)	23 ($< 10^{-120}$)	35 ($< 10^{-270}$)	50 ($< 10^{-537}$)

Table 6: Performance table of $S = \{$QS10, TOCS, ROCiniS, pROCiniS, CROCS$\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (5, 15)$.

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.1}$						
0.025	54.6324	66.8434	65.0704	67.6892	**67.9541**	63.9289
	-177 (≈ 1)		-26 (≈ 1)	13 ($< 10^{-37}$)	17 ($< 10^{-63}$)	-43 (≈ 1)
0.05	60.631	79.6073	77.515	80.8308	**81.2462**	75.0604
	-293 (≈ 1)		-36 (≈ 1)	22 ($< 10^{-105}$)	29 ($< 10^{-188}$)	-77 (≈ 1)
0.075	66.2421	87.3117	85.51	88.5196	**88.8993**	82.6218
	-353 (≈ 1)		-37 (≈ 1)	26 ($< 10^{-151}$)	35 ($< 10^{-264}$)	-93 (≈ 1)
0.1	71.1357	91.4965	90.2486	92.6587	**92.9416**	87.2917
	-369 (≈ 1)		-31 (≈ 1)	30 ($< 10^{-203}$)	38 ($< 10^{-318}$)	-97 (≈ 1)
$M_{0.15}$						
0.025	53.8151	66.751	65.0398	66.6866	**67.1261**	64.0735
	-187 (≈ 1)		-26 (≈ 1)	-1 (0.833)	6 ($< 10^{-9}$)	-40 (≈ 1)
0.05	59.1924	80.0336	77.9334	79.9378	**80.5922**	75.8877
	-320 (≈ 1)		-36 (≈ 1)	-2 (0.955)	10 ($< 10^{-23}$)	-71 (≈ 1)
0.075	64.7978	88.4979	86.8231	88.4484	**89.0565**	84.3398
	-396 (≈ 1)		-36 (≈ 1)	-1 (0.863)	13 ($< 10^{-36}$)	-86 (≈ 1)
0.1	70.1593	93.3566	92.2868	93.3859	**93.7761**	89.8222
	-425 (≈ 1)		-29 (≈ 1)	1 (0.202)	12 ($< 10^{-34}$)	-90 (≈ 1)
$M_{0.2}$						
0.025	53.1601	**66.6077**	64.9542	65.6685	66.2252	64.1396
	-194 (≈ 1)		-25 (≈ 1)	-14(≈ 1)	-6 (≈ 1)	-37 (≈ 1)
0.05	57.9304	**79.6874**	78.0325	78.372	79.2708	76.2831
	-332 (≈ 1)		-29 (≈ 1)	-23 (≈ 1)	-7 (≈ 1)	-58 (≈ 1)
0.075	63.3024	**88.6911**	87.3592	87.3416	88.1658	85.0747
	-420 (≈ 1)		-29 (≈ 1)	-29 (≈ 1)	-12 (≈ 1)	-76 (≈ 1)
0.1	68.5969	**93.8452**	93.0379	92.8091	93.4091	90.8846
	-457 (≈ 1)		-23 (≈ 1)	-29 (≈ 1)	-13 (≈ 1)	-79 (≈ 1)
$M_{0.25}$						
0.025	52.6167	**66.0668**	64.7367	64.5509	65.2254	64.1651
	-194 (≈ 1)		-20 (≈ 1)	-23 (≈ 1)	-13 (≈ 1)	-28 (≈ 1)
0.05	56.9333	**79.4307**	78.1272	77.0549	78.1201	76.4954
	-342 (≈ 1)		-23 (≈ 1)	-41 (≈ 1)	-23 (≈ 1)	-50 (≈ 1)
0.075	61.9752	**88.4256**	87.6503	86.0857	87.1106	85.5309
	-433 (≈ 1)		-17 (≈ 1)	-50 (≈ 1)	-28 (≈ 1)	-61 (≈ 1)
0.1	67.2462	**93.8548**	93.4804	91.9583	92.7422	91.5724
	-475 (≈ 1)		-11 (≈ 1)	-52 (≈ 1)	-31 (≈ 1)	-62 (≈ 1)

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.3}$						
0.025	52.385	**65.8212**	64.6872	63.8978	64.61	64.297
	-193 (≈ 1)		-17 (≈ 1)	-28 (≈ 1)	-18 (≈ 1)	-23 (≈ 1)
0.05	56.4635	**78.9432**	78.073	75.9177	77.081	76.6645
	-340 (≈ 1)		-15 (≈ 1)	-51 (≈ 1)	-32 (≈ 1)	-39 (≈ 1)
0.075	61.1824	**88.1345**	87.7255	85.0109	86.1485	85.9144
	-438 (≈ 1)		-9 (≈ 1)	-65 (≈ 1)	-42 (≈ 1)	-47 (≈ 1)
0.1	66.3447	**93.7883**	93.7784	91.1884	92.0939	91.9995
	-486 (≈ 1)		0 (0.614)	-70 (≈ 1)	-47 (≈ 1)	-49 (≈ 1)
$M_{0.35}$						
0.025	52.2349	**65.5657**	64.5091	63.362	64.0945	64.3913
	-192 (≈ 1)		-16 (≈ 1)	-33 (≈ 1)	-22 (≈ 1)	-17 (≈ 1)
0.05	56.1588	**78.569**	77.906	75.0373	76.25	76.8751
	-338 (≈ 1)		-11 (≈ 1)	-59 (≈ 1)	-39 (≈ 1)	-29 (≈ 1)
0.075	60.8209	**87.8082**	87.673	84.1555	85.3681	86.2183
	-437 (≈ 1)		-3 (0.998)	-74 (≈ 1)	-51 (≈ 1)	-33 (≈ 1)
0.1	66.0295	93.6097	**93.8405**	90.6152	92.3571	91.5666
	-486 (≈ 1)		7 ($< 10^{-12}$)	-79 (≈ 1)	-55 (≈ 1)	-35 (≈ 1)
$M_{0.4}$						
0.025	52.3129	**65.2131**	64.0843	62.9659	63.5902	64.4291
	-185 (≈ 1)		-17 (≈ 1)	-33 (≈ 1)	-24 (≈ 1)	-12 (≈ 1)
0.05	56.2501	**78.263**	77.6079	74.5366	75.6444	77.0975
	-332 (≈ 1)		-11 (≈ 1)	-62 (≈ 1)	-44 (≈ 1)	-20 (≈ 1)
0.075	60.98	**87.5823**	87.5663	83.7024	84.8323	86.5298
	-430 (≈ 1)		0 (0.634)	-78 (≈ 1)	-56 (≈ 1)	-22 (≈ 1)
0.1	66.1484	93.4763	**93.7997**	90.2224	91.1473	92.753
	-481 (≈ 1)		9 ($< 10^{-21}$)	-84 (≈ 1)	-62 (≈ 1)	-20 (≈ 1)

Table 7: Performance table of $S = \{$QS10, TOCS, ROCiniS, pROCiniS, CROCS$\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (5, 25)$.

