Model Evaluation in the Dark: Robust Classifier Metrics with Missing Labels

Danial Dervovic
J.P. Morgan AI Research
315 Argyle Street, Glasgow, UK
danial.dervovic@jpmorgan.com

Michael Cashmore
J.P. Morgan AI Research
315 Argyle Street, Glasgow, UK
michael.cashmore@jpmorgan.com

Abstract

Missing data in supervised learning is well-studied, but the specific issue of *missing labels* during model evaluation has been overlooked. Ignoring samples with missing values, a common solution, can introduce bias, especially when data is Missing Not At Random (MNAR). We propose a multiple imputation technique for evaluating classifiers using metrics such as precision, recall, and ROC-AUC. This method not only offers point estimates but also a predictive distribution for these quantities when labels are missing. We empirically show that the predictive distribution's location and shape are generally correct, even in the MNAR regime. Moreover, we establish that this distribution is approximately Gaussian and provide finite-sample convergence bounds. Additionally, a robustness proof is presented, confirming the validity of the approximation under a realistic error model.

1 INTRODUCTION

Missing data, the absence of values in a dataset, remains a persistent challenge in Machine Learning (ML). It introduces bias, hampers generalisation, and ultimately diminishes a model's predictive performance.

Multiple mechanisms can lead to data being missing (Rubin, 1976), most commonly Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). MCAR corresponds to the causes of the missing data, or "missingness", being completely unrelated to the data itself. MAR means missingness is related to the observed data but not the missing data. If neither MCAR nor MAR hold, we have MNAR where missingness is related to the unobserved data. These mechanisms can be formalised using graphical models, as in Figure 1. Deducing whether data from a particular source are missing due to MCAR, MAR or MNAR is difficult without access to the data-generating mechanism (d'Haultfoeuille, 2010).

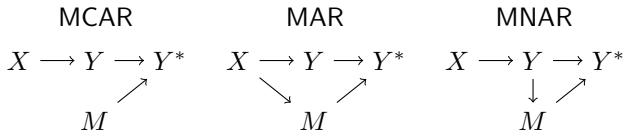

Figure 1: Illustration of missingness mechanisms in relation to this work, with arrows indicating causation. $X \in \mathcal{X}$ are input features, $Y \in \mathcal{Y}$ is the label, $M \in \{0, 1\}$ is the missingness flag and $Y^* \in \mathcal{Y} \cup \{\texttt{NA}\}$ is the masked label.

There are numerous approaches to remove or impute missing values, such as ignoring samples with missing values, known as listwise deletion (Allison, 2001); imputing a random, static, or statistically derived value (Donders et al., 2006); or training another model to impute a predicted value (Little et al., 2019; Song and Shepperd, 2007). Without careful handling of the missing data, the predictive power of ML models can be reduced (Ayilara et al., 2019).

This work considers another issue for ML in the presence of missing data, which is *evaluating* models, specifically when the missing values are ground-truth labels at test-time. Unless labels are MCAR then any technique to ignore or impute missing values can lead to a biased estimate (Williams, 2015). This is an issue that can appear in model monitoring contexts (Kenthapadi et al., 2022), when there is a delay in receiving labels from a production ML system. For example, in evaluating a fraud detection model we may receive some ground-truth labels before others; or some users

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

may delay providing feedback to a recommender system.

In particular, for the setting where some fraction of labels are missing at test-time and the task is evaluating the performance of a binary classifier, we provide algorithms – Performance Estimation by Multiple Imputation (PEMI), in Algorithm 1, and PEMI-Gauss in Algorithm 2 – for correctly describing a decision-maker's ignorance about the value of various performance measures. Notably, these include precision, recall and ROC-AUC amongst others. To the authors' knowledge this work is the first to tackle missing labels and ML model evaluation simultaneously.

The description of the ignorance of a performance measure's value is naturally given as a probability distribution. We rigorously prove finite-sample convergence bounds to a normal distribution in Theorem 3. The algorithms are further proven in Theorem 4 to be immune to a realistic amount of noise in the imputation process. As an intermediate step, we prove finite-sample convergence bounds of correlated Gaussian random variables to a Gaussian distribution (Theorem 2), which may be of independent interest. Experiments on real datasets in Section C show that PEMI and PEMI-Gauss are effective over a range of missingness mechanisms when even a large amount of labels (30%) are missing.

Related Work. Missing data has been considered in breadth and depth within the statistical literature, primarily in the context of inference (Josse and Reiter, 2018; Little et al., 2019). Approaches fall broadly into two camps: likelihood-based and imputation-based. Likelihood-based methods, such as those for binomial regression (Ibrahim and Lipsitz, 1996) or Generalised Linear Models (Ibrahim et al., 2001) use Expectation Maximisation (EM) (Dempster et al., 1977).

Multiple Imputation is the most common imputation-based approach within the ML context (Bertsimas et al., 2018), proposing to replace the missing covariates with a function of the non-missing data. This is done several times non-deterministically to account for variance. Within supervised learning, Josse et al. (2024); Bertsimas et al. (2018) examine so-called Impute-then-Regress methods. The work by Morvan et al. (2024) generalises these results, showing that Impute-then-Regress procedures are asymptotically Bayes consistent for all missing data mechanisms and almost all imputation functions, regardless of the distribution of $\langle X, Y \rangle$ and the number of missing covariates. We note that these works assume only the covariates X have missing data, whereas we are focused on missing labels Y. These asymptotic, general results are supplemented by works (Ayme et al., 2024; Lobo et al., 2024) that specialise to specific models and missingness mechanisms to obtain sharpened bounds.

Focusing on missing labels, Semi-supervised Learning (SSL) (Scudder, 1965) leverages unlabelled alongside labelled data (Chapelle et al., 2006) to improve performance on supervised learning tasks. Work on SSL with awareness of the causal diagram of missingness mechanisms is in its infancy. Notable examples include the doubly robust methods of Hu et al. (2022) and Sportisse et al. (2023), where they attempt direct modelling of the missingness mechanism. The SSL literature generally assumes access to a fully labelled test dataset, whereas we are specifically interested in the distribution over evaluation metrics induced by the uncertainty over missing test labels.

Active Learning (Settles, 2009; Hino, 2020) considers the scenario where there are few labels and the decision-maker simultaneously learns a model and a sample-efficient policy for querying an oracle to gain more labels for learning. Mishler et al. (2023) consider active learning in the MNAR setting for training rather than evaluation.

Within the Computer Vision community, there is work on accounting for missing test labels on tasks that are not classification. For example, ranking models on unknown domains (Sun et al., 2021) and for domain adaptive instance segmentation (Guan and Yuan, 2024).

Amoukou et al. (2024) provide estimated quantiles of an ML model's loss function at test-time when labels are absent, towards solving the Sequential Harmful Shift Detection problem. These estimates are used in a sequential testing framework to alert when the test distribution differs from the training distribution.

To the authors' knowledge this work is the first focusing specifically on the predictive distribution for classification metrics in the presence of missing labels.

2 PROBLEM SETUP

In this paper we consider binary classification, in which labels Y might be missing and input features X are complete. Data $\langle X, Y, M \rangle \in \mathcal{X} \times \mathcal{Y} \times \{0, 1\}$ are drawn iid (independently and identically distributed) from distribution $\mathcal{D}$, and $\mathcal{Y} = \{0, 1\}$. The observed label is denoted by $Y^* \in \mathcal{Y} \cup \{\texttt{NA}\}$ and $M = 1$ indicates a missing label, that is

$$Y^* = \begin{cases} Y, & M = 0; \\ \texttt{NA}, & M = 1. \end{cases}$$

We assume that there is a trained model $\hat{y} : \mathcal{X} \to [0, 1]$, where the output $\hat{y}(x)$ is interpreted as the model's

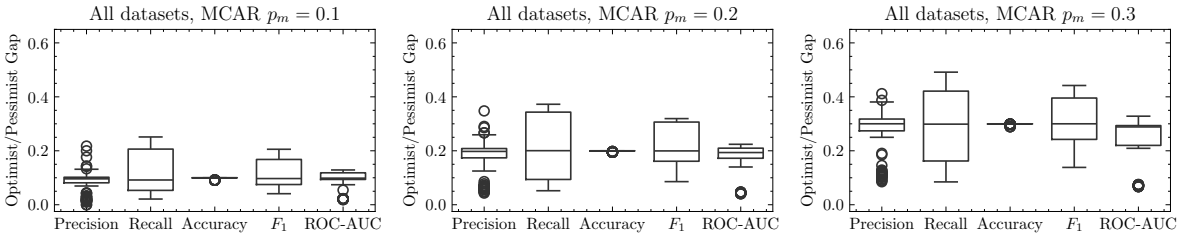

Figure 2: Gap between optimistic and pessimistic bounds on classifier performance measures $\widehat{Q}_n^{(S)}$. Each plot (left-to-right) corresponds to a fraction of missing labels $p_m \in \langle 0.1, 0.2, 0.3 \rangle$.

subjective probability that the label associated with $x \in \mathcal{X}$ is 1. Equipped with a threshold $\tau \in (0,1)$, $\hat{y}$ induces a classifier $\psi : \mathcal{X} \to \{0,1\}$ by returning $\psi(x) = 1$ if $\hat{y}(x) \geq \tau$ and $\psi(x) = 0$ otherwise.

Suppose we have a validation set, $D_n = \{\langle x_i, y_i^*, m_i \rangle\}_{i=1}^n$, drawn iid from $\mathcal{D}$. We wish to use this validation set to produce unbiased estimates of some performance metric:

$$\mathcal{Q}(\hat{y}) = \mathbb{E}_{\mathcal{D}}[Q(X, Y, \hat{y}(X))]$$

where Q is a probabilistic query[1] corresponding to the metric. For example, accuracy has the probabilistic query $Q_{\text{acc}} = P(Y = \psi(X))$. We denote an estimator admitted by $\mathcal{Q}$ using the validation set D_n as $\widehat{Q}_n(\hat{y})$, or simply $\widehat{Q}_n$, when $\hat{y}$ is clear from context. For the accuracy example, when there are no missing labels, that is, $m_i = 0 \Rightarrow y_i^* = y_i$ for all $i \in \{1, \ldots, n\}$, we have:

$$\begin{aligned} \widehat{Q}_{\text{acc},n} &= \tfrac{1}{n} \textstyle\sum_{i=1}^n \mathbb{1}\{y_i = \psi(x_i)\} \qquad (1) \\ &= \tfrac{1}{n} \Big(\textstyle\sum_{i:y_i=0} \mathbb{1}\{\psi(x_i) = 0\} + \sum_{i:y_i=1} \mathbb{1}\{\psi(x_i) = 1\} \Big) \end{aligned}$$

where $\mathbb{1}\{E\}$ is the indicator function for an event E. Eq. (1) equals the traditional formula for accuracy:

$$\tfrac{1}{n}(\text{True Positives} + \text{True Negatives}).$$

If there is no missing data, i.e. $M = 0$ always, then $\mathcal{Q}$ usually admits an *unbiased*, *consistent* estimator using the validation set D_n. That is,

$$\mathbb{E}_{\mathcal{D}}\Big[\widehat{Q}_n(\hat{y})\Big] = \mathcal{Q}(\hat{y}) \quad \text{and}$$

$$\lim_{n\to\infty} \mathbb{P}_{\mathcal{D}}\Big[\Big|\widehat{Q}_n(\hat{y}) - \mathcal{Q}(\hat{y})\Big| > \epsilon\Big] = 0 \quad \text{for all} \quad \epsilon > 0.$$

It is straightforward to show that the vanilla estimator for accuracy, $\widehat{Q}_{\text{acc},n}$, is both unbiased and consistent in the nonmissing labels setting, which for completeness we demonstrate in Proposition 5 (Appendix).

[1]See Appendix A for exposition on probabilistic queries.

Missing Labels. Often in practise we have missing data labels in our validation set D_n, that is, $M \neq 0$ in general when data is drawn from $\mathcal{D}$. Unless the labels are MCAR then computing $\widehat{Q}_n$ using the data $(D_n \,|\, M = 0)$, that is, ignoring datapoints with missing labels, will in general give a biased estimate for $\mathcal{Q}$ (Williams, 2015). We wish to produce an unbiased estimate for $\mathcal{Q}$, or better yet, a distribution over plausible values centred on an unbiased estimate.

For simplicity, without loss of generality we assume that the first k labels within D_n are known, giving the set of known labels $\mathcal{K} = \{1, 2, \ldots, k\}$ and the remaining indices as masked labels $\overline{\mathcal{K}} = \{k+1, k+2, \ldots, n\}$.

3 NAÏVE ESTIMATION

For a given D_n there are 2^{n-k} possible outcomes describing the true pattern of labels $y_1, \ldots, y_n$. The set of potential scenarios is defined as $\mathcal{S} = \big\{\langle y_1, \ldots, y_k, s'_1, \ldots s'_{n-k}\rangle \;|\; s'_j \in \{0,1\} \text{ for all } j \in \{1, \ldots, n-k\}\big\}$. In words, $\mathcal{S} \subset \{0,1\}^n$ is the set of label assignments to D_n consistent with the observed labels. Each $s \in \mathcal{S}$ induces imputed evaluation data $D_n^{(s)}$ by substituting elements of the scenario bitstring s as realised values of Y in D_n. Subsequently, an estimator assuming no missing data $\widehat{Q}_n$ is computed on the imputed dataset $D_n^{(s)}$, yielding the imputed estimator $\widehat{Q}_n^{(s)}$. Denote by S the random variable describing our ignorance of the true scenario.

We will ask two questions, namely:

1. What are the upper and lower bounds on $\widehat{Q}_n^{(S)}$?
2. What is the probability distribution over $\widehat{Q}_n^{(S)}$ describing our ignorance of $\widehat{Q}_n$?

From the iid assumption on $\mathcal{D}$, we can surmise that the distribution over scenarios is a product distribution over independent label distributions. We define the random variable Y_i corresponding to the decision-maker's knowledge of the label of a point in D_n like

so:

$$Y_i \sim \begin{cases} \mathcal{B}(p_i), & i \in \overline{\mathcal{K}}; \\ \mathbb{1}\{Y_i = y_i\}, & i \in \mathcal{K}, \end{cases} \tag{2}$$

where $\mathcal{B}(p)$ is a Bernoulli distribution with parameter $p \in (0,1)$ and the $y_i \in \{0,1\}$ are constants for $i \in \mathcal{K}$. The parameters p_i for $i \in \overline{\mathcal{K}}$ represent the conditional probability $\mathbb{P}[Y_i = 1 \,|\, X = x_i, M = 1]$. The scenario random variable $S = \langle Y_1, Y_2, \ldots, Y_n \rangle$ has the product distribution $\mathbb{P}[S = s] = \mathbb{P}[Y_1 = s_1] \cdots \mathbb{P}[Y_n = s_n]$ for $s = s_1 s_2 \cdots s_n$.

Computational Considerations. When there are few missing test labels, i.e. $n - k \lesssim 15$, one can enumerate all scenarios $s \in \mathcal{S}$. In this case it is straightforward to answer question 1; we obtain lower and upper bounds by sorting the values in $\{\widehat{Q}_n^{(s)}\}_{s \in \mathcal{S}}$. The number of scenarios scales exponentially in the number of missing labels $n - k$, precluding this brute force approach in practical problems.

Bounds on Imputed Estimator. For a classifier $\psi : \mathcal{X} \to \{0,1\}$ it is straightforward to produce optimistic and pessimistic bounds for any imputed metric estimator $\widehat{Q}_n^{(S)}$. Indeed for all missing datapoints $i \in \overline{\mathcal{K}}$, we imagine that an oracle can choose the scenario s that takes place, substituting $s_i = \psi(x_i)$ for the optimistic bound and $s_i = 1 - \psi(x_i)$ for the pessimistic bound. The oracle evaluates the metric $\widehat{Q}_n^{(s)}$ as one would in the nonmissing case. The optimistic bound corresponds to the case when the classifier correctly classifies every missing datapoint, the pessimistic bound is the opposite.

In Figure 2 we show the width of these bounds, evaluated in the MCAR setting on the same experimental setup as in Section 6. We plot the distribution of the widths of the gap between the optimistic and pessimistic values for: precision, recall, accuracy, F_1-score and ROC-AUC. Given that the stated metrics take values in the $[0,1]$ interval, these bounds are clearly too loose to be of use to a practitioner. Accordingly, we would prefer a distribution showing how *plausible* any particular value for our chosen metric is.

4 PERFORMANCE ESTIMATION BY MULTIPLE IMPUTATION

Recall from (2) that for test points with missing labels $i \in \overline{\mathcal{K}}$, Bernoulli distributions $\mathcal{B}(p_i)$ describe knowledge of the labels Y_i. This induces a distribution over the performance metric $\widehat{Q}_n^{(S)}$, suggesting a straightforward algorithm, Performance Estimation by Multiple Imputation (PEMI), to approximate $\widehat{Q}_n^{(S)}$ that is formalised in Algorithm 1. Namely, sample the relevant Bernoulli random variables yielding a bitstring s, evaluate $\widehat{Q}_n^{(s)}$, then repeat, recording the values in an empirical cdf (cumulative distribution function).

Algorithm 1 Performance Estimation by Multiple Imputation (PEMI)

Input: Metric estimator $\widehat{Q}_n$, nonmissing labels $\{y_i \mid i \in \mathcal{K}\}$, missing Bernoulli parameters $\{p_i \mid i \in \overline{\mathcal{K}}\}$, number of samples $B \in \mathbb{N}$.
Output: Empirical cdf $\widehat{F}$

$\widehat{F} \leftarrow (t \mapsto 0)$ ▷ Zero function
for $b \in \{1, \ldots, B\}$ **do**
  $s \leftarrow \langle\,\rangle$ ▷ $\langle\,\rangle$ is empty string
  for $i \in \langle 1, \ldots, n \rangle$ **do**
    if $i \in \overline{\mathcal{K}}$ **then**
      Sample $s_i \sim \mathcal{B}(p_i)$
      $s \leftarrow s \mathbin{+\!\!+} \langle s_i \rangle$ ▷ $+\!\!+$ is string concatenation
    else
      $s \leftarrow s \mathbin{+\!\!+} \langle y_i \rangle$
  $\widehat{F} \leftarrow \widehat{F} + \left(t \mapsto \frac{1}{B} \mathbb{1}\{\widehat{Q}_n^{(s)} \le t\}\right)$
return $\widehat{F}$

Assignment of Bernoulli Parameters for PEMI. In practise, it is unlikely that we would know the true values of the $\{p_i\}$. However, if we have a *well-calibrated* model for these probabilities then these can be substituted in lieu of the true Bernoulli parameters into Algorithm 1. A model being well-calibrated means that the predicted probability of an event reflects the true frequency of the event's occurrence (DeGroot and Fienberg, 1983). For example, consider a model predicting if it will rain or not. Suppose 1000 inputs to this model receive a predicted 20% chance of rain. Then out of these, roughly 200 will observe rain if the model is well-calibrated.

By default, many modern ML models are not well-calibrated (Zadrozny and Elkan, 2001; Guo et al., 2017; Kuleshov et al., 2018). In this work we use the scaling-binning calibrator of Kumar et al. (2019) to provide estimates of the $\{p_i\}$ for use within Algorithm 1 to give an empirical distribution of $\widehat{Q}_n^{(S)}$. At the time of writing, there is early work (Kweon and Yu, 2024; Gong and Ma, 2025) on MNAR-aware model calibration in the context of recommender systems, but for our purposes the scaling-binning calibrator is sufficient.

If one wishes to be less committal in the assumptions placed upon the missing label distributions, one can invoke the *Maximum Entropy Principle* (MaxEnt) of Jaynes (1957), whereby the distribution describing our uncertainty about a random variable is the one maximising Shannon entropy, subject to testable information. Within our setting, where the Y_i take values in $\{0,1\}$ and are drawn iid, we have $p_i = \frac{1}{2}$ when no constraints are imposed. We may also impose that

Table 1: CM-based classification metrics with associated probabilistic queries.

Metric	CM Formula	Probabilistic Query
Precision	$\frac{\mathrm{TP}}{\mathrm{TP+FP}}$	$\frac{P(Y=1,\psi(X)=1)}{P(Y=1,\psi(X)=1)+P(Y=0,\psi(X)=1)} = \frac{P(Y=1,\psi(X)=1)}{P(\psi(X)=1)}$
Recall	$\frac{\mathrm{TP}}{\mathrm{TP+FN}}$	$\frac{P(Y=1,\psi(X)=1)}{P(Y=1,\psi(X)=1)+P(Y=1,\psi(X)=0)}$
Accuracy	$\mathrm{TP} + \mathrm{TN}$	$P(Y=1,\psi(X)=1) + P(Y=0,\psi(X)=0)$
F_1 Score	$\frac{2\mathrm{TP}}{2\mathrm{TP+FP+FN}}$	$\frac{2P(Y=1,\psi(X)=1)}{2P(Y=1,\psi(X)=1)+P(Y=0,\psi(X)=1)+P(Y=1,\psi(X)=0)}$

Table 2: CM estimators. For a variable $z \in \{0,1\}$, let $\overline{z} = 1 - z$. Recall $Y_i \sim \mathcal{B}(p_i)$ for all $i \in \overline{\mathcal{K}}$.

CM	Estimator
TP	$\sum_{i=1}^n \mathbb{1}\{y_i = 1 \wedge \psi_i = 1\} = \sum_{i\in\mathcal{K}} y_i\psi_i + \sum_{i\in\overline{\mathcal{K}}} Y_i\psi_i$
FN	$\sum_{i=1}^n \mathbb{1}\{y_i = 1 \wedge \psi_i = 0\} = \sum_{i\in\mathcal{K}} y_i\overline{\psi_i} + \sum_{i\in\overline{\mathcal{K}}} Y_i\overline{\psi_i}$
FP	$\sum_{i=1}^n \mathbb{1}\{y_i = 0 \wedge \psi_i = 1\} = \sum_{i\in\mathcal{K}} \overline{y_i}\psi_i + \sum_{i\in\overline{\mathcal{K}}} \overline{Y_i}\psi_i$
TN	$\sum_{i=1}^n \mathbb{1}\{y_i = 0 \wedge \psi_i = 0\} = \sum_{i\in\mathcal{K}} \overline{y_i}\overline{\psi_i} + \sum_{i\in\overline{\mathcal{K}}} \overline{Y_i}\psi_i$

the mean over the training set and test set are equal, giving $p_i = N_+/N$ where N is the total number of examples in the training set and N_+ the number of positives. Choosing p_i as the calibrated probability estimate is equivalent to imposing $X_i = x_i$ as testable constraints in the MaxEnt framework.

Performance Metrics. We shall see that for several common classification metrics, $\widehat{Q}_n^{(S)}$ is either a weighted sum of Bernoulli random variables or a ratio thereof. In both cases, we are able to show that the distribution is approximately Gaussian (Lemma 1 and Theorem 3). The first case is a standard result; the ratio of sums of Bernoulli variables is novel to the authors' knowledge. We also show in Theorem 4 that these results are robust under a realistic noise model for the calibrator.

Confusion matrix-based metrics. In this work, we mainly focus on evaluation metrics based on the confusion matrix (CM), as defined below.

$$\begin{aligned}\mathrm{CM} &= \begin{bmatrix} \text{True Positives} & \text{False Negatives} \\ \text{False Positives} & \text{True Negatives} \end{bmatrix} = n \cdot \begin{bmatrix} \mathrm{TP} & \mathrm{FN} \\ \mathrm{FP} & \mathrm{TN} \end{bmatrix} \\ &= n \cdot \begin{bmatrix} P(Y=1,\psi(X)=1) & P(Y=1,\psi(X)=0) \\ P(Y=0,\psi(X)=1) & P(Y=0,\psi(X)=0) \end{bmatrix}\end{aligned}$$

The elements of the CM may be thought of as the product of the number of evaluation points n with an estimator of certain probabilistic queries using D_n. Common classification metrics comprise simple arithmetic combinations of elements of the CM, and as such correspond to probabilistic queries too. These probabilistic queries are shown in Table 1. In Table 2 we show the estimators of the CM elements. When $\overline{\mathcal{K}} = \emptyset$ we recover the nonmissing estimators for the CM elements. When $\overline{\mathcal{K}} \neq \emptyset$ the estimators of the CM elements are now random variables. We list their means and covariance matrix in Lemma 6 in the appendix.

Thus, under the iid assumption, substitution of the CM estimators from Table 2 into Table 1 leads us to the following: The classification metrics precision, recall, accuracy and F_1-score are distributed as sums of Bernoulli random variables or (correlated) ratios thereof.

Rank-based metrics. There are other performance metrics for binary classifiers that do not depend on the confusion matrix, but instead of the *ranking* between data points, that is, how $\hat{y}(x)$ and $\hat{y}(x')$ compare with one another for $x, x' \in \mathcal{X}$. One such metric is the area under the curve of the receiver operating characteristic (Hanley and McNeil, 1983; Bradley, 1997), denoted as ROC-AUC, which we shall also focus on in this work. Indeed, ROC-AUC corresponds to an estimation of the following probabilistic query

$$P(\hat{y}(X) \geq \hat{y}(X') \mid Y = 1,\ Y' = 0), \tag{3}$$

where $\langle X, Y\rangle$ and $\langle X', Y'\rangle$ are understood to come from independent draws from $\mathcal{D}$ (Bradley, 1997). We describe the estimator $\widehat{Q}^{(S)}_{\text{roc-auc},n}$, in Eq. (53) and its mean and variance in Remark 14 (relegated to Appendix for brevity).

5 APPROXIMATE DISTRIBUTIONS

Theoretical results in the sequel have proof in Appendix Section B.

Sums of Bernoulli Random Variables. It is well-known that a sum of Bernoulli random variables under certain conditions is well-approximated by a Gaussian distribution. To be more precise, "well-approximated" means close in Kolmogorov-Smirnov (KS) distance. For random variables U, V with cdfs F_U, F_V, the KS distance between their distributions is given by

$$\mathrm{d}_{\mathsf{KS}}(F_U, F_V) = \sup_{t\in\mathbb{R}} |F_U(t) - F_V(t)|.$$

We can bound the KS distance of a weighted sum of Bernoulli variables to an appropriately-scaled Gaussian, denoted by $\mathcal{N}(\mu, \sigma^2)$ using Lemma 1.

Lemma 1. *Let $Z = \sum_{i=1}^n a_i Y_i$, where $Y_i \sim \mathcal{B}(p_i)$ and mutually independent, $p_i \in (0,1)$ and $a_i \in \mathbb{R}\setminus\{0\}$. Then, we have that*

$$\mathrm{d}_{\mathsf{KS}}(F_Z, \mathcal{N}(\mu_z, \sigma_z^2)) \leq \frac{C_0}{\sqrt{nv_*}} \frac{1 + a^*}{2a_*}, \quad \textit{where}$$

$$v_* = \min_i\{p_i(1-p_i)\},\quad a_* = \min_i\{|a_i|\},\ a^* = \max_i\{|a_i|\},$$
$$\mu_z = \textstyle\sum_{i=1}^n a_i p_i,\ \sigma_z^2 = \sum_{i=1}^n a_i p_i(1-p_i)$$

and $C_0 = 0.5600$ is a universal constant.

Consultation of Tables 1 and 2 tells us that the estimators for Accuracy and Precision take the format $Z + \text{const.}$ (with $a_* = a^* = 1$), leading to $O(n^{-1/2})$ scaling in the KS distance to a Gaussian (for fixed v_*) from Lemma 1.

Ratios of Normal Variables. For the remaining CM-based metrics – Recall and F_1 score – we see from Tables 1 and 2 that each will be distributed according to a ratio of (correlated) sums of Bernoullis. From Lemma 1 we can treat the sums in the numerator and denominator as Gaussian random variables and consider the distribution of the ratio. It is known (Marsaglia, 1965) that the ratio of correlated normal random variables is poorly behaved in general, in that no moments of the resulting distribution exist. Nonetheless, it has been shown empirically that in many cases this distribution is approximately normal itself (Marsaglia, 2006). We present the first rigorous bounds on the distance of this ratio distribution from normality.