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.1}$						
0.025	52.7597	59.9867	62.2493	64.0278	64.0271	**64.0288**
	-103 (≈ 1)		33 ($< 10^{-236}$)	59 ($< 10^{-755}$)	59 ($< 10^{-755}$)	59 ($< 10^{-756}$)
0.05	57.0208	68.4202	72.6142	75.203	75.2004	**75.2125**
	-167 (≈ 1)		65 ($< 10^{-921}$)	107 ($< 10^{-2471}$)	107 ($< 10^{-2469}$)	107 ($< 10^{-2478}$)
0.075	61.2192	74.7703	80.0344	82.8188	**82.8203**	82.815
	-205 (≈ 1)		89 ($< 10^{-1723}$)	139 ($< 10^{-4212}$)	139 ($< 10^{-4214}$)	139 ($< 10^{-4208}$)
0.1	64.9877	79.2668	84.9207	87.5898	87.5907	**87.5921**
	-225 (≈ 1)		104 ($< 10^{-2364}$)	158 ($< 10^{-5443}$)	158 ($< 10^{-5444}$)	158 ($< 10^{-5446}$)
$M_{0.15}$						
0.025	52.2991	60.0253	62.3006	64.0962	**64.0963**	64.0876
	-110 (≈ 1)		33 ($< 10^{-239}$)	59 ($< 10^{-767}$)	59 ($< 10^{-767}$)	59 ($< 10^{-764}$)
0.05	56.1201	69.1739	73.5174	76.1635	76.1626	**76.1788**
	-191 (≈ 1)		68 ($< 10^{-1005}$)	111 ($< 10^{-2674}$)	111 ($< 10^{-2673}$)	111 ($< 10^{-2686}$)
0.075	60.47	76.487	81.9182	84.7614	84.7605	**84.7751**
	-244 (≈ 1)		95 ($< 10^{-1947}$)	148 ($< 10^{-4762}$)	148 ($< 10^{-4760}$)	148 ($< 10^{-4779}$)
0.1	64.9164	82.1097	87.9204	**90.3814**	90.3796	90.3805
	-276 (≈ 1)		115 ($< 10^{-2881}$)	170 ($< 10^{-6265}$)	170 ($< 10^{-6262}$)	170 ($< 10^{-6264}$)
$M_{0.2}$						
0.025	51.9603	60.1291	62.1588	64.2771	64.275	**64.2781**
	-116 (≈ 1)		29 ($< 10^{-191}$)	60 ($< 10^{-797}$)	60 ($< 10^{-796}$)	61 ($< 10^{-798}$)
0.05	55.5704	69.5013	73.5715	76.5493	76.5514	**76.5618**
	-204 (≈ 1)		64 ($< 10^{-886}$)	112 ($< 10^{-2741}$)	112 ($< 10^{-2743}$)	112 ($< 10^{-2751}$)
0.075	59.787	77.3812	82.683	85.6851	85.6858	**85.7064**
	-268 (≈ 1)		94 ($< 10^{-1913}$)	151 ($< 10^{-4975}$)	151 ($< 10^{-4976}$)	152 ($< 10^{-5003}$)
0.1	64.3048	83.5219	88.9761	91.7685	**91.7729**	91.7557
	-309 (≈ 1)		112 ($< 10^{-2726}$)	177 ($< 10^{-6822}$)	177 ($< 10^{-6830}$)	177 ($< 10^{-6798}$)
$M_{0.25}$						
0.025	52.0727	60.2909	62.0081	64.355	**64.3629**	64.3547
	-117 (≈ 1)		25 ($< 10^{-137}$)	59 ($< 10^{-766}$)	59 ($< 10^{-769}$)	59 ($< 10^{-766}$)
0.05	55.3901	69.7734	73.5322	76.9357	76.9393	**76.9533**
	-210 (≈ 1)		59 ($< 10^{-758}$)	115 ($< 10^{-2853}$)	115 ($< 10^{-2856}$)	115 ($< 10^{-2867}$)
0.075	59.498	77.9661	82.6597	**86.2772**	86.2747	86.2649
	-282 (≈ 1)		83 ($< 10^{-1516}$)	153 ($< 10^{-5111}$)	153 ($< 10^{-5108}$)	153 ($< 10^{-5095}$)
0.1	64.0417	84.3343	89.3345	**92.4145**	92.4087	92.4083
	-328 (≈ 1)		105 ($< 10^{-2377}$)	178 ($< 10^{-6903}$)	178 ($< 10^{-6891}$)	178 ($< 10^{-6891}$)