Theorem 2 (Ratio of Correlated Gaussians)**.** *Let Z and W be jointly Gaussian random variables, i.e. $(Z, W) \sim \mathcal{N}(m, \Sigma)$, where $m = [\mu_z, \mu_w]^\mathsf{T}$ and $\Sigma = \begin{bmatrix} \sigma_z^2 & \rho\sigma_z\sigma_w \\ \rho\sigma_w\sigma_z & \sigma_w^2 \end{bmatrix}$. Moreover, let the the distribution of $T := Z/W$ be described by $G(t) = \mathbb{P}[T \leq t]$. Then, under conditions*

1. *$\mathbb{P}(W > 0) \to 1$,. i.e. $\mu_w/\sigma_w \gg 1$;*
2. *$2|\mu_z|\left|\frac{\sigma_w}{\mu_w} - \frac{\sigma_z}{\mu_z}\right| \gg 1$;*

the distribution of Z/W satisfies

$$\mathrm{d}_{\mathsf{KS}}(G, \mathcal{N}(\mu, \sigma^2)) \leq \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_w^2(|\mu_z| + \sigma_z^2) + \mu_w^2}{\sigma\mu_w^3},$$

where $\mu = \frac{\mu_z}{\mu_w}$, $\sigma^2 = \frac{{\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\,\rho\sigma_z\sigma_w\mu_z\mu_w}{{\mu_w}^4}$.

In Remark 9 (Appendix) an interpretation of each term in the bound of Theorem 2 is given.

Ratio Distribution of Sum of Weighted Bernoulli Variables. To show a rigorous bound on the distance of the predictive distributions of Recall and F_1-score, we cannot assume the numerator and denominator are Gaussian, but explicitly consider them as sums of Bernoulli random variables. We combine the result of Theorem 2 with Lemma 1 to prove Theorem 3, an explicit bound on KS distance from a normal distribution applicable to these performance metrics.

Theorem 3 (Gaussian Approximation)**.** *Consider the random variables*

$$\widetilde{Z} = \alpha + \textstyle\sum_{i=1}^n a_i Y_i, \quad \widetilde{W} = \beta + \sum_{i=1}^n b_i Y_i$$

for $0 < \alpha \leq \beta$, $a_i \geq 0$, $b_i > 0$, $b_i \geq a_i$, $Y_i \sim \mathcal{B}(p_i)$ and mutually independent for $p_i \in (0, 1)$. Moreover,

Algorithm 2 Performance Estimation by Multiple Imputation - Gaussian Approximation (PEMI-Gauss)

Input: Metric estimator $\widehat{Q}_n$, nonmissing labels $\{y_i \mid i \in \mathcal{K}\}$, missing Bernoulli parameters $\{p_i \mid i \in \overline{\mathcal{K}}\}$.
Output: Estimated CDF $\widehat{F}$
1: Compute Gaussian params μ, σ^2 using Table 1, Table 2, Lemma 6, Theorem 3 and Remark 14.
2: $\widehat{F} \leftarrow \left(t \mapsto \Phi(\frac{t-\mu}{\sigma})\right)$ ▷ Φ is the standard normal cdf
3: **return** $\widehat{F}$

$v_* = \min_i\{p_i(1-p_i)\}$, $a_* = \min_i\{|a_i| \mid a_i > 0\}$, $b^* = \max_i\{|b_i|\}$ and $n_a = |\{a_i \mid a_i > 0\}| \leq n$. The ratio distribution $\widetilde{G}(t) = \mathbb{P}[\widetilde{Z}/\widetilde{W} \leq t]$ is approximately Gaussian, that is

$$\mathrm{d}_{\mathsf{KS}}(\widetilde{G}, \mathcal{N}(\mu, \sigma)) \leq \frac{C_0}{\sqrt{n_a v_*}} \frac{1 + b^*}{a_*} + \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_w^2(|\mu_z| + \sigma_z^2) + \mu_w^2}{\sigma\mu_w^3},$$

where $C_0 = 0.5600$ is a universal constant and

$$\mu_z = \alpha + \textstyle\sum_{i=1}^n a_i p_i, \quad \sigma_z^2 = \sum_{i=1}^n a_i^2 p_i(1-p_i),$$
$$\mu_w = \beta + \textstyle\sum_{i=1}^n b_i p_i, \quad \sigma_w^2 = \sum_{i=1}^n b_i^2 p_i(1-p_i),$$
$$\rho = \tfrac{1}{\sigma_z\sigma_w} \textstyle\sum_{i=1}^n a_i b_i p_i(1-p_i),$$
$$\mu = \tfrac{\mu_z}{\mu_w}, \quad \sigma^2 = \tfrac{{\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\,\rho\sigma_z\sigma_w\mu_z\mu_w}{{\mu_w}^4}.$$

The bound in Theorem 3 is a finite-sample bound. It is not immediately obvious from the bound what the asymptotic behaviour is from the functional form. The assumptions are too general to provide the asymptotic behaviour on all cases – consider, e.g. $p_i \to 0$ for some $i \in [n]$, which can make the bound arbitrarily large. Nonetheless, as a coherence check, we can make several small assumptions and recover $O(n^{-1/2})$ asymptotic behaviour, as shown in Corollary 11 (Appendix).

Gaussian Approximation for ROC-AUC. For ROC-AUC, the probabilistic query is also a ratio of sums of Bernoullis (see Eq. 53 in Appendix for details), but now there is dependence between the elements of the sums, such that we cannot directly invoke Berry-Esseen type bounds as in Lemma 1. As a corollary the bound of Theorem 3 doesn't necessarily hold. Under certain mixing properties (Bradley, 2007; Dedecker et al., 2007; Doukhan, 2012) likely satisfied here, one can derive similar convergence bounds – see the paper by Tikhomirov (1981) for a classical example. For the ROC-AUC example this is very involved and delegated to future work. Nonetheless, we use the mean and variance computed in Remark 14 (Appendix).

PEMI-Gauss Algorithm. We formalise approximating the limiting behaviour of the PEMI algorithm in Algorithm 2, Performance Estimation by Multiple

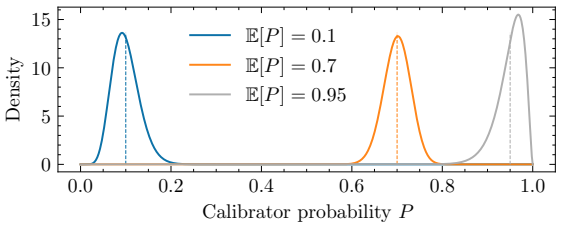

Figure 3: Illustration of flexibility afforded by Theorem 4. For $\mathbb{V}[P] = 0.0009$, we show the allowed variation in a calibrator's output P when the expectation $\mathbb{E}[P]$ is fixed to the given true Bernoulli probabilities $p \in \{0.1, 0.7, 0.95\}$.

Imputation - Gaussian Approximation (PEMI-Gauss). PEMI-Gauss works with guarantees for any CM-based performance metric, but may also be applied whenever a mean and variance can be computed. PEMI-Gauss offers advantages over the empirical cdf provided by PEMI in terms of simplicity, interpretability, analytical convenience, and computational efficiency, in that we no longer generate $B \cdot (n-k)$ random bits.

Robustness. The performance guarantees offered by PEMI-Gauss to CM-based classification metrics implicitly assume that we have *perfectly calibrated* Bernoulli parameters, which is unrealistic for real-world use cases. We are able to prove that the same guarantees hold when the calibrated Bernoulli parameters are subject to a realistic noise model.

Theorem 4 (Robustness)**.** *The conclusion of Theorem 3 remains valid if we relax the assumption that $Y_i \sim \mathcal{B}(p_i)$ for $p_i \in (0,1)$ to the following: the random variables $Y_i \sim \mathcal{B}(P_i)$, where $P_i \sim \text{Beta}(\alpha_i, \beta_i)$ such that $\mathbb{E}[P_i] = p_i$ and $\mathbb{V}[P_i] < v_*$ for all $i \in [n]$.*

Theorem 4 is a formal assertion that using PEMI-Gauss with a calibrator that is correct *on average* results in correct predictive distributions. As a realistic illustration: suppose that in the evaluation set the smallest (resp. largest) true Bernoulli parameter is $p_* = 0.0001$ (resp. $p^* = 0.9999$). The maximum variance allowed by Theorem 4 of the underlying Beta distributions is therefore $v_* = 0.000999$. For a Bernoulli parameter p, we can plot the pdf (probability density function) of the distribution of $P \sim \text{Beta}(\alpha_p, \beta_p)$ corresponding to imposing $\mathbb{E}[P] = p$ and $\mathbb{V}[P] = 0.0009 < v_*$ (which uniquely defines α_p, β_p). Consider the input data $X_p = \{X \mid \mathbb{P}[Y = 1 \mid X,\ M = 1] \approx p\}$, that has missing labels and true Bernoulli parameter p. If the calibrated predictions follow the distribution of the random variable P over X_p then the guarantee of Theorem 4 holds. We plot in Figure 3 the density of P for several values of p and note that the spread of the distributions is quite generous, that is, the calibrator can err significantly and Theorem 4 ensures the distribution output by PEMI-Gauss has high fidelity.

6 EXPERIMENTS

Recall that PEMI and PEMI-Gauss return a predictive distribution for estimators $\widehat{Q}_n^{(S)}$ when labels are missing in the evaluation set D_n. We study the quality of these algorithms along the following dimensions:

1. Centre and shape of predictive distribution.
2. Effectiveness across missingness mechanisms.

Setup. For a given dataset $D = \{\langle x_1, y_1\rangle, \ldots, \langle x_N, y_N\rangle\}$ we partition randomly into 10 stratified folds $D^1, \ldots, D^{10}$, that is, the fraction of positive labels N_+/N is the same in each fold D^u and the same as in D. Within a training set $D \setminus D^u$, 90% of the data is randomly chosen to train an XGBoost (Chen and Guestrin, 2016) model[2] $\hat{y}$ and the remaining 10% is chosen to fit a scaling-Binning calibrator (Kumar et al., 2019) $\hat{c} : [0,1] \to [0,1]$ with 10 bins (default). Each test fold D^u is subsequently split into two further random (non-stratified) sub-folds, $D^{u,1}$ and $D^{u,2}$. The same model trained and calibrated on $D \setminus D^u$ is evaluated on two test sets: *i.* D^u with labels missing from $D^{u,1}$; and *ii.* D^u with labels missing from $D^{u,2}$. For a given dataset this gives 20 replications in total and ensures that the missing labels are independent of one another.

We consider missingness proportions $p_m \in \{0.1, 0.2, 0.3\}$, masking labels in two settings: *i.* **MCAR.** Randomly sample a fraction p_m from the relevant sub-fold $D^{u,\cdot}$; *ii.* **MNAR.** Randomly sample a fraction p_m from the relevant sub-fold $D^{u,\cdot}$, ensuring a fraction $\eta \in (0,1)$ has a positive label. For masked datapoints $i \in \overline{\mathcal{K}}$, we reserve the true labels y_i for evaluation and set $y_i^* = \texttt{NA}$ in the test sets $D^{u,\cdot}$. For the MNAR experiments, we independently consider $\eta \in \{0.1, 0.2, 0.4, 0.6, 0.8, 0.9\}$.

For each sub-fold $D^{u,\cdot}$, there are three ways we choose the Bernoulli parameters input to PEMI and PEMI-Gauss for all $i \in \overline{\mathcal{K}}$.

1. Set $p_i = \frac{1}{2}$ (MaxEnt – no testable constraints);
2. Set $p_i = \frac{N_+}{N}$ (MaxEnt – fixed mean);
3. Set $p_i = \hat{c}(\hat{y}(x_i))$ (calibrated);

The six datasets under consideration are Dota2 Games Results (Tridgell, 2016), IMDB.drama (Read et al., 2012; IMDB, 2024), Bank Marketing (Moro and Cortez, 2014), Diabetes (Kahn, 1994), German Credit (Hofmann, 1994) and Adult (Becker and Kohavi, 1996).

Evaluation. We turn to the forecasting literature for assessing the predictive distribution of $\widehat{Q}_n^{(S)}$. Prob-

[2]Default hyperparameters are used with Categorical mode enabled.

ability Integral Transform (PIT) is the idea that given a random variable U with cdf F_U, the random variable $V = F_U(U)$ has a standard uniform distribution, which we denote by $\mathcal{U}[0,1]$. Diebold et al. (1998) show that for any sequence of random variables $T_1, T_2, ...$ with no forward dependence – formally, $T_r \perp \langle T_{r+1}, T_{r+2}, \ldots\rangle \mid \langle T_1, T_2, \ldots, T_r\rangle$ for all r – then $F_{T_1}(T_1), F_{T_2}(T_2), \ldots \sim \mathcal{U}[0,1]$ independently. This result is used to evaluate forecast distributions $\widehat{F}_r$, by evaluating on realisations of T_r, that is, computing $\widehat{F}_r(t_r)$ and comparing the empirical cdf (aggregated over r) visually against the cdf of a $\mathcal{U}[0,1]$ random variable[3]. Let $\widehat{Q}^u$ be the estimator $\widehat{Q}(\hat{y})$ evaluated on D^u, i.e. the ground truth for fold u. From Diebold et al. (1998) we expect the correct distribution of $\widehat{Q}_n^{(S)}$ evaluated at the ground truth values, $F_{\widehat{Q}_n^{(S)}}(\widehat{Q}^u)$, to be distributed according to $\mathcal{U}[0,1]$, for any dataset, fold or metric.