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.3}$						
0.025	51.9196	60.2795	61.8119	64.5453	**64.5519**	64.5414
	-119 (≈ 1)		22 ($< 10^{-109}$)	62 ($< 10^{-845}$)	62 ($< 10^{-848}$)	62 ($< 10^{-843}$)
0.05	55.2825	70.0868	73.4078	77.2987	77.2966	**77.3108**
	-216 (≈ 1)		52 ($< 10^{-593}$)	116 ($< 10^{-2916}$)	116 ($< 10^{-2914}$)	116 ($< 10^{-2926}$)
0.075	59.363	78.2777	82.6384	86.6667	86.668	**86.6718**
	-289 (≈ 1)		78 ($< 10^{-1316}$)	156 ($< 10^{-5289}$)	156 ($< 10^{-5291}$)	156 ($< 10^{-5296}$)
0.1	63.682	84.8769	89.4167	92.8908	**92.8911**	92.8892
	-343 (≈ 1)		96 ($< 10^{-2001}$)	180 ($< 10^{-7060}$)	180 ($< 10^{-7061}$)	180 ($< 10^{-7057}$)
$M_{0.35}$						
0.025	51.9202	60.4391	61.7136	64.6388	**64.6455**	64.6298
	-121 (≈ 1)		18 ($< 10^{-76}$)	61 ($< 10^{-820}$)	61 ($< 10^{-823}$)	61 ($< 10^{-817}$)
0.05	55.1776	70.2187	73.1388	**77.4825**	77.478	77.4731
	-220 (≈ 1)		46 ($< 10^{-459}$)	117 ($< 10^{-2969}$)	117 ($< 10^{-2966}$)	117 ($< 10^{-2961}$)
0.075	59.2723	78.4174	82.4804	86.9557	**86.9581**	86.9449
	-292 (≈ 1)		72 ($< 10^{-1142}$)	160 ($< 10^{-5532}$)	160 ($< 10^{-5536}$)	159 ($< 10^{-5517}$)
0.1	63.4811	85.1164	89.3843	93.0803	93.0785	**93.0835**
	-350 (≈ 1)		90 ($< 10^{-1781}$)	181 ($< 10^{-7093}$)	181 ($< 10^{-7089}$)	181 ($< 10^{-7099}$)
$M_{0.4}$						
0.025	52.0469	60.4855	61.5774	**64.7917**	64.7895	64.7641
	-120 (≈ 1)		16 ($< 10^{-57}$)	63 ($< 10^{-863}$)	63 ($< 10^{-862}$)	63 ($< 10^{-852}$)
0.05	55.2612	70.169	72.9707	77.5238	77.5273	**77.5338**
	-218 (≈ 1)		44 ($< 10^{-421}$)	118 ($< 10^{-3044}$)	118 ($< 10^{-3047}$)	119 ($< 10^{-3053}$)
0.075	59.2241	78.6469	82.4156	87.1406	**87.1452**	87.1312
	-297 (≈ 1)		67 ($< 10^{-986}$)	159 ($< 10^{-5527}$)	160 ($< 10^{-5534}$)	159 ($< 10^{-5514}$)
0.1	63.3733	85.2506	89.299	93.2793	**93.2821**	93.2637
	-354 (≈ 1)		86 ($< 10^{-1605}$)	183 ($< 10^{-7307}$)	183 ($< 10^{-7313}$)	183 ($< 10^{-7274}$)

Table 8: Performance table of $S = \{$QS10, TOCS, ROCiniS, pROCiniS, CROCS$\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (15, 15)$.

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.1}$						
0.025	52.8053	59.9933	62.278	**64.0769**	64.0764	64.0627
	-102 (≈ 1)		33 ($< 10^{-241}$)	60 ($< 10^{-771}$)	59 ($< 10^{-771}$)	59 ($< 10^{-766}$)
0.05	57.0069	68.4419	72.5538	75.1499	75.1492	**75.1663**
	-167 (≈ 1)		64 ($< 10^{-885}$)	105 ($< 10^{-2416}$)	105 ($< 10^{-2415}$)	106 ($< 10^{-2428}$)
0.075	61.2186	74.6492	79.9627	82.6394	82.6415	**82.6599**
	-203 (≈ 1)		90 ($< 10^{-1750}$)	138 ($< 10^{-4130}$)	138 ($< 10^{-4133}$)	138 ($< 10^{-4153}$)
0.1	65.0076	79.1688	84.8239	87.5221	**87.523**	87.5197
	-223 (≈ 1)		104 ($< 10^{-2355}$)	159 ($< 10^{-5461}$)	159 ($< 10^{-5462}$)	158 ($< 10^{-5458}$)
$M_{0.15}$						
0.025	52.2129	60.1674	62.3332	**64.1026**	64.0952	64.0985
	-113 (≈ 1)		31 ($< 10^{-217}$)	57 ($< 10^{-717}$)	57 ($< 10^{-715}$)	57 ($< 10^{-716}$)
0.05	56.0746	69.1463	73.4905	76.0732	**76.0736**	76.0712
	-191 (≈ 1)		68 ($< 10^{-1004}$)	110 ($< 10^{-2622}$)	110 ($< 10^{-2623}$)	110 ($< 10^{-2621}$)
0.075	60.4664	76.5426	81.9918	**84.766**	84.7657	84.7558
	-245 (≈ 1)		95 ($< 10^{-1965}$)	147 ($< 10^{-4709}$)	147 ($< 10^{-4708}$)	147 ($< 10^{-4696}$)
0.1	64.8071	82.1881	87.9383	90.4687	90.4723	**90.4738**
	-278 (≈ 1)		114 ($< 10^{-2828}$)	170 ($< 10^{-6311}$)	171 ($< 10^{-6317}$)	171 ($< 10^{-6320}$)
$M_{0.2}$						
0.025	52.0649	60.2348	62.1698	**64.3377**	64.3387	64.3326
	-116 (≈ 1)		28 ($< 10^{-174}$)	60 ($< 10^{-781}$)	60 ($< 10^{-781}$)	60 ($< 10^{-779}$)
0.05	55.5927	69.5836	73.5876	76.6288	76.6276	**76.6341**
	-204 (≈ 1)		63 ($< 10^{-858}$)	112 ($< 10^{-2744}$)	112 ($< 10^{-2743}$)	112 ($< 10^{-2748}$)
0.075	59.7767	77.4328	82.631	85.7583	85.7594	**85.7677**
	-269 (≈ 1)		92 ($< 10^{-1839}$)	152 ($< 10^{-5014}$)	152 ($< 10^{-5016}$)	152 ($< 10^{-5027}$)
0.1	64.4642	83.6085	89.0608	91.7622	**91.7637**	91.7608
	-309 (≈ 1)		112 ($< 10^{-2739}$)	175 ($< 10^{-6688}$)	175 ($< 10^{-6691}$)	175 ($< 10^{-6685}$)
$M_{0.25}$						
0.025	51.9684	60.3052	61.7969	**64.4926**	64.4925	64.4786
	-119 (≈ 1)		22 ($< 10^{-104}$)	61 ($< 10^{-814}$)	61 ($< 10^{-814}$)	61 ($< 10^{-809}$)
0.05	55.4662	69.8989	73.3223	77.0364	77.0408	**77.0435**
	-211 (≈ 1)		54 ($< 10^{-629}$)	114 ($< 10^{-2841}$)	114 ($< 10^{-2844}$)	114 ($< 10^{-2846}$)
0.075	59.5807	78.037	82.5769	86.3658	**86.372**	86.3709
	-282 (≈ 1)		81 ($< 10^{-1418}$)	154 ($< 10^{-5151}$)	154 ($< 10^{-5159}$)	154 ($< 10^{-5158}$)
0.1	64.0489	84.4528	89.2709	92.4907	**92.4934**	92.4802
	-330 (≈ 1)		101 ($< 10^{-2211}$)	178 ($< 10^{-6881}$)	178 ($< 10^{-6886}$)	178 ($< 10^{-6860}$)