For metrics {Precision, Recall, Accuracy, F_1-score, ROC-AUC} we compute 120 ground-truth quantities $\widehat{Q}^u$. Then, for each predictive distribution $\widehat{F}_{\mathcal{A}}$ provided by algorithm $\mathcal{A}$ we evaluate $\widehat{F}_{\mathcal{A}}(\widehat{Q}^u)$ and record the values in an empirical cdf $\widehat{F}_{\mathcal{A}}^{\text{PIT}}$. The distance of $\widehat{F}_{\mathcal{A}}^{\text{PIT}}$ is then measured against $\mathcal{U}[0,1]$ in the following ways: Wasserstein-1 (W_1) distance; and KS distance, where for real-valued r.v.s U, V we have

$$\mathrm{d}_{W_1}(F_U, F_V) = \int_{-\infty}^{+\infty} |F_U(t) - F_V(t)|\,\mathrm{d}t\,.$$

The W_1 and KS distances capture the quality of the shape of the predictive distributions. We also measure the central tendency of the predicted distribution by root mean-squared error (RMSE) and mean absolute error (MAE). For all distances lower values are better.

Baselines. We are not aware of any prior works studying this problem, so as a baseline we use the bootstrap distribution of the estimator $\widehat{Q}^u$ for a given metric on the nonmissing data only, with $B = 10000$ replicates as recommended by Hesterberg (2015).

Results. In Tables 3 and 4 we show the W_1 and MAE fidelity measures for the MCAR setting when the fraction of missing labels p_m is 30% of the test set. In all cases we see that either PEMI or PEMI-Gauss with calibrated $\{p_i\}$ is the best choice. Tables 14 and 15 (Appendix) show the calibrated multiple-imputation based methods to be mostly overlapping at the $\alpha = 0.9$ confidence level. Note that the Gaussian approximation works well, in that the numbers for PEMI-Gauss are similar or better than the PEMI. The effect of poor calibration on the multiple imputation methods is stark, showing significantly diminished performance as compared with their calibrated counterparts. In the MCAR setting, we would expect that performance measures focusing on the location of the predictive distribution, such as MAE and MSE, will show little difference between the PEMI methods and bootstrap. This is because there is no bias in the missing data. On the other hand, we would expect that for measures considering shape also – namely KS and W_1 distance – for there to be some impact on performance due to ignoring missing data. Indeed this is largely borne out in the tables in Appendix Section C.1.

Figure 4 shows the effect of varying the MNAR class imbalance η on the predictive quality of PEMI-Gauss and the bootstrap distribution the for W_1 and MAE distances. There is an exception for the Precision metric and W_1-distance with $\eta \in \{0.1, 0.2\}$ – here the uncalibrated models outperform. The shape of the predictive distribution for the Precision metric using well calibrated Bernoulli probabilities is actually worse for small η than more crude estimates for the p_i. In the parlance of Section 4, Precision measures TP/P. P is fixed by the classifier outputs ψ_i so it is the shape of the distribution of $\text{TP} = \sum_{i\in\mathcal{K}} y_i\psi_i + \sum_{i\in\overline{\mathcal{K}}} Y_i\psi_i$ that controls the Precision distribution. Here, Y_i are more likely to be zero due to small η. The result is that the calibrated probabilities p_i are overestimated and this drives the difference. The cruder estimates of p_i are less opinionated so give a smaller error.

[3] A plot $x = y$ on the Cartesian plane with $x, y \in [0,1]$.

Table 3: (Distribution shape) $\mathrm{d}_{W_1}(\widehat{F}_{\mathcal{A}}^{\text{PIT}}, \mathcal{U}[0,1])$ for different algorithms in MCAR setting, $p_m = 0.3$. Best algorithm in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		0.108729	0.089308	0.068415	0.089793	0.443732
PEMI $B = 5$	$p = \frac{1}{2}$	0.205000	0.361667	0.386667	0.380833	0.375000
	p calibrated	0.077500	0.072500	0.055833	0.074167	**0.085000**
	$p = N_+/N$	0.357500	0.364167	0.340833	0.355833	0.377500
PEMI $B = 10$	$p = \frac{1}{2}$	0.258750	0.417083	0.441667	0.411667	0.431667
	p calibrated	0.063750	0.047083	**0.035000**	0.032083	0.096250
	$p = N_+/N$	0.408333	0.405833	0.378333	0.412500	0.428333
PEMI $B = 100$	$p = \frac{1}{2}$	0.285333	0.450958	0.484583	0.469167	0.481917
	p calibrated	**0.030717**	**0.041592**	0.060375	**0.020367**	0.118675
	$p = N_+/N$	0.448625	0.453083	0.423167	0.458667	0.470542
PEMI-Gauss	$p = \frac{1}{2}$	0.264135	0.456714	0.484778	0.472469	0.485517
	p calibrated	0.058313	0.043286	0.042600	0.022275	0.129226
	$p = N_+/N$	0.437487	0.456702	0.418207	0.459145	0.477043

Table 4: (Distribution central tendency) MAE of centre of predictive distributions in MCAR setting, $p_m = 0.3$. Best algorithm in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		0.018494	0.015045	0.011883	0.013720	0.098082
PEMI $B = 5$	$p = \frac{1}{2}$	0.044845	0.103826	0.074209	0.092816	0.117412
	p calibrated	0.015533	**0.010099**	0.008119	0.010729	0.010893
	$p = N_+/N$	0.063529	0.062823	0.035999	0.061992	0.076707
PEMI $B = 10$	$p = \frac{1}{2}$	0.044080	0.102523	0.074850	0.091197	0.117487
	p calibrated	0.015096	0.011005	0.008198	0.009548	0.011014
	$p = N_+/N$	0.063623	0.061134	0.035974	0.061332	0.077073
PEMI $B = 100$	$p = \frac{1}{2}$	0.045993	0.102531	0.074734	0.092006	0.117845
	p calibrated	0.014831	0.010373	**0.007727**	0.009215	**0.010091**
	$p = N_+/N$	0.062079	0.060404	0.035468	0.062175	0.077259
PEMI-Gauss	$p = \frac{1}{2}$	0.045475	0.102887	0.074606	0.091768	0.117791
	p calibrated	**0.014629**	0.010520	0.007762	**0.009094**	0.010327
	$p = N_+/N$	0.062586	0.060855	0.035542	0.061928	0.077555

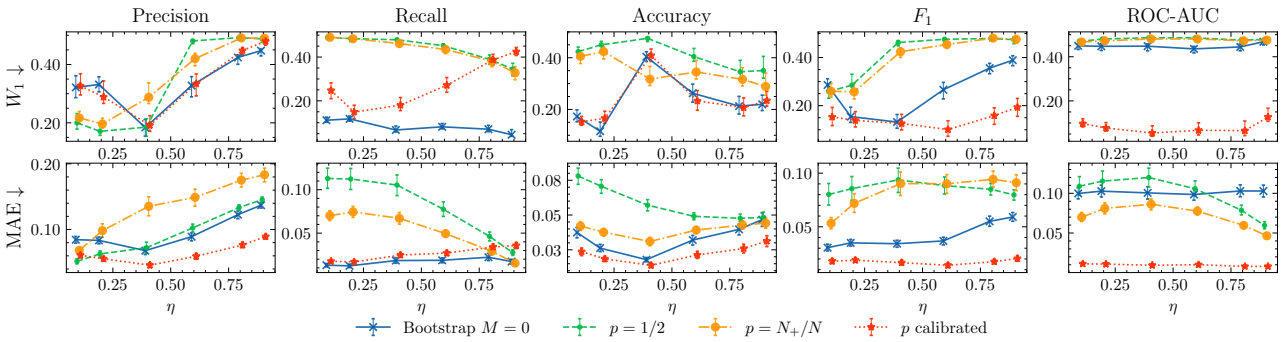

Figure 4: Effect of varying MNAR class imbalance $\eta \in (0, 1)$ on fidelity using PEMI-Gauss and baseline for $p_m = 0.3$. Each column represents a metric estimator, with the top row showing W_1-distance and bottom row showing MAE. Error bars are bootstrapped confidence intervals at the $\alpha = 0.9$ level. Fidelities are on y-axis (lower is better) with η varying on the x-axis; x-axis values are slightly jittered to help separate each series visually.

For the Recall metric, we have the bootstrap distribution of nonmissing data outperforming PEMI-Gauss. There is a similar reason for this as in the previous paragraph. Due to small η the PEMI predictive distribution is underestimating the true value. Recall is given by $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$. Overestimation in the denominator dominates overestimation in the numerator so we have that the overall predictive distribution underestimates.

The gap between the Bootstrap and PEMI-Gauss is less pronounced in the MNAR setting as compared with MCAR, but still significant, for all but the Recall metric where the bootstrap outperforms. For all estimation methods, performance in the MNAR setting is worse than MCAR as we may expect.

We show the results of these experiments in full for $p_m \in \{0.1, 0.2, 0.3\}$, alongside results for KS distance and RMSE in Appendix C. The data are qualitatively similar and the same conclusions can be drawn.

7 CONCLUSION

In this paper we demonstrate – via the PEMI and PEMI-Gauss algorithms – that multiple imputation can be successfully applied to model evaluation when test-time labels are missing. Crucially, the theoretical and empirical results in Sections 5 and 6 show that a well-calibrated imputation model is key to the success of this approach. We observe that state-of-the-art calibrators are sufficient for this purpose. However, PEMI and PEMI-Gauss are more effective with certain classification metrics, so we caution users of this method to evaluate on their own use cases prior to deployment.

For future work, there are interesting directions including, but not limited to:

- Greater breadth of binary classification metrics (especially rank-based) and proving guarantees for these.
- Extension to multi-class classification. We might consider extending from Y_i being Bernoulli distributed to multinoulli (categorical) distributed. For the robustness results, we used the conjugate prior to the Bernoulli distribution, the Beta distribution. In a similar way we might consider the conjugate to the multinoulli, that is, the Dirichlet distribution. On the evaluation metrics themselves, their multi-class variants are typically treated the same as their binary counterparts, using a "one vs rest" logic.
- Expanding PEMI to regression models and relevant metrics. Here, the continuous, potentially unbounded nature of the label space $\mathcal{Y}$ may present some difficulties.

References

P.D. Allison. 2001. *Missing Data*. SAGE Publications.

Salim I. Amoukou, Tom Bewley, Saumitra Mishra, Freddy Lecue, Daniele Magazzeni, and Manuela

Veloso. 2024. Sequential Harmful Shift Detection Without Labels. In *Advances in Neural Information Processing Systems*.

Olawale F Ayilara, Lisa Zhang, Tolulope T Sajobi, Richard Sawatzky, Eric Bohm, and Lisa M Lix. 2019. Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and quality of life outcomes* 17 (2019), 1–9.

Alexis Ayme, Claire Boyer, Aymeric Dieuleveut, and Erwan Scornet. 2024. Random features models: a way to study the success of naive imputation. arXiv:2402.03839 [math.ST]

Barry Becker and Ronny Kohavi. 1996. Adult. UCI Machine Learning Repository. `https://doi.org/10.24432/C5XW20`

Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. 2018. From Predictive Methods to Missing Data Imputation: An Optimization Approach. *Journal of Machine Learning Research* 18, 196 (2018), 1–39.

Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition* 30, 7 (1997), 1145–1159.

R.C. Bradley. 2007. *Introduction to Strong Mixing Conditions*. Number v. 1 in Introduction to Strong Mixing Conditions. Kendrick Press.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. 2006. *Semi-Supervised Learning*. MIT Press.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*.

J. Dedecker, P. Doukhan, G. Lang, J.R. Leon, S. Louhichi, and C. Prieur. 2007. *Weak Dependence: With Examples and Applications*. Springer New York.

Morris H. DeGroot and Stephen E. Fienberg. 1983. The Comparison and Evaluation of Forecasters. *Journal of the Royal Statistical Society. Series D (The Statistician)* 32, 1/2 (1983), 12–22.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)* 39, 1 (1977), 1–22.

Francis X. Diebold, Todd A. Gunther, and Anthony S. Tay. 1998. Evaluating Density Forecasts with Applications to Financial Risk Management. *International Economic Review* 39, 4 (1998), 863–883.

A Rogier T Donders, Geert JMG Van Der Heijden, Theo Stijnen, and Karel GM Moons. 2006. A gentle introduction to imputation of missing values. *Journal of clinical epidemiology* 59, 10 (2006), 1087–1091.

P. Doukhan. 2012. *Mixing: Properties and Examples*. Springer New York.

Xavier d'Haultfoeuille. 2010. A new instrumental method for dealing with endogenous selection. *Journal of Econometrics* 154, 1 (2010), 1–15.

C. G. Esseen. 1956. A moment inequality with an application to the central limit theorem. *Scandinavian Actuarial Journal* 1956, 2 (1956), 160–170.

Matthias Feurer, Jan N. van Rijn, Arlind Kadra, Pieter Gijsbers, Neeratyoy Mallik, Sahithya Ravi, Andreas Müller, Joaquin Vanschoren, and Frank Hutter. 2021. OpenML-Python: an extensible Python API for OpenML. *Journal of Machine Learning Research* 22, 100 (2021), 1–5.

Shuxia Gong and Chen Ma. 2025. Gradient-Based Multiple Robust Learning Calibration on Data Missing-Not-at-Random via Bi-Level Optimization. *Entropy* 27, 2 (2025).

Licong Guan and Xue Yuan. 2024. Label-free model evaluation and weighted uncertainty sample selection for domain adaptive instance segmentation. *Engineering Applications of Artificial Intelligence* 127 (2024), 107204.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML'17)*. 1321–1330.

J A Hanley and B J McNeil. 1983. A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* 148, 3 (1983), 839–843.

Jack C. Hayya, Don Armstrong, and Nicolas Gressis. 1975. A Note on the Ratio of Two Normally Distributed Variables. *Management Science* 21 (1975), 1338–1341.

Tim C. Hesterberg. 2015. What Teachers Should Know About the Bootstrap: Resampling in the Undergraduate Statistics Curriculum. *The American Statistician* 69, 4 (2015), 371–386.

Hideitsu Hino. 2020. Active Learning: Problem Settings and Recent Developments. arXiv:2012.04225 [cs.LG]

Hans Hofmann. 1994. Statlog (German Credit Data). UCI Machine Learning Repository. `https://doi.org/10.24432/C5NC77`

Xinting Hu, Yulei Niu, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. 2022. On Non-Random Missing Labels in Semi-Supervised Learning. In *The Tenth International Conference on Learning Representations, ICLR 2022*.

Joseph G. Ibrahim, Ming-Hui Chen, and Stuart R. Lipsitz. 2001. Missing Responses in Generalised Linear Mixed Models When the Missing Data Mechanism is Nonignorable. *Biometrika* 88, 2 (2001), 551–564.

Joseph G. Ibrahim and Stuart R. Lipsitz. 1996. Parameter Estimation from Incomplete Data in Binomial Regression When the Missing Data Mechanism is Nonignorable. *Biometrics* 52, 3 (1996), 1071–1078.

IMDB. 2024. Information courtesy of IMDb (`https://www.imdb.com`). Used with permission.

E. T. Jaynes. 1957. Information Theory and Statistical Mechanics. *Phys. Rev.* 106 (May 1957), 620–630. Issue 4.

Julie Josse, Jacob M. Chen, Nicolas Prost, Erwan Scornet, and Gaël Varoquaux. 2024. On the consistency of supervised learning with missing values. arXiv:1902.06931 [stat.ML]

Julie Josse and Jerome P. Reiter. 2018. Introduction to the Special Section on Missing Data. *Statist. Sci.* 33, 2 (2018), 139–141.

Michael Kahn. 1994. Diabetes. UCI Machine Learning Repository. `https://doi.org/10.24432/C5T59G`

Krishnaram Kenthapadi, Himabindu Lakkaraju, Pradeep Natarajan, and Mehrnoosh Sameki. 2022. Model Monitoring in Practice: Lessons Learned and Open Challenges. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. 4800–4801.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. 2018. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*. PMLR, 2796–2804.

Ananya Kumar, Percy Liang, and Tengyu Ma. 2019. Verified Uncertainty Calibration. In *Advances in Neural Information Processing Systems (NeurIPS '19)*.

Wonbin Kweon and Hwanjo Yu. 2024. Doubly Calibrated Estimator for Recommendation on Data Missing Not at Random. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*. 3810–3820.

R. Little, D. Rubin, and an O'Reilly Media Company Safari. 2019. *Statistical Analysis with Missing Data., 3rd Edition*. Wiley.

Angel D Reyero Lobo, Alexis Ayme, Claire Boyer, and Erwan Scornet. 2024. Harnessing pattern-by-pattern linear classifiers for prediction with missing data. arXiv:2405.09196 [math.ST]

George Marsaglia. 1965. Ratios of Normal Variables and Ratios of Sums of Uniform Variables. *J. Amer. Statist. Assoc.* 60, 309 (1965), 193–204.

George Marsaglia. 2006. Ratios of Normal Variables. *Journal of Statistical Software* 16, 4 (2006), 1–10.

Alan Mishler, Mohsen Ghassemi, Alec Koppel, and Sumitra Ganesh. 2023. Active Learning with Missing Not At Random Outcomes. In *NeurIPS 2023 Workshop on Adaptive Experimental Design and Active Learning in the Real World*.

Karthika Mohan, Judea Pearl, and Jin Tian. 2013. Graphical Models for Inference with Missing Data. In *Advances in Neural Information Processing Systems (NeurIPS '13)*.

Rita P. Moro, S. and P. Cortez. 2014. Bank Marketing. UCI Machine Learning Repository. `https://doi.org/10.24432/C5K306`

Marine Le Morvan, Julie Josse, Erwan Scornet, and Gaël Varoquaux. 2024. What's a good imputation to predict with missing values?. In *Proceedings of the 35th International Conference on Neural Information Processing Systems (NeurIPS '21)*.

Jesse Read, Albert Bifet, Bernhard Pfahringer, and Geoff Holmes. 2012. Batch-Incremental versus Instance-Incremental Learning in Dynamic and Evolving Data. In *Advances in Intelligent Data Analysis XI*. Springer Berlin Heidelberg, Berlin, Heidelberg, 313–323.

Donald B Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.

H. Scudder. 1965. Probability of error of some adaptive pattern-recognition machines. *IEEE Transactions on Information Theory* 11, 3 (1965), 363–371.

Burr Settles. 2009. *Active Learning Literature Survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

Irina G Shevtsova. 2010. An improvement of convergence rate estimates in the Lyapunov theorem.. In *Doklady Mathematics*, Vol. 82.

Qinbao Song and Martin Shepperd. 2007. Missing data imputation techniques. *International journal of business intelligence and data mining* 2, 3 (2007), 261–291.

Aude Sportisse, Hugo Schmutz, Olivier Humbert, Charles Bouveyron, and Pierre-Alexandre Mattei. 2023. Are labels informative in semi-supervised learning? estimating and leveraging the missing-data mechanism. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*.

Xiaoxiao Sun, Yunzhong Hou, Weijian Deng, Hongdong Li, and Liang Zheng. 2021. Ranking Models in Unlabeled New Environments. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV '21)*.

A. N. Tikhomirov. 1981. On the Convergence Rate in the Central Limit Theorem for Weakly Dependent Random Variables. *Theory of Probability & Its Applications* 25, 4 (1981), 790–809.

Stephen Tridgell. 2016. Dota2 Games Results. UCI Machine Learning Repository. `https://doi.org/10.24432/C5W593`

Richard Williams. 2015. Missing data part 1: Overview, traditional methods. *Notre Dame: University of Notre Dame* (2015).

Bianca Zadrozny and Charles Elkan. 2001. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML '01)*. 609–616.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [**Yes**]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [**Yes**] Time complexity not provided as is trivial.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [**No**] Source code was made available to reviewers.
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [**Yes**]
   (b) Complete proofs of all theoretical results. [**Yes**], see Section B in Appendix.
   (c) Clear explanations of any assumptions. [**Yes**]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [**No**] Source code was made available to reviewers.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [**Yes**]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [**Yes**]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [**Yes**] See beginning of Section C in Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [**Yes**]
   (b) The license information of the assets, if applicable. [**Yes**] All datasets used have permissive licences.
   (c) New assets either in the supplemental material or as a URL, if applicable. [**Not Applicable**]
   (d) Information about consent from data providers/curators. [**No**] They are all commonly used public datasets with many citations.
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [**Not Applicable**]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [**Not Applicable**]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [**Not Applicable**]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [**Not Applicable**]

A Probabilistic Queries

Within this work we invoke the concept of a probabilistic query, which we use in the same manner as Mohan et al. (2013). Since this is not well-known, we describe it briefly here.

Suppose we have random variables $X_1, X_2, \ldots$. A *primitive probabilistic query* $P : X_1, X_2, \ldots \to [0, 1]$ describes a new random variable corresponding to a question we wish to ask about the random variables, that has a probability as an answer. For example, $P(X_1 = X_2)$ is the query corresponding to the question *"what is the probability that r.v.* X_1 *equals r.v.* X_2*?"*. We can combine primitive probabilistic queries using elementary algebra to obtain a general probabilistic query, e.g.,

$$\frac{P(X_1 = X_2)}{P(X_2 < X_3)}.$$

Primitive probabilistic queries relate to *probabilities* in that under a distribution $\mathcal{D}$, the expectation of the primitive query P equals the probability $\mathbb{P}$, that is,

$$\mathbb{E}_{\mathcal{D}}[P(E)] = \mathbb{P}_{\mathcal{D}}[E] \text{ for all events } E.$$

This is a notational convenience to distinguish the mathematical form of a probabilistic question $P(E)$ from the actual realised numerical probability $\mathbb{P}[E]$, allowing the same query to be evaluated under different distributions. For example, consider distributions $\mathcal{D}_1$ and $\mathcal{D}_2$ such that $\mathcal{D}_1 \neq \mathcal{D}_2$. The probabilistic query $P(X_1 = X_2)$ can be discussed as an object separate to the probabilities $\mathbb{P}_{\mathcal{D}_1}[X_1 = X_2] \neq \mathbb{P}_{\mathcal{D}_2}[X_1 = X_2]$.

B Proofs and Ancillary Results

For a positive integer $n \in \mathbb{N}$, let $[n] = \{1, 2, \ldots, n\}$. A function $f(x) = O(g(x))$ if there exist $B > 0$, $x_0 \in \mathbb{R}$ such that $|f(x)| \leq Bg(x)$ for all $x \geq x_0$.

B.1 Vanilla Accuracy Estimator

Proposition 5. *The vanilla accuracy estimator,*

$$\widehat{Q}_{acc,n} = \frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{y_i = \psi(x_i)\},$$

is both an unbiased and consistent estimator of the performance metric $\mathcal{Q}_{acc} = \mathbb{E}_{\mathcal{D}}[P(\psi(X) = Y)]$.

Proof. Unbiasedness.

$$\begin{aligned}
\mathbb{E}\Big[\widehat{Q}_{\text{acc},n}\Big] &= \mathbb{E}\left[\frac{1}{n} \sum_{i \in [n]} \mathbb{1}\{y_i = \psi(x_i)\}\right] \\
&= \frac{1}{n} \sum_{i \in [n]} \mathbb{E}[\mathbb{1}\{y_i = \psi(x_i)\}] && \text{(linearity of expectation)} \\
&= \frac{1}{n} \sum_{i \in [n]} \mathbb{P}[Y = \psi(X)] \\
&= \frac{1}{n} \sum_{i \in [n]} \mathbb{E}[P(Y = \psi(X))] && \text{(Definition of probabilistic query)} \\
&= \mathbb{E}[P(Y = \psi(X))] = \mathcal{Q}_{\text{acc}}.
\end{aligned}$$

Consistency.

$$\begin{aligned}
\mathbb{V}\left[\widehat{Q}_{\text{acc},n}\right] &= \mathbb{V}\left[\frac{1}{n}\sum_{i\in[n]} \mathbb{1}\{y_i = \psi(x_i)\}\right] \\
&= \frac{1}{n^2}\sum_{i\in[n]} \mathbb{V}[\mathbb{1}\{y_i = \psi(x_i)\}] && (Y_i \text{ are mutually independent}) \\
&= \frac{1}{n^2}\sum_{i\in[n]} \mathbb{P}[Y = \psi(X)](1 - \mathbb{P}[Y = \psi(X)]) \\
&= \frac{1}{n}\mathbb{P}[Y = \psi(X)](1 - \mathbb{P}[Y = \psi(X)]).
\end{aligned}$$

As $\mathbb{V}\left[\widehat{Q}_{\text{acc},n}\right] = O(1/n)$, we have that $\widehat{Q}_{\text{acc},n}$ is a consistent estimator of $\mathcal{Q}_{\text{acc}}$ from Chebyshev's inequality. □

B.2 Confusion Matrix Estimators

Lemma 6. *Suppose we have missing evaluation data ($\overline{\mathcal{K}} \neq \emptyset$). The mean and covariance of the confusion matrix element estimators ($\widehat{TP}, \widehat{FN}, \widehat{FP}, \widehat{TN}$) are*

$$\mu_{CM} = \begin{matrix} TP \\ FN \\ FP \\ TN \end{matrix}\begin{pmatrix} \sum_{i\in\mathcal{K}} y_i\psi_i + \sum_{i\in\overline{\mathcal{K}}} p_i\psi_i \\ \sum_{i\in\mathcal{K}} y_i\overline{\psi_i} + \sum_{i\in\overline{\mathcal{K}}} p_i\overline{\psi_i} \\ \sum_{i\in\mathcal{K}} \overline{y_i}\psi_i + \sum_{i\in\overline{\mathcal{K}}} (1-p_i)\psi_i \\ \sum_{i\in\mathcal{K}} \overline{y_i}\overline{\psi_i} + \sum_{i\in\overline{\mathcal{K}}} (1-p_i)\overline{\psi_i} \end{pmatrix},$$

$$\Sigma_{CM} = \begin{matrix} \\ TP \\ FN \\ FP \\ TN \end{matrix}\begin{matrix} \begin{matrix} TP \end{matrix} & \begin{matrix} FN \end{matrix} & \begin{matrix} FP \end{matrix} & \begin{matrix} TN \end{matrix} \end{matrix}\begin{pmatrix} \sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\psi_i & 0 & -\sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\psi_i & 0 \\ 0 & \sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\overline{\psi_i} & 0 & -\sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\overline{\psi_i} \\ -\sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\psi_i & 0 & \sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\psi_i & 0 \\ 0 & -\sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\overline{\psi_i} & 0 & \sum_{i\in\overline{\mathcal{K}}} p_i(1-p_i)\overline{\psi_i} \end{pmatrix},$$

where for a variable $z \in \{0,1\}$, $\overline{z} = 1 - z$.

Proof. Each element is a straightforward consequence of the following: linearity of expectation; covariance of linear combinations; the fact that $\mathbb{E}[Y] = p$, $\mathbb{V}[Y] = p(1-p)$ for $Y \sim \mathcal{B}(p)$; $\text{Cov}[U, V] = 0$ for independent r.v.s U, V. □

B.3 Sums of Bernoulli Variables

We shall require the following classic result.