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.3}$						
0.025	52.0966	60.4116	61.527	**64.5955**	64.5934	64.5878
	-119 (≈ 1)		16 ($< 10^{-59}$)	61 ($< 10^{-814}$)	61 ($< 10^{-813}$)	61 ($< 10^{-811}$)
0.05	55.4257	70.0568	73.031	**77.351**	77.3476	77.3328
	-214 (≈ 1)		47 ($< 10^{-474}$)	117 ($< 10^{-2984}$)	117 ($< 10^{-2981}$)	117 ($< 10^{-2968}$)
0.075	59.4718	78.329	82.384	**86.81**	86.8083	86.8028
	-288 (≈ 1)		72 ($< 10^{-1134}$)	158 ($< 10^{-5429}$)	158 ($< 10^{-5427}$)	158 ($< 10^{-5419}$)
0.1	63.8719	84.9447	89.1916	92.9359	**92.9364**	92.9319
	-341 (≈ 1)		89 ($< 10^{-1742}$)	180 ($< 10^{-7052}$)	180 ($< 10^{-7053}$)	180 ($< 10^{-7044}$)
$M_{0.35}$						
0.025	52.0055	60.4265	61.4676	**64.7257**	64.7163	64.7186
	-120 (≈ 1)		15 ($< 10^{-52}$)	63 ($< 10^{-860}$)	63 ($< 10^{-856}$)	63 ($< 10^{-857}$)
0.05	55.4179	70.2287	72.7862	77.5797	77.5813	**77.5856**
	-217 (≈ 1)		40 ($< 10^{-351}$)	118 ($< 10^{-3045}$)	118 ($< 10^{-3046}$)	118 ($< 10^{-3050}$)
0.075	59.4158	78.6093	82.2603	87.129	**87.1336**	87.1282
	-293 (≈ 1)		65 ($< 10^{-922}$)	160 ($< 10^{-5554}$)	160 ($< 10^{-5561}$)	160 ($< 10^{-5553}$)
0.1	63.6248	85.2072	89.1124	93.208	**93.2086**	93.2043
	-350 (≈ 1)		83 ($< 10^{-1482}$)	182 ($< 10^{-7222}$)	182 ($< 10^{-7223}$)	182 ($< 10^{-7214}$)
$M_{0.4}$						
0.025	51.9371	60.5425	61.2204	64.8059	**64.8085**	64.7869
	-123 (≈ 1)		10 ($< 10^{-23}$)	62 ($< 10^{-846}$)	62 ($< 10^{-847}$)	62 ($< 10^{-839}$)
0.05	55.3621	70.3693	72.5732	77.7221	77.7233	**77.7291**
	-220 (≈ 1)		35 ($< 10^{-261}$)	119 ($< 10^{-3057}$)	119 ($< 10^{-3058}$)	119 ($< 10^{-3063}$)
0.075	59.2526	78.7275	82.0071	**87.2744**	87.2699	87.2744
	-298 (≈ 1)		58 ($< 10^{-743}$)	161 ($< 10^{-5624}$)	161 ($< 10^{-5618}$)	161 ($< 10^{-5624}$)
0.1	63.4981	85.4048	89.0213	**93.4732**	93.4679	93.4617
	-355 (≈ 1)		77 ($< 10^{-1276}$)	186 ($< 10^{-7486}$)	185 ($< 10^{-7475}$)	185 ($< 10^{-7461}$)

Table 9: Performance table of $S = \{$QS10, TOCS, ROCiniS, pROCiniS, CROCS$\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (25, 25)$.

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.1}$						
0.025	52.352	57.8046	66.0224	67.8647	**68.1268**	64.0882
	-78 (≈ 1)		120 ($< 10^{-3112}$)	147 ($< 10^{-4708}$)	151 ($< 10^{-4964}$)	91 ($< 10^{-1804}$)
0.05	56.4419	64.4646	78.3938	80.8141	**81.2257**	75.1184
	-116 (≈ 1)		218 ($< 10^{-10326}$)	259 ($< 10^{-14606}$)	266 ($< 10^{-15423}$)	164 ($< 10^{-5848}$)
0.075	60.5394	69.7164	86.0813	88.4986	**88.8664**	82.5899
	-136 (≈ 1)		279 ($< 10^{-16892}$)	327 ($< 10^{-23178}$)	334 ($< 10^{-24252}$)	214 ($< 10^{-9912}$)
0.1	64.0456	73.5169	90.5439	92.7001	**92.9862**	87.2991
	-145 (≈ 1)		314 ($< 10^{-21358}$)	362 ($< 10^{-28465}$)	369 ($< 10^{-29520}$)	246 ($< 10^{-13095}$)
$M_{0.15}$						
0.025	52.2609	57.8158	65.5071	66.7175	**67.1768**	64.0363
	-79 (≈ 1)		112 ($< 10^{-2720}$)	130 ($< 10^{-3665}$)	137 ($< 10^{-4062}$)	90 ($< 10^{-1768}$)
0.05	55.8742	65.0747	78.3222	79.9201	**80.6001**	75.9374
	-133 (≈ 1)		208 ($< 10^{-9393}$)	235 ($< 10^{-12004}$)	247 ($< 10^{-13231}$)	168 ($< 10^{-6164}$)
0.075	59.7179	71.0597	86.8283	88.4174	**89.0078**	84.3402
	-169 (≈ 1)		273 ($< 10^{-16244}$)	305 ($< 10^{-20251}$)	317 ($< 10^{-21891}$)	226 ($< 10^{-11055}$)
0.1	63.4478	75.8302	91.9917	93.3567	**93.7656**	89.7839
	-190 (≈ 1)		311 ($< 10^{-21009}$)	343 ($< 10^{-25594}$)	353 ($< 10^{-27096}$)	261 ($< 10^{-14852}$)
$M_{0.2}$						
0.025	52.0021	57.9104	64.9282	65.6087	**66.1909**	64.1701
	-84 (≈ 1)		102 ($< 10^{-2259}$)	112 ($< 10^{-2727}$)	121 ($< 10^{-3164}$)	91 ($< 10^{-1792}$)
0.05	55.4382	65.2638	77.7476	78.4399	**79.3747**	76.28
	-142 (≈ 1)		196 ($< 10^{-8308}$)	207 ($< 10^{-9323}$)	223 ($< 10^{-10803}$)	171 ($< 10^{-6373}$)
0.075	59.2291	71.7116	86.7369	87.3001	**88.1175**	85.071
	-186 (≈ 1)		262 ($< 10^{-14895}$)	273 ($< 10^{-16195}$)	290 ($< 10^{-18209}$)	230 ($< 10^{-11443}$)
0.1	62.8975	76.9786	92.3073	92.8408	**93.4111**	90.9289
	-217 (≈ 1)		301 ($< 10^{-19630}$)	313 ($< 10^{-21324}$)	327 ($< 10^{-23247}$)	269 ($< 10^{-15688}$)
$M_{0.25}$						
0.025	51.9829	58.0464	64.3531	64.5397	**65.2783**	64.2548
	-86 (≈ 1)		92 ($< 10^{-1821}$)	94 ($< 10^{-1932}$)	105 ($< 10^{-2405}$)	90 ($< 10^{-1764}$)
0.05	55.2406	65.3903	77.0576	76.9195	**78.0167**	76.2994
	-147 (≈ 1)		182 ($< 10^{-7214}$)	180 ($< 10^{-7035}$)	198 ($< 10^{-8534}$)	170 ($< 10^{-6259}$)
0.075	58.9249	72.1052	86.3204	86.0681	**87.091**	85.4726
	-196 (≈ 1)		248 ($< 10^{-13327}$)	243 ($< 10^{-12801}$)	263 ($< 10^{-15018}$)	231 ($< 10^{-11612}$)
0.1	62.5599	77.6762	92.1582	91.9572	**92.7273**	91.4951
	-234 (≈ 1)		286 ($< 10^{-17782}$)	281 ($< 10^{-17198}$)	300 ($< 10^{-19511}$)	271 ($< 10^{-15905}$)

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.3}$						
0.025	51.8604	58.0852	63.987	63.8377	**64.5611**	64.3052
	-88 (≈ 1)		86 ($< 10^{-1593}$)	83 ($< 10^{-1513}$)	94 ($< 10^{-1923}$)	90 ($< 10^{-1772}$)
0.05	55.1093	65.6515	76.6339	75.8976	**77.0747**	76.6226
	-152 (≈ 1)		171 ($< 10^{-6382}$)	159 ($< 10^{-5514}$)	179 ($< 10^{-6936}$)	171 ($< 10^{-6368}$)
0.075	58.8637	72.5209	85.9562	85.0105	**86.1332**	85.9048
	-203 (≈ 1)		234 ($< 10^{-11916}$)	216 ($< 10^{-10129}$)	238 ($< 10^{-12271}$)	233 ($< 10^{-11815}$)
0.1	62.4228	78.2885	92.0175	91.194	**92.0913**	91.9942
	-246 (≈ 1)		273 ($< 10^{-16190}$)	254 ($< 10^{-13988}$)	275 ($< 10^{-16398}$)	272 ($< 10^{-16125}$)
$M_{0.35}$						
0.025	51.8998	58.1542	63.5615	63.2984	64.0218	**64.3258**
	-89 (≈ 1)		78 ($< 10^{-1335}$)	74 ($< 10^{-1207}$)	85 ($< 10^{-1575}$)	90 ($< 10^{-1745}$)
0.05	55.0249	65.8863	76.075	75.0916	76.2376	**76.8666**
	-157 (≈ 1)		159 ($< 10^{-5475}$)	143 ($< 10^{-4426}$)	161 ($< 10^{-5660}$)	172 ($< 10^{-6410}$)
0.075	58.6227	72.7957	85.4717	84.2119	85.3912	**86.1387**
	-211 (≈ 1)		221 ($< 10^{-10569}$)	197 ($< 10^{-8388}$)	219 ($< 10^{-10420}$)	234 ($< 10^{-11850}$)
0.1	62.265	78.872	91.7462	90.5802	91.5316	**92.3829**
	-258 (≈ 1)		257 ($< 10^{-14362}$)	230 ($< 10^{-11504}$)	252 ($< 10^{-13804}$)	272 ($< 10^{-16108}$)
$M_{0.4}$						
RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
0.025	51.8875	58.1743	63.1818	62.8936	63.5561	**64.433**
	-89 (≈ 1)		72 ($< 10^{-1144}$)	68 ($< 10^{-1015}$)	78 ($< 10^{-1323}$)	91 ($< 10^{-1796}$)
0.05	54.9124	66.1669	75.7744	74.5806	75.6632	**77.1095**
	-163 (≈ 1)		150 ($< 10^{-4867}$)	130 ($< 10^{-3689}$)	148 ($< 10^{-4750}$)	172 ($< 10^{-6402}$)
0.075	58.4559	73.2721	85.1367	83.7395	84.8881	**86.5661**
	-221 (≈ 1)		207 ($< 10^{-9282}$)	180 ($< 10^{-7053}$)	202 ($< 10^{-8859}$)	235 ($< 10^{-11960}$)
0.1	62.0305	79.3659	91.607	90.2587	91.2052	**92.763**
	-269 (≈ 1)		246 ($< 10^{-13116}$)	215 ($< 10^{-10004}$)	236 ($< 10^{-12130}$)	274 ($< 10^{-16251}$)

Table 10: Performance table of $S = \{$QS10, TOCS, ROCiniS, pROCiniS, CROCS$\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (25, 5)$.