Theorem 7 (Berry-Esseen, (Esseen, 1956; Shevtsova, 2010))**.** *Let $T_1, T_2, \ldots$ be independent random variables with $\mathbb{E}[T_i] = 0$, $\mathbb{E}[T_i^2] = \sigma_i^2 > 0$ and $\mathbb{E}[|T_i|^3] = \rho_i < \infty$. Also, let*

$$S_n = \frac{T_1 + T_2 + \cdots + T_n}{\sqrt{\sigma_1^2 + \sigma_1^2 + \cdots + \sigma_n^2}}$$

and F_{S_n} be the cdf of S_n. Then, the Berry-Esseen theorem states there exists a universal constant $C_0 \in [0.4097, 0.5600]$ such that $\mathrm{d}_{\mathsf{KS}}(F_{S_n}, \mathcal{N}(0,1)) \leq C_0\psi_0$, where

$$\psi_0 = \left(\sum_{i=1}^n \sigma_i^2\right)^{-3/2} \cdot \sum_{i=1}^n \rho_i.$$

B.3.1 Proof of Lemma 1

Lemma 1. *Let $Z = \sum_{i=1}^n a_i Y_i$, where $Y_i \sim \mathcal{B}(p_i)$ and mutually independent, $p_i \in (0,1)$ and $a_i \in \mathbb{R} \setminus \{0\}$. Then, we have that*

$$\mathrm{d}_{\mathsf{KS}}(F_Z, \mathcal{N}(\mu_z, \sigma_z^2)) \leq \frac{C_0}{\sqrt{nv_*}} \frac{1 + a^*}{2a_*}, \quad \text{where}$$

$$v_* = \min_i\{p_i(1-p_i)\}, \quad a_* = \min_i\{|a_i|\},\ a^* = \max_i\{|a_i|\},$$
$$\mu_z = \textstyle\sum_{i=1}^n a_i p_i,\ \sigma_z^2 = \sum_{i=1}^n a_i p_i(1-p_i)$$

and $C_0 = 0.5600$ *is a universal constant.*

Proof. We proceed with a direct application of the Berry-Esseen theorem. First define the random variables $T_i = \frac{a_i}{\sigma_z}(Y_i - p_i)$ and $S_n = \sum_{i=1}^n T_i$. We then have

$$\mathbb{E}[T_i] = 0,\ \mathbb{E}[T_i^2] = \frac{a_i^2}{\sigma_z^2}p_i(1-p_i),\ \mathbb{E}[|T_i|^3] = \frac{a_i^3}{\sigma_z^3}p_i(1-p_i)(1-2p_i+2p_i^2). \tag{4}$$

Towards applying Berry-Esseen, we then evaluate

$$\begin{aligned}
\psi_0 &= \left(\sum_{i=1}^n \frac{a_i^2}{\sigma_z^2}p_i(1-p_i)\right)^{-3/2} \cdot \sum_{i=1}^n \frac{a_i^3}{\sigma_z^3}p_i(1-p_i)(1-2p_i+2p_i^2) && (5)\\
&= \left(\frac{1}{\sigma_z^2}\right)^{-3/2}\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{-3/2} \cdot \left(\frac{1}{\sigma_z^3}\right)\sum_{i=1}^n a_i^3 p_i(1-p_i)\left(\frac{1}{2} + 2(p_i - \frac{1}{2})^2\right) && \text{(Factorising)}\\
&= \frac{1}{2}\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{-1/2} + 2\cdot\frac{\sum_{i=1}^n a_i^3 p_i(1-p_i)(p_i-\frac{1}{2})^2}{\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{3/2}} && \text{(Splitting sum and } \sigma_z \text{ cancellation)}\\
&\leq \frac{1}{2}\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{-1/2} + \frac{1}{2}\cdot\frac{\sum_{i=1}^n a_i^3 p_i(1-p_i)}{\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{3/2}} && \text{(Since } (p_i-\tfrac{1}{2})^2 \leq \tfrac{1}{4})\\
&\leq \frac{1}{2}\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{-1/2} + \frac{a^*}{2}\cdot\frac{\sum_{i=1}^n a_i^2 p_i(1-p_i)}{\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{3/2}} && \text{(Since } a_i \leq a^* \text{ for all } i \in [n])\\
&= \frac{1+a^*}{2}\left(\sum_{i=1}^n a_i^2 p_i(1-p_i)\right)^{-1/2} && (6)\\
&\leq \frac{1+a^*}{2}\left(\sum_{i=1}^n a_*^2 p_i(1-p_i)\right)^{-1/2} && \text{(Since } a_i^2 \geq a_*^2 \text{ for all } i \in [n])\\
&= \frac{1+a^*}{2a_*}\left(\sum_{i=1}^n p_i(1-p_i)\right)^{-1/2}. && (7)
\end{aligned}$$

Consider that

$$\sum_{i=1}^n p_i(1-p_i) \geq \sum_{i=1}^n v_* = nv_* \tag{8}$$
$$\iff \left(\sum_{i=1}^n p_i(1-p_i)\right)^{-1/2} \leq (nv_*)^{-1/2} \tag{9}$$

Substituting (7) and (9) into the Berry-Esseen theorem yields $d_{\mathsf{KS}}(F_{S_n}, \mathcal{N}(0,1)) \leq \frac{C_0}{\sqrt{nv_*}}\frac{1+a^*}{2a_*}$. This implies $|\mathbb{P}(S_n \leq t) - \Phi(t)| \leq \frac{C_0}{\sqrt{nv_*}}\frac{1+a^*}{2a_*}$ for all $t \in \mathbb{R}$. Choosing $t = \frac{z-\mu_z}{\sigma_z}$ for any $z \in \mathbb{R}$ we have that

$$\left|\mathbb{P}(S_n \leq \frac{z-\mu_z}{\sigma_z}) - \Phi(\frac{z-\mu_z}{\sigma_z})\right| = \left|\mathbb{P}(Z \leq z) - \Phi(\frac{z-\mu_z}{\sigma_z})\right|$$

since

$$S_n \leq \frac{z-\mu_z}{\sigma_z} \iff \sum_{i=1}^{n} \frac{a_i}{\sigma_z}(X_i - p_i) \leq \frac{z-\mu_z}{\sigma_z} \tag{10}$$

$$\iff \sum_{i=1}^{n} \frac{a_i}{\sigma_z} X_i - \sum_{i=1}^{n} \frac{a_i}{\sigma_z} p_i \leq \frac{z-\mu_z}{\sigma_z} \tag{11}$$

$$\iff \frac{Z-\mu_z}{\sigma_z} \leq \frac{z-\mu_z}{\sigma_z} \iff Z \leq z \tag{12}$$

and the result follows. □

B.4 Ratio of Correlated Gaussians

B.4.1 Proof of Theorem 2

We require the following ancillary result for proving the convergence bound of a ratio of correlated normal random variables.

Lemma 8. *Consider a real number $a \in \mathbb{R}$ and function $s : \mathbb{R} \to \mathbb{R}$ with Taylor series $s(t) = s_0 + s_1(t-t_0) + s_2(t-t_0)^2 + O((t-t_0)^3)$ such that $s_0 > 0$. We can expand*

$$\frac{a}{s(t)} = \frac{a}{s_0} - \frac{a s_1}{s_0^2}(t-t_0) + a\left(\frac{s_1^2}{s_0^3} - \frac{s_2}{s_0^2}\right)(t-t_0)^2 + O((t-t_0)^3).$$

Proof. Given the Taylor series expansion of $s(t)$:

$$s(t) = s_0 + s_1(t-t_0) + s_2(t-t_0)^2 + O((t-t_0)^3),$$

we want to find the Taylor series expansion of $\frac{a}{s(t)}$.

First, we rewrite $s(t)$ in a more convenient form:

$$s(t) = s_0\left(1 + \frac{s_1}{s_0}(t-t_0) + \frac{s_2}{s_0}(t-t_0)^2 + O((t-t_0)^3)\right).$$

Let $\epsilon = \frac{s_1}{s_0}(t-t_0) + \frac{s_2}{s_0}(t-t_0)^2 + O((t-t_0)^3)$. Then:

$$s(t) = s_0(1+\epsilon).$$

The reciprocal of $s(t)$ is:

$$\frac{1}{s(t)} = \frac{1}{s_0(1+\epsilon)} = \frac{1}{s_0} \cdot \frac{1}{1+\epsilon}.$$

Using the Taylor series expansion for $\frac{1}{1+\epsilon}$ around $\epsilon = 0$:

$$\frac{1}{1+\epsilon} = 1 - \epsilon + \epsilon^2 - \epsilon^3 + O(\epsilon^4).$$

Substituting back $\epsilon = \frac{s_1}{s_0}(t-t_0) + \frac{s_2}{s_0}(t-t_0)^2 + O((t-t_0)^3)$, we get:

$$\frac{1}{s(t)} = \frac{1}{s_0}\Bigg(1 - \left(\frac{s_1}{s_0}(t-t_0) + \frac{s_2}{s_0}(t-t_0)^2 + O((t-t_0)^3)\right) + \left(\frac{s_1}{s_0}(t-t_0) + \frac{s_2}{s_0}(t-t_0)^2 + O((t-t_0)^3)\right)^2 + O((t-t_0)^3)\Bigg).$$

Expanding the square term and collecting like powers of $(t - t_0)$, we get:

$$\frac{1}{s(t)} = \frac{1}{s_0}\left(1 - \frac{s_1}{s_0}(t-t_0) - \frac{s_2}{s_0}(t-t_0)^2 + \left(\frac{s_1}{s_0}(t-t_0)\right)^2 + O((t-t_0)^3)\right).$$

Simplifying further:

$$\frac{1}{s(t)} = \frac{1}{s_0} - \frac{s_1}{s_0^2}(t-t_0) + \left(\frac{s_1^2}{s_0^3} - \frac{s_2}{s_0^2}\right)(t-t_0)^2 + O((t-t_0)^3).$$

Finally, multiplying by a, we get the Taylor series expansion for $\frac{a}{s(t)}$:

$$\frac{a}{s(t)} = \frac{a}{s_0} - \frac{as_1}{s_0^2}(t-t_0) + a\left(\frac{s_1^2}{s_0^3} - \frac{s_2}{s_0^2}\right)(t-t_0)^2 + O((t-t_0)^3).$$

So, the Taylor series expansion of $\frac{a}{s(t)}$ around $t = t_0$ is:

$$\frac{a}{s(t)} = \frac{a}{s_0} - \frac{as_1}{s_0^2}(t-t_0) + a\left(\frac{s_1^2}{s_0^3} - \frac{s_2}{s_0^2}\right)(t-t_0)^2 + O((t-t_0)^3).$$

□

Theorem 2 (Ratio of Correlated Gaussians)**.** *Let Z and W be jointly Gaussian random variables, i.e. $(Z, W) \sim \mathcal{N}(m, \Sigma)$, where $m = [\mu_z, \mu_w]^\mathsf{T}$ and $\Sigma = \begin{bmatrix} \sigma_z^2 & \rho\sigma_z\sigma_w \\ \rho\sigma_w\sigma_z & \sigma_w^2 \end{bmatrix}$. Moreover, let the the distribution of $T := Z/W$ be described by $G(t) = \mathbb{P}[T \leq t]$. Then, under conditions*

1. *$\mathbb{P}(W > 0) \to 1$,. i.e. $\mu_w/\sigma_w \gg 1$;*
2. *$2|\mu_z|\left|\frac{\sigma_w}{\mu_w} - \frac{\sigma_z}{\mu_z}\right| \gg 1$;*

the distribution of Z/W satisfies

$$\mathrm{d}_{\mathsf{KS}}(G, \mathcal{N}(\mu, \sigma^2)) \leq \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_w^2(|\mu_z| + \sigma_z^2) + \mu_w^2}{\sigma\mu_w^3},$$

$$\textit{where} \quad \mu = \tfrac{\mu_z}{\mu_w}, \quad \sigma^2 = \tfrac{{\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\,\rho\sigma_z\sigma_w\mu_z\mu_w}{{\mu_w}^4}.$$

Proof. We proceed via the standard Geary-Hinkley transformation (Hayya et al., 1975), then bound the Kolmogorov-Smirnov distance from a Normal distribution using an upper bound on the standard Normal CDF and Local-Lipschitz type argument.

Using the axioms of probability we have

$$G(t) := \mathbb{P}(Z/W \leq t) = \mathbb{P}(Z - tW < 0, W > 0) + \mathbb{P}(Z - tW \geq 0, W \leq 0)$$

$$= \mathbb{P}(Z - tW < 0 | W > 0)\mathbb{P}(W > 0) + \mathbb{P}(Z - tW \geq 0 | W \leq 0)\mathbb{P}(W \leq 0)$$

$$\to \mathbb{P}(Z - tW < 0 | W > 0)\mathbb{P}(W > 0) \to \mathbb{P}(Z - tW < 0)$$

where we invoke the condition $\mathbb{P}(W > 0) \to 1$.

The random variable $D := Z - tW$ is itself a normal distribution with parameters $\mu_d = \mu_z - t\mu_w$, $\sigma_d^2 = \sigma_z^2 - 2tc + t^2\sigma_w^2$, where $c := \rho\sigma_z\sigma_w$. Thus we have

$$\mathbb{P}(Z - tW < 0) = \Phi(\frac{-\mu_d}{\sigma_d}) = \Phi\left(\frac{t - \frac{\mu_z}{\mu_w}}{\mu_w^{-1}\sqrt{\sigma_z^2 - 2tc + t^2\sigma_w^2}}\right) = \Phi\left(\frac{t-\mu}{s(t)}\right),$$

where we have defined $s(t) := \mu_w^{-1}\sqrt{\sigma_z^2 - 2tc + t^2\sigma_w^2}$.