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.1}$						
0.025	52.4609	58.1806	64.2311	65.8896	**65.982**	63.9716
	-81 (≈ 1)		88 ($< 10^{-1677}$)	112 ($< 10^{-2743}$)	114 ($< 10^{-2810}$)	84 ($< 10^{-1534}$)
0.05	56.5179	65.1682	75.8462	78.2073	**78.398**	75.1307
	-125 (≈ 1)		166 ($< 10^{-5956}$)	205 ($< 10^{-9098}$)	208 ($< 10^{-9385}$)	154 ($< 10^{-5149}$)
0.075	60.7686	70.7859	83.6439	86.1489	**86.3664**	82.8754
	-149 (≈ 1)		217 ($< 10^{-10206}$)	264 ($< 10^{-15170}$)	269 ($< 10^{-15660}$)	203 ($< 10^{-8918}$)
0.1	64.6259	74.8757	88.4036	90.7852	**90.9941**	87.7132
	-158 (≈ 1)		247 ($< 10^{-13259}$)	298 ($< 10^{-19327}$)	303 ($< 10^{-19934}$)	233 ($< 10^{-11770}$)
$M_{0.15}$						
0.025	52.0916	58.2761	63.9599	65.736	**65.9083**	64.1711
	-88 (≈ 1)		82 ($< 10^{-1479}$)	109 ($< 10^{-2568}$)	111 ($< 10^{-2690}$)	86 ($< 10^{-1592}$)
0.05	55.7957	65.8354	76.156	78.5761	**78.9358**	76.115
	-145 (≈ 1)		161 ($< 10^{-5619}$)	201 ($< 10^{-8785}$)	207 ($< 10^{-9325}$)	160 ($< 10^{-5572}$)
0.075	59.9207	72.2282	84.7596	87.2721	**87.6311**	84.6811
	-184 (≈ 1)		216 ($< 10^{-10103}$)	265 ($< 10^{-15219}$)	272 ($< 10^{-16060}$)	214 ($< 10^{-9964}$)
0.1	64.0409	77.2636	90.3118	92.5497	**92.8556**	90.2325
	-205 (≈ 1)		250 ($< 10^{-13611}$)	302 ($< 10^{-19800}$)	309 ($< 10^{-20774}$)	249 ($< 10^{-13420}$)
$M_{0.2}$						
0.025	51.9848	58.3114	63.628	65.277	**65.5899**	64.2095
	-90 (≈ 1)		77 ($< 10^{-1292}$)	101 ($< 10^{-2234}$)	106 ($< 10^{-2443}$)	86 ($< 10^{-1594}$)
0.05	55.368	66.0766	75.7971	78.0599	78.5679	**76.3442**
	-155 (≈ 1)		151 ($< 10^{-4979}$)	189 ($< 10^{-7748}$)	197 ($< 10^{-8466}$)	160 ($< 10^{-5586}$)
0.075	59.2498	72.7553	84.7956	87.105	87.5609	**85.3245**
	-202 (≈ 1)		208 ($< 10^{-9417}$)	253 ($< 10^{-13940}$)	263 ($< 10^{-14967}$)	218 ($< 10^{-10357}$)
0.1	63.2109	78.4515	90.7138	92.6395	93.0176	**91.2377**
	-237 (≈ 1)		240 ($< 10^{-12522}$)	285 ($< 10^{-17679}$)	295 ($< 10^{-18839}$)	252 ($< 10^{-13808}$)
$M_{0.25}$						
0.025	51.8221	58.3664	63.2386	64.7258	**65.1075**	64.2398
	-93 (≈ 1)		71 ($< 10^{-1084}$)	92 ($< 10^{-1858}$)	98 ($< 10^{-2092}$)	85 ($< 10^{-1582}$)
0.05	55.1238	66.2358	75.3774	77.3627	**77.9616**	76.5404
	-161 (≈ 1)		142 ($< 10^{-4393}$)	175 ($< 10^{-6642}$)	185 ($< 10^{-7424}$)	161 ($< 10^{-5647}$)
0.075	58.8481	73.2135	84.4727	86.4672	**87.0406**	85.6721
	-214 (≈ 1)		195 ($< 10^{-8255}$)	234 ($< 10^{-11853}$)	245 ($< 10^{-13039}$)	218 ($< 10^{-10322}$)
0.1	62.7808	79.1133	90.7379	92.4942	**92.9473**	91.8471
	-254 (≈ 1)		230 ($< 10^{-11464}$)	271 ($< 10^{-15963}$)	282 ($< 10^{-17293}$)	256 ($< 10^{-14188}$)