We evaluate

$$\left|\mathbb{P}(Z/W \le t) - \Phi\left(\frac{t-\mu}{\sigma}\right)\right| = \left|\Phi\left(\frac{t-\mu}{s(t)}\right) - \Phi\left(\frac{t-\mu}{\sigma}\right)\right|$$

$$= \left|\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(t-\mu)/\sigma} \exp\left(-\frac{z^2}{2}\right)\mathrm{d}z - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{(t-\mu)/s(t)} \exp\left(-\frac{z^2}{2}\right)\mathrm{d}z\right|$$

$$= \left|\frac{1}{\sqrt{2\pi}}\int_{(t-\mu)/s(t)}^{(t-\mu)/\sigma} \exp\left(-\frac{z^2}{2}\right)\mathrm{d}z\right| = \left|\frac{1}{\sqrt{2\pi}}\int_{(t-\mu)/s(t)}^{(t-\mu)/\sigma} \frac{1}{z}\cdot z\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z\right|$$

$$= \left|\frac{1}{\sqrt{2\pi}}\left(\frac{-1}{z}\cdot\exp\left(-\frac{z^2}{2}\right)\right)\Bigg|_{(t-\mu)/s(t)}^{(t-\mu)/\sigma} + \frac{1}{\sqrt{2\pi}}\int_{(t-\mu)/s(t)}^{(t-\mu)/\sigma} \frac{1}{z}\exp\left(-\frac{z^2}{2}\right)\mathrm{d}z\right| \le \left|\frac{1}{\sqrt{2\pi}}\left(\frac{-1}{z}\cdot\exp\left(-\frac{z^2}{2}\right)\right)\Bigg|_{(t-\mu)/s(t)}^{(t-\mu)/\sigma}\right|$$

where we use integration-by-parts. Evaluating at the limits gives

$$\begin{aligned}\left|\mathbb{P}(Z/W \le t) - \Phi\left(\frac{t-\mu}{\sigma}\right)\right| &= \frac{1}{\sqrt{2\pi}}\left|\frac{s(t)}{t-\mu}\exp\left(-\frac{(t-\mu)^2}{2s(t)^2}\right) - \frac{\sigma}{t-\mu}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\right| \\ &\le \frac{1}{\sqrt{2\pi}}\frac{1}{|t-\mu|}\left|s(t)\exp\left(-\frac{(t-\mu)^2}{2s(t)^2}\right) - \sigma\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)\right| \\ &= \frac{1}{\sqrt{2\pi}}\frac{1}{|t-\mu|}|f(t)| \end{aligned} \tag{13}$$

where we have defined $f(t) := s(t)\exp\big(-(t-\mu)^2/2s(t)^2\big) - \sigma\exp\big(-(t-\mu)^2/2\sigma^2\big)$. We will consider the Taylor expansion of $s(t)$ and substitute into the definition of $f(t)$ to establish our bounds. Consider now the Taylor expansion of $s(t)$ around $t=\mu$, whence

$$s(t) = \sigma + s_1(t-\mu) + s_2(t-\mu)^2 + \cdots$$

Using Lemma 8, we expand $(t-\mu)/s(t)$ as

$$\begin{aligned}\frac{t-\mu}{s(t)} &= \frac{t-\mu}{\sigma} - \frac{(t-\mu)s_1}{\sigma^2}(t-\mu) + (t-\mu)\left(\frac{s_1^2}{\sigma^3} - \frac{s_2}{\sigma^2}\right)(t-\mu)^2 + O((t-\mu)^3) \\ &= \frac{t-\mu}{\sigma} - \frac{s_1}{\sigma^2}(t-\mu)^2 + O((t-\mu)^3). \end{aligned} \tag{14}$$

Recall the power series

$$\exp\left(-\frac{z^2}{2}\right) = \sum_{k=0}^{\infty}\frac{(-1)^k z^{2k}}{2^k k!} = 1 - \frac{z^2}{2} + \frac{z^4}{8} - \cdots \tag{15}$$

which converges for all $z \in (-\infty, \infty)$. We substitute (14) into (15) yielding

$$\exp\left(-\frac{(t-\mu)^2}{2s(t)^2}\right) = 1 - \frac{1}{2}\left[\frac{t-\mu}{\sigma} - \frac{s_1}{\sigma^2}(t-\mu)^2 + O((t-\mu)^3)\right] + O((t-\mu)^4). \tag{16}$$

Multiplying by the Taylor expansion for $s(t)$ yields

$$s(t)\exp\left(-\frac{(t-\mu)^2}{2s(t)^2}\right) = \big(\sigma + s_1(t-\mu) + s_2(t-\mu)^2 + O((t-\mu)^3)\big) \tag{17}$$

$$\cdot\left(1 - \frac{1}{2}\left[\frac{t-\mu}{\sigma} - \frac{s_1}{\sigma^2}(t-\mu)^2 + O((t-\mu)^3)\right] + O((t-\mu)^4)\right) \tag{18}$$

$$= \sigma\left[1 - \frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] + s_1(t-\mu) + s_2(t-\mu)^2 + O((t-\mu)^3) \tag{19}$$

Similarly, we have using (15) that

$$\sigma \exp\left(\frac{-(t-\mu)^2}{2\sigma^2}\right) = \sigma\left[1 - \frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right] + O((t-\mu)^4). \tag{20}$$

Subtracting (20) from (17) yields

$$f(t) = s_1(t-\mu) + s_2(t-\mu)^2 + O((t-\mu)^3). \tag{21}$$

The function $s(t)$ is the square root of a quadratic function, and therefore is very close to a linear function. As such we expect the quadratic terms of $s(t)$ to dominate, that is, we expect in the neighbourhood of $t = \mu$

$$\left|s_0 + s_1(t-t_0) + s_2(t-t_0)^2\right| \geq \left|\sum_{k=3}^{\infty} s_k (t-t_0)^k\right|.$$

Moreover, as $f(t)$ is a smooth function of $s(t)$ and t we expect this property to hold for the expansion of $f(t)$ in (21), that is, in the neighbourhood of $t = \mu$

$$\left|f_0 + f_1(t-\mu) + f_2(t-\mu)^2\right| \geq \left|\sum_{k=3}^{\infty} f_k (t-\mu)^k\right|, \tag{22}$$

where $f(t) = \sum_{k=0}^{\infty} f_k (t-\mu)^k$ is the Taylor expansion of $f(t)$ around $t = \mu$. Indeed, one can verify that $f_0 = f(\mu) = 0$ and $f_1 = s_1$, $f_2 = s_2$.

Thus we have

$$\begin{aligned}
|f(t)| &= \left|s_1(t-\mu) + s_2(t-\mu)^2 + \sum_{k=3}^{\infty} f_k (t-\mu)^k\right| && (23)\\
&\leq \left|s_1(t-\mu) + s_2(t-\mu)^2\right| + \left|\sum_{k=3}^{\infty} f_k (t-\mu)^k\right| && \text{(triangle inequality)}\\
&\leq 2\left|s_1(t-\mu) + s_2(t-\mu)^2\right| && \text{(Using (22))}
\end{aligned}$$

It can be shown by direct differentiation that

$$s_1 = \frac{\sqrt{{\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\,c\mu_w\mu_z}\left(\mu_z{\sigma_w}^2 - c\mu_w\right)}{\mu_w{\mu_z}^2{\sigma_w}^2 + {\mu_w}^3{\sigma_z}^2 - 2\,c{\mu_w}^2\mu_z} = \frac{\mu_z{\sigma_w}^2 - c\mu_w}{\sqrt{{\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\,c\mu_w\mu_z}\mu_w} = \frac{\mu_z\sigma_w^2 - c\mu_w}{\mu_w^3\sigma}$$

$$\begin{aligned}
s_2 &= \frac{\left({\mu_w}^2{\sigma_w}^2{\sigma_z}^2 - c^2{\mu_w}^2\right)\sqrt{{\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\,c\mu_w\mu_z}}{2\left({\mu_z}^4{\sigma_w}^4 + {\mu_w}^4{\sigma_z}^4 - 4\,c\mu_w{\mu_z}^3{\sigma_w}^2 + 4\,c^2{\mu_w}^2{\mu_z}^2 + 2\left({\mu_w}^2{\mu_z}^2{\sigma_w}^2 - 2\,c{\mu_w}^3\mu_z\right){\sigma_z}^2\right)}\\
&= \frac{(\sigma_w\sigma_z + c)(\sigma_w\sigma_z - c){\mu_w}^2}{2\left({\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\,c\mu_w\mu_z\right)^{\frac{3}{2}}} = \frac{\sigma_w^2\sigma_z^2(1-\rho^2)}{2\mu_w^4\sigma^3}
\end{aligned}$$

Thus we have

$$\begin{aligned}
|f| &\leq 2\left|\frac{\mu_z\sigma_w^2 - c\mu_w}{\mu_w^3\sigma}(t-\mu) + \frac{\sigma_w^2\sigma_z^2(1-\rho^2)}{2\mu_w^4\sigma^3}(t-\mu)^2\right| && (24)\\
&\leq 2\left|\frac{\mu_z\sigma_w^2 - c\mu_w}{\mu_w^3\sigma}(t-\mu)\right| + 2\left|\frac{\sigma_w^2\sigma_z^2(1-\rho^2)}{2\mu_w^4\sigma^3}(t-\mu)^2\right| && \text{(triangle inequality)}\\
&= 2\left|\frac{\mu_z\sigma_w^2 - c\mu_w}{\mu_w^3\sigma}(t-\mu)\right| + 2\left|\frac{\sigma_w^2\sigma_z^2(1-\rho^2)}{\mu_w^3\sigma}(t-\mu)\right| \cdot \left|\frac{t-\mu}{2\mu_w\sigma^2}\right| && (25)
\end{aligned}$$

Consider the interval $I = [\mu - 2|\mu_w|\sigma^2, \mu + 2|\mu_w|\sigma^2]$. Observe that for $t \in I$, $\left|\frac{t-\mu}{2\mu_w\sigma^2}\right| \leq 1$. For $t \notin I$, we have from Assumption 2 in the theorem statement that $f(t)$ is negligible, because

$$
\begin{aligned}
f(t) \to 0 &\impliedby (2|\mu_w|\sigma^2 \gg \sigma) \wedge (t \notin I) && \text{(exp}(\cdot)\text{ terms in } f(t) \text{ vanish)} \\
&\iff (2|\mu_w|\sigma \gg 1) \wedge (t \notin I) \\
&\iff (2|\mu_w|)^{-1}\sqrt{\mu_z^2\sigma_w^2 + \mu_w^2\sigma_z^2 - 2\rho\sigma_z\sigma_w\mu_z\mu_w} \gg 1) \wedge (t \notin I) && \text{(definition of } \sigma) \\
&\impliedby (2|\mu_w|)^{-1}\sqrt{\mu_z^2\sigma_w^2 + \mu_w^2\sigma_z^2 - 2\sigma_z\sigma_w\mu_z\mu_w} \gg 1) \wedge (t \notin I) && (\rho \in [-1, 1]) \\
&\iff (2|\mu_w|)^{-1}\sqrt{(\mu_z\sigma_w - \mu_w\sigma_z)^2} \gg 1) \wedge (t \notin I) && \text{(factoring inside } \sqrt{\cdot}) \\
&\iff (2|\mu_w|)^{-1}|\mu_z\sigma_w - \mu_w\sigma_z| \gg 1) \wedge (t \notin I) \\
&\iff (2|\mu_z|)\left|\frac{\sigma_w}{\mu_z} - \frac{\sigma_z}{\mu_z}\right| \gg 1) \wedge (t \notin I), && \text{(Rearranging terms)}
\end{aligned}
$$

which is precisely Assumption 2. We substitute the statement $\left|\frac{t-\mu}{2\mu_w\sigma^2}\right| \leq 1$ into (24) yielding

$$
\begin{aligned}
|f| &\leq \left|\frac{\mu_z\sigma_w^2 - c\mu_w}{\mu_w^3\sigma}(t-\mu)\right| + \left|\frac{\sigma_w^2\sigma_z^2(1-\rho^2)}{\mu_w^3\sigma}\right| \\
|f| &\leq \left|\frac{\mu_z\sigma_w^2 - c\mu_w}{\mu_w^3\sigma}\right||t-\mu| + \left|\frac{\sigma_w^2\sigma_z^2(1-\rho^2)}{\mu_w^3\sigma}\right||t-\mu| \qquad (26)
\end{aligned}
$$

Substituting (26) into (13) yields

$$
\begin{aligned}
\left|\mathbb{P}(Z/W \leq t) - \Phi\left(\frac{t-\mu}{\sigma}\right)\right| &\leq \frac{2}{\sqrt{2\pi}}\left(\left|\frac{\mu_z\sigma_w^2 - c\mu_w}{\mu_w^3\sigma}\right| + \left|\frac{\sigma_w^2\sigma_z^2(1-\rho^2)}{\mu_w^3\sigma}\right|\right) && (27) \\
&\leq \sqrt{\frac{2}{\pi}}\frac{1}{\sigma|\mu_w|^3}\left(\left|\mu_z\sigma_w^2 - c\mu_w\right| + \sigma_w^2\sigma_z^2(1-\rho^2)\right) && (\rho \in (-1,1)) \\
&\leq \sqrt{\frac{2}{\pi}}\frac{1}{\sigma|\mu_w|^3}\left(\left|\mu_z\sigma_w^2 - \rho\mu_w\sigma_w\sigma_z\right| + \sigma_w^2\sigma_z^2(1-\rho^2)\right) && (c = \rho\sigma_w\sigma_z) \\
&\leq \sqrt{\frac{2}{\pi}}\frac{1}{\sigma|\mu_w|^3}\left(\left|\mu_z\sigma_w^2\right| + |\rho\mu_w\sigma_w\sigma_z| + \sigma_w^2\sigma_z^2(1-\rho^2)\right) && \text{(triangle inequality)} \\
&= \sqrt{\frac{2}{\pi}}\frac{1}{\sigma|\mu_w|^3}\left(\left|\mu_z\sigma_w^2\right| - \sigma_w^2\sigma_z^2(|\rho| - \frac{\mu_w}{2\sigma_w\sigma_z})^2 + \frac{\mu_w^2}{4} + \sigma_w^2\sigma_z^2\right) \\
& && \text{(completing the square)} \\
&\leq \sqrt{\frac{2}{\pi}}\frac{1}{\sigma|\mu_w|^3}\left(\left|\mu_z\sigma_w^2\right| + \frac{\mu_w^2}{4} + \sigma_w^2\sigma_z^2\right) && (28) \\
&\leq \sqrt{\frac{2}{\pi}}\frac{1}{\sigma|\mu_w|^3}\left(\left|\mu_z\sigma_w^2\right| + \mu_w^2 + \sigma_w^2\sigma_z^2\right) && (29)
\end{aligned}
$$

and the result follows since t is arbitrary and $\mu_w > 0$ by assumption. □

Remark 9. *There is a nice interpretation of each of the terms in the upper bound of Theorem 2.*

- $\frac{1}{\mu_w^3}$ *: the further the mean of the denominator from zero the better the approximation. As* μ_w *grows there is less and less support to the distribution of the ratio below zero.*
- $\frac{1}{\sigma}$ *: the narrower the ratio distribution, the worse the approximation. We can think of broader distributions being more "forgiving".*
- $\left|\mu_z\sigma_w^2\right|$*: the closer the mean of the numerator to zero, the better the approximation.*
- μ_w^2 *: weakens the* $\frac{1}{\mu_w^3}$ *term to* $\frac{1}{\mu_w}$*.*
- $\sigma_w^2\sigma_z^2$ *: penalty term for larger variances.*