RN	QS10	QS	TOCS	ROCiniS	pROCiniS	CROCS
$M_{0.3}$						
0.025	51.9051	58.4368	62.9972	64.384	**64.833**	64.3651
	-93 (≈ 1)		66 ($< 10^{-949}$)	86 ($< 10^{-1623}$)	93 ($< 10^{-1881}$)	86 ($< 10^{-1613}$)
0.05	55.0463	66.5254	74.9611	76.7688	**77.4243**	76.7603
	-166 (≈ 1)		131 ($< 10^{-3736}$)	161 ($< 10^{-5611}$)	172 ($< 10^{-6397}$)	161 ($< 10^{-5601}$)
0.075	58.7609	73.5828	84.275	86.101	**86.7432**	86.1015
	-222 (≈ 1)		185 ($< 10^{-7467}$)	221 ($< 10^{-10575}$)	233 ($< 10^{-11829}$)	221 ($< 10^{-10575}$)
0.1	62.4759	79.7387	90.5917	92.1843	**92.6747**	92.2397
	-269 (≈ 1)		216 ($< 10^{-10126}$)	253 ($< 10^{-13939}$)	265 ($< 10^{-15283}$)	255 ($< 10^{-14087}$)
$M_{0.35}$						
0.025	51.8132	58.4793	62.7513	64.0942	**64.5006**	64.4329
	-95 (≈ 1)		62 ($< 10^{-833}$)	82 ($< 10^{-1446}$)	87 ($< 10^{-1665}$)	86 ($< 10^{-1627}$)
0.05	54.7688	66.6555	74.6245	76.3421	**76.985**	76.9518
	-172 (≈ 1)		124 ($< 10^{-3327}$)	152 ($< 10^{-5002}$)	162 ($< 10^{-5727}$)	162 ($< 10^{-5688}$)
0.075	58.4279	73.9825	83.9505	85.697	86.3373	**86.3559**
	-233 (≈ 1)		173 ($< 10^{-6498}$)	206 ($< 10^{-9260}$)	219 ($< 10^{-10424}$)	219 ($< 10^{-10459}$)
0.1	62.0888	80.1681	90.4645	91.9405	92.4244	**92.5442**
	-282 (≈ 1)		206 ($< 10^{-9191}$)	240 ($< 10^{-12542}$)	252 ($< 10^{-13795}$)	255 ($< 10^{-14118}$)
$M_{0.4}$						
0.025	51.7845	58.5859	62.4781	63.8332	64.2046	**64.4203**
	-97 (≈ 1)		56 ($< 10^{-691}$)	76 ($< 10^{-1262}$)	82 ($< 10^{-1449}$)	85 ($< 10^{-1564}$)
0.05	54.9121	66.8873	74.3998	76.0579	76.6236	**77.0851**
	-174 (≈ 1)		117 ($< 10^{-2958}$)	144 ($< 10^{-4481}$)	153 ($< 10^{-5081}$)	161 ($< 10^{-5602}$)
0.075	58.333	74.3194	83.6796	85.4053	85.9965	**86.5946**
	-239 (≈ 1)		162 ($< 10^{-5737}$)	195 ($< 10^{-8300}$)	207 ($< 10^{-9311}$)	219 ($< 10^{-10408}$)
0.1	61.9926	80.6044	90.2713	91.7942	92.2422	**92.8356**
	-291 (≈ 1)		194 ($< 10^{-8158}$)	229 ($< 10^{-11431}$)	240 ($< 10^{-12536}$)	255 ($< 10^{-14107}$)

Table 11: Performance table of $S = \{$QS10, TOCS, ROCiniS, pROCiniS, CROCS$\}$ for $r = 1000000$, $N = 1000$, $(\alpha, \beta) = (15, 5)$.

References

Ahmed M Alaa and Mihaela Van Der Schaar. Bayesian inference of individualized treatment effects using multi-task gaussian processes. Advances in neural information processing systems, 30, 2017.

Susan Athey and Guido W Imbens. Machine learning methods for estimating heterogeneous causal effects. stat, 1050(5):1–26, 2015.

Susan Athey, Guido W Imbens, and Stefan Wager. Approximate residual balancing: debiased inference of average treatment effects in high dimensions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4):597–623, 2018.

Daniel Baier and Björn Stöcker. Profit uplift modeling for direct marketing campaigns: approaches and applications for online shops. Journal of Business Economics, 92(4):645–673, 2022.

Donald Bamber. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of mathematical psychology, 12(4):387–415, 1975.

Mouloud Belbahri, Alejandro Murua, Olivier Gandouet, and Vahid Partovi Nia. Qini-based uplift regression. The Annals of Applied Statistics, 15(3):1247–1272, 2021.

Jeroen Berrevoets, James Jordon, Ioana Bica, Mihaela van der Schaar, et al. Organite: Optimal transplant donor organ offering using an individual treatment effect. Advances in neural information processing systems, 33:20037–20050, 2020.

Artem Betlei, Eustache Diemert, and Massih-Reza Amini. Uplift prediction with dependent feature representation in imbalanced treatment and control conditions. In Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13–16, 2018, Proceedings, Part V 25, pages 47–57. Springer, 2018.

Corinna Cortes and Mehryar Mohri. Confidence intervals for the area under the roc curve. Advances in neural information processing systems, 17, 2004.

Alicia Curth and Mihaela Van der Schaar. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In International Conference on Artificial Intelligence and Statistics, pages 1810–1818. PMLR, 2021.

Richard B Darlington. Comparing two groups by simple graphs. Psychological Bulletin, 79 (2):110, 1973.

Simon De Vos, Christopher Bockel-Rickermann, Stefan Lessmann, and Wouter Verbeke. Uplift modeling with continuous treatments: A predict-then-optimize approach. arXiv preprint arXiv:2412.09232, 2024.

Floris Devriendt, Darie Moldovan, and Wouter Verbeke. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big data, 6(1):13–41, 2018.

Floris Devriendt, Jente Van Belle, Tias Guns, and Wouter Verbeke. Learning to rank for uplift modeling. IEEE Transactions on Knowledge and Data Engineering, 2020.

Eustache Diemert, Artem Betlei, Christophe Renaudin, and Massih-Reza Amini. A large scale benchmark for uplift modeling. In KDD, 2018.

Diemert, Eustache, Betlei, Artem, Renaudin, Christophe, and Massih-Reza, Amini. A large scale benchmark for uplift modeling. In Proceedings of the AdKDD and TargetAd Workshop, KDD, London, United Kingdom, August 20, 2018. ACM, 2018. URL `https://ailab.criteo.com/criteo-uplift-prediction-dataset/`.

Lori E Dodd and Margaret S Pepe. Partial auc estimation and regression. Biometrics, 59 (3):614–623, 2003.

Tom Fawcett and Alexandru Niculescu-Mizil. Pav and the roc convex hull. Machine Learning, 68:97–106, 2007.

Carlos Fernández and Foster Provost. Causal classification: Treatment effect vs. outcome prediction. NYU Stern School of Business, 2019.

Robin M Gubela, Stefan Lessmann, and Szymon Jaroszewicz. Response transformation and profit decomposition for revenue uplift modeling. European Journal of Operational Research, 283(2):647–661, 2020.