B.5 Ratios of Correlated Sums of Bernoullis

B.5.1 Proof of Theorem 3

Theorem 3 (Gaussian Approximation)**.** *Consider the random variables*

$$\widetilde{Z} = \alpha + \sum_{i=1}^n a_i Y_i, \quad \widetilde{W} = \beta + \sum_{i=1}^n b_i Y_i$$

for $0 < \alpha \leq \beta$, $a_i \geq 0$, $b_i > 0$, $b_i \geq a_i$, $Y_i \sim \mathcal{B}(p_i)$ *and mutually independent for* $p_i \in (0,1)$. *Moreover,* $v_* = \min_i\{p_i(1-p_i)\}$, $a_* = \min_i\{|a_i| \mid a_i > 0\}$, $b^* = \max_i\{|b_i|\}$ *and* $n_a = |\{a_i \mid a_i > 0\}| \leq n$. *The ratio distribution* $\widetilde{G}(t) = \mathbb{P}[\widetilde{Z}/\widetilde{W} \leq t]$ *is approximately Gaussian, that is*

$$\mathrm{d}_{\mathsf{KS}}(\widetilde{G}, \mathcal{N}(\mu,\sigma)) \leq \frac{C_0}{\sqrt{n_a v_*}}\frac{1+b^*}{a_*} + \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_w^2(|\mu_z| + \sigma_z^2) + \mu_w^2}{\sigma\mu_w^3},$$

where $C_0 = 0.5600$ *is a universal constant and*

$$\begin{aligned}
\mu_z &= \alpha + \textstyle\sum_{i=1}^n a_i p_i, \quad \sigma_z^2 = \sum_{i=1}^n a_i^2 p_i(1-p_i),\\
\mu_w &= \beta + \textstyle\sum_{i=1}^n b_i p_i, \quad \sigma_w^2 = \sum_{i=1}^n b_i^2 p_i(1-p_i),\\
\rho &= \tfrac{1}{\sigma_z\sigma_w}\textstyle\sum_{i=1}^n a_i b_i p_i(1-p_i),\\
\mu &= \tfrac{\mu_z}{\mu_w}, \quad \sigma^2 = \tfrac{{\mu_z}^2{\sigma_w}^2 + {\mu_w}^2{\sigma_z}^2 - 2\rho\sigma_z\sigma_w\mu_z\mu_w}{{\mu_w}^4}.
\end{aligned}$$

Proof. We wish to determine the ratio distribution $\widetilde{Z}/\widetilde{W}$. From Lemma 1 we have that $\widetilde{Z} \approx Z$ and $\widetilde{W} \approx W$ are approximately Gaussian. Moreover, for fixed $t \in \mathbb{R}$ we have that $\widetilde{Z} - t\widetilde{W}$ is close to the Gaussian distributed $Y = Z - tW$. Indeed we have

$$\begin{aligned}
\left|\mathbb{P}(\widetilde{Z}/\widetilde{W} \leq t) - \Phi(\frac{t-\mu}{\sigma})\right| &= \left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \Phi(\frac{t-\mu}{\sigma})\right| && \text{(re-arranging terms in first } \mathbb{P}(\,\cdot\,))\\
&= \left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \mathbb{P}(Z - tW \leq 0) + \mathbb{P}(Z - tW \leq 0) - \Phi(\frac{t-\mu}{\sigma})\right| && (30)\\
&\leq \left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \mathbb{P}(Z - tW \leq 0)\right| + \left|\mathbb{P}(Z - tW \leq 0) - \Phi(\frac{t-\mu}{\sigma})\right| && (31)\\
&\leq \left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \mathbb{P}(Z - tW \leq 0)\right| + \left|\mathbb{P}(Z/W \leq t) - \Phi(\frac{t-\mu}{\sigma})\right| && (32)
\end{aligned}$$

For the first term of (32), we have

$$\begin{aligned}
&\left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \mathbb{P}(Z - tW \leq 0)\right| && (33)\\
&= \left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \mathbb{P}(Z - t\widetilde{W} \leq 0) + \mathbb{P}(Z - t\widetilde{W} \leq 0) - \mathbb{P}(Z - tW \leq 0)\right| && (34)\\
&\leq \left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \mathbb{P}(Z - t\widetilde{W} \leq 0)\right| + \left|\mathbb{P}(Z - t\widetilde{W} \leq 0) - \mathbb{P}(Z - tW \leq 0)\right| && \text{(triangle inequality)}\\
&= \left|\mathbb{P}(\widetilde{Z} \leq t\widetilde{W}) - \mathbb{P}(Z \leq t\widetilde{W})\right| + \left|\mathbb{P}(W \geq Z/t) - \mathbb{P}(\widetilde{W} \geq Z/t)\right| && \text{(rearranging inequalities in } \mathbb{P}(\,\cdot\,))\\
&= \left|\mathbb{P}(\widetilde{Z} \leq t\widetilde{W}) - \mathbb{P}(Z \leq t\widetilde{W})\right| + \left|1 - \mathbb{P}(W \leq Z/t) - (1 - \mathbb{P}(\widetilde{W} \leq Z/t))\right| && (35)\\
&= \left|\mathbb{P}(Z \leq t\widetilde{W}) - \mathbb{P}(\widetilde{Z} \leq t\widetilde{W})\right| + \left|\mathbb{P}(W < Z/t) - \mathbb{P}(\widetilde{W} < Z/t)\right| && (36)\\
&= \left|\mathbb{P}(Z \leq t\widetilde{W}) - \mathbb{P}(\widetilde{Z} \leq t\widetilde{W})\right| + \left|\mathbb{P}(W \leq Z/t) - \mathbb{P}(\widetilde{W} \leq Z/t)\right|, && (37)
\end{aligned}$$

where (37) follows (36) from continuity of W and $\widetilde{W}$. Lemma 1 provides bounds on $\left|\mathbb{P}(Z \leq z) - \mathbb{P}(\widetilde{Z} \leq z)\right|$ and

$\left|\mathbb{P}(W \leq w) - \mathbb{P}(\widetilde{W} \leq w)\right|$ for arbitrary $z, w \in \mathbb{R}$, giving

$$\left|\mathbb{P}(\widetilde{Z} - t\widetilde{W} \leq 0) - \mathbb{P}(Z - tW \leq 0)\right| \leq \frac{C_0}{\sqrt{n_a v_*}} \frac{1+a^*}{2a_*} + \frac{C_0}{\sqrt{n v_*}} \frac{1+b^*}{2b_*} \tag{38}$$

$$\leq \frac{C_0}{\sqrt{n_a v_*}} \left(\frac{1+a^*}{2a_*} + \frac{1+b^*}{2b_*} \right) \leq \frac{C_0}{\sqrt{n_a v_*}} \frac{1+b^*}{a_*} \tag{39}$$

For the second term of (32) we invoke Theorem 2, which yields

$$\left|\mathbb{P}(Z/W \leq t) - \Phi(\frac{t-\mu}{\sigma})\right| \leq \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_w^2(|\mu_z| + \sigma_z^2) + \mu_w^2}{\sigma \mu_w^3}, \tag{40}$$

and from Lemma 1 we have that

$$\mu_z = \alpha + \sum_{i=1}^n a_i p_i, \quad \sigma_z^2 = \sum_{i=1}^n a_i^2 p_i(1-p_i), \quad \mu_w = \beta + \sum_{i=1}^n b_i p_i, \quad \sigma_w^2 = \sum_{i=1}^n b_i^2 p_i(1-p_i). \tag{41}$$

The quantity $\mathrm{Cov}[\widetilde{Z}, \widetilde{W}]$ is readily computed as $\sum_{i=1}^n a_i b_i p_i(1-p_i)$, from which we compute ρ using the standard formula $\rho_{X,Y} = \mathrm{Cov}[X,Y]/(\sigma_X \sigma_Y)$. The mean of the approximate distribution $\mu = \mu_z/\mu_w$ as given in Theorem 2.

Substituting (38) and (40) into (32) gives

$$\left|\mathbb{P}(\widetilde{Z}/\widetilde{W} \leq t) - \Phi(\frac{t-\mu}{\sigma})\right| \leq \frac{C_0}{\sqrt{n_a v_*}} \frac{1+b^*}{a_*} + \sqrt{\frac{2}{\pi}} \cdot \frac{\sigma_w^2(|\mu_z| + \sigma_z^2) + \mu_w^2}{\sigma \mu_w^3} \tag{42}$$

and the result follows since t is arbitrary. □

B.5.2 Theorem 3 Asymptotics

We require the following lemma.

Lemma 10. *Let $s(t) := \mu_w^{-1}\sqrt{\sigma_z^2 - 2tc + t^2\sigma_w^2}$, where $\mu_w, \sigma_w, \sigma_z > 0$ and $c \in [-\sigma_w\sigma_z, +\sigma_w, \sigma_z]$. Then, $s(t) \geq s(c/\sigma_w^2) = \mu_w^{-1}\sqrt{\sigma_z^2 - c^2/\sigma_w^2}$ for all $t \in \mathbb{R}$, that is, c/σ_w^2 minimises $s(t)$.*

Proof. The first derivative of $s(t)$ is

$$\frac{\partial s}{\partial t} = \frac{\sigma_w^2 t - c}{\mu_w(\sigma_z^2 - 2tc + t^2\sigma_w^2)^{1/2}},$$

with $\frac{\partial s}{\partial t} = 0$ having unique solution $t = c/\sigma_w^2$. The second derivative

$$\left.\frac{\partial^2 s}{\partial t^2}\right|_{t=c/\sigma_w^2} = \left.\frac{\sigma_w^2\sigma_z^2 - c^2}{\mu_w(\sigma_z^2 - 2tc + t^2\sigma_w^2)^{3/2}}\right|_{t=c/\sigma_w^2} = \frac{\sigma_w^2\sigma_z^2 - c^2}{\mu_w(\sigma_z^2(1-\rho^2))^{3/2}} = \frac{\sigma_w^2}{\mu_w\sigma_z\sqrt{1-\rho^2}} > 0$$

from which we deduce $t = c/\sigma_w^2$ minimises $s(t)$. □

Corollary 11 ($O(n^{-1/2})$ scaling)**.** *If in the assumptions of Theorem 3 we further impose the following:*

- *$b_i = 1$ for all $i \in [n]$*
- *$a_i = 0$ for all $i \in [n]$, apart from a constant fraction $r \in (0,1)$ of a_i that are equal to one, that is, $n_a = \lfloor rn \rfloor$.*
- *$p_i = p$ for all $i \in [n]$*

then we have that the ratio distribution $\widetilde{G}(t) = \mathbb{P}(\widetilde{Z}/\widetilde{W} \leq t)$ satisfies $\mathrm{d}_{\mathsf{KS}}(\widetilde{G}, \mathcal{N}(\mu, \sigma^2)) \sim O(n^{-1/2})$ as n grows large.

Proof. We shall directly use the bounds of Theorem 3. We have that

$$\begin{aligned} \mu_z &= \alpha + n_a p, \quad \sigma_z^2 = n_a p(1-p), \\ \mu_w &= \beta + np, \quad \sigma_w^2 = np(1-p), \\ \rho &= \tfrac{1}{\sqrt{n n_a p(1-p)}} n_a p(1-p), \\ a_* &= b^* = 1 \quad v_* = p(1-p) \end{aligned} \tag{43}$$

Moreover,

$$\sigma \geq \mu_w^{-1}\sqrt{\sigma_z^2 - \frac{c^2}{\sigma_w^2}} \qquad \text{(Lemma 10)}$$
$$= \frac{n_a}{np}\sqrt{p(1-p)\left(1-\frac{1}{n^2}\right)} \geq \frac{r(n-1)}{np}\sqrt{p(1-p)\left(1-\frac{1}{n^2}\right)} \qquad \text{(Using Eq. (43))}$$

Therefore, we have

$$\frac{1}{\sigma\mu_w^3} \leq \frac{1}{(\beta+np)^3}\frac{np}{r(n-1)}\left(p(1-p)\left(1-\frac{1}{n^2}\right)\right)^{-\frac{1}{2}} \underset{n\to\infty}{\longrightarrow} \frac{1}{n^3p^2r}(p(1-p))^{-\frac{1}{2}} \tag{44}$$

Moreover, we have

$$\begin{aligned}\sigma_w^2(|\mu_z|+\sigma_z^2)+\mu_w^2 &= np(1-p)[\lfloor nr\rfloor p + \lfloor nr\rfloor p(1-p)] + n^2p^2(1-p)^2 \\ &\leq n^2p^2r + n^2p^2(1-p)^2(1+r) \qquad (45)\\ &\leq 3n^2p^2 \qquad (0<r,p<1)\end{aligned}$$

Eqs. (44) and (45) combined imply

$$\frac{\sigma_w^2(|\mu_z|+\sigma_z^2)+\mu_w^2}{\sigma\mu_w^3} \underset{n\to\infty}{\longrightarrow} O(n^{-1}). \tag{46}$$

Direct substitution of Eqs. (46) and (43) into Theorem 3 gives

$$\mathrm{d_{KS}}(F_{\widetilde{Z}/\widetilde{W}}, \mathcal{N}(\mu,\sigma^2)) \leq \frac{2C_0}{\sqrt{rnp(1-p)}} + O(n^{-1}) \underset{n\to\infty}{\longrightarrow} O(n^{-1/2}), \tag{47}$$

which follows since C_0, r and p are constants. □

B.6 Robustness

B.6.1 Proof of Theorem 4

Theorem 4 (Robustness)**.** *The conclusion of Theorem 3 remains valid if we relax the assumption that $Y_i \sim \mathcal{B}(p_i)$ for $p_i \in (0,1)$ to the following: the random variables $Y_i \sim \mathcal{B}(P_i)$, where $P_i \sim \mathrm{Beta}(\alpha_i,\beta_i)$ such that $\mathbb{E}[P_i] = p_i$ and $\mathbb{V}[P_i] < v_*$ for all $i \in [n]$.*

Proof. We shall demonstrate that if $\mathbb{V}[P_i] < v_*$, then $X_i \sim \mathcal{B}(p_i)$ for all $i \in [n]$ which is the exact setting for Theorem 3.

A $\mathrm{Beta}(\alpha,\beta)$-distributed random variable P can be alternatively specified by its expectation and variance, whence