Robin Marco Gubela, Stefan Lessmann, Johannes Haupt, Annika Baumann, Tillmann Radmer, and Fabian Gebert. Revenue uplift modeling. 2017.

Pierre Gutierrez and Jean-Yves Gérardy. Causal inference and uplift modelling: A review of the literature. In International conference on predictive applications and APIs, pages 1–13. PMLR, 2017.

James A Hanley, Barbara J McNeil, et al. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. Radiology, 148(3):839–843, 1983.

RA Hilgers. Distribution-free confidence bounds for roc curves. Methods of information in medicine, 30(02):96–101, 1991.

Jennifer L Hill. Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics, 20(1):217–240, 2011.

K. Hillstrom. The minethatdata e-mail analytics and data mining challenge. `http://blog.minethatdata.com/2008/03/minethatdata-e-mail-analytics-and-data.html`, 2008. Retrieved on 02.04.2012.

Paul W Holland. Statistics and causal inference. Journal of the American statistical Association, 81(396):945–960, 1986.

Maciej Jaskowski and Szymon Jaroszewicz. Uplift modeling for clinical trial data. In ICML workshop on clinical data analysis, volume 46, pages 79–95, 2012.

Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the national academy of sciences, 116(10):4156–4165, 2019.

Finn Kuusisto, Vitor Santos Costa, Houssam Nassif, Elizabeth Burnside, David Page, and Jude Shavlik. Support vector machines for differential prediction. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II 14, pages 50–65. Springer, 2014.

Victor SY Lo. The true lift model: a novel data mining approach to response modeling in database marketing. ACM SIGKDD Explorations Newsletter, 4(2):78–86, 2002.

Sofus Macskassy and Foster Provost. Confidence bands for roc curves: Methods and an empirical study. Proceedings of the First Workshop on ROC Analysis in AI. August 2004., 2004.

Alessandro Marchese, Hans de Ferrante, Jeroen Berrevoets, and Sam Verboven. Dynamite: Optimal time-sensitive organ offers using ite. Machine Learning for Health (ML4H), pages 696–713, 2025.

Simon J Mason and Nicholas E Graham. Areas beneath the relative operating characteristics (roc) and relative operating levels (rol) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, 128(584):2145–2166, 2002.

Diego Olaya, Kristof Coussement, and Wouter Verbeke. A survey and benchmarking study of multitreatment uplift modeling. Data Mining and Knowledge Discovery, 34(2):273–308, 2020a.

Diego Olaya, Jonathan Vásquez, Sebastián Maldonado, Jaime Miranda, and Wouter Verbeke. Uplift modeling for preventing student dropout in higher education. Decision Support Systems, 134:113320, 2020b.

Charles S Peirce. The numerical measure of the success of predictions. Science, (93):453–454, 1884.

Foster Provost and Tom Fawcett. Robust classification for imprecise environments. Machine learning, 42:203–231, 2001.

Nicholas Radcliffe. Using control groups to target on predicted lift: Building and assessing uplift model. Direct Marketing Analytics Journal, pages 14–21, 2007.

Nicholas J Radcliffe and Patrick D Surry. Real-world uplift modelling with significance-based uplift trees. White Paper TR-2011-1, Stochastic Solutions, pages 1–33, 2011.

Evy Rombaut and Marie-Anne Guerry. The effectiveness of employee retention through an uplift modeling approach. International Journal of Manpower, 41(8):1199–1220, 2020.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1):41–55, 1983.

Enrique F Schisterman, Neil J Perkins, Aiyi Liu, and Howard Bondell. Optimal cut-point and its corresponding youden index to discriminate individuals using pooled blood samples. Epidemiology, 16(1):73–81, 2005.

Pranab Kumar Sen. A note on asymptotically distribution-free confidence bounds for p {X< Y}, based on two independent samples. Sankhyā: The Indian Journal of Statistics, Series A, pages 95–102, 1967.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In International Conference on Machine Learning, pages 3076–3085. PMLR, 2017.

Huiyang Shao, Qianqian Xu, Zhiyong Yang, Peisong Wen, Gao Peifeng, and Qingming Huang. Weighted roc curve in cost space: Extending auc to cost-sensitive learning. Advances in Neural Information Processing Systems, 36:17357–17368, 2023.

Samuel Sanford Shapiro and Martin B Wilk. An analysis of variance test for normality (complete samples). Biometrika, 52(3/4):591–611, 1965.

Adrian J Simpson and Mike J Fitter. What is the best index of detectability? Psychological Bulletin, 80(6):481, 1973.

Eugene Somoza and Douglas Mossman. Roc curves and the binormal assumption. The Journal of Neuropsychiatry and Clinical Neurosciences, 1991.

Shashikala Sukhatme and CA Beam. Stratification in nonparametric roc studies. Biometrics, pages 149–163, 1994.

D Van Dantzig. two sample test. Indagationes Mathematicae, 13:1–8, 1951.

Wouter Verbeke, Diego Olaya, Jeroen Berrevoets, Sam Verboven, and Sebastián Maldonado. The foundations of cost-sensitive causal classification. arXiv preprint arXiv:2007.12582, 2020.

Sam Verboven and Niels Martin. Combining the clinical and operational perspectives in heterogeneous treatment effect inference in healthcare processes. In International Conference on Process Mining, pages 327–339. Springer, 2022.

Nisus Writer and Others. Information: Data Exploration with Information Theory, 2021. URL `https://cran.r-project.org/web/packages/Information/Information.pdf`. R package version 0.2.1.

Steve Yadlowsky, Scott Fleming, Nigam Shah, Emma Brunskill, and Stefan Wager. Evaluating treatment prioritization rules via rank-weighted average treatment effects. Journal of the American Statistical Association, pages 1–14, 2024.

Lihui Zhao, Lu Tian, Tianxi Cai, Brian Claggett, and Lee-Jen Wei. Effectively selecting a target population for a future comparative study. Journal of the American Statistical Association, 108(502):527–539, 2013.

Kelly H Zou, Julia R Fielding, Stuart G Silverman, and Clare MC Tempany. Hypothesis testing i: proportions. Radiology, 226(3):609–613, 2003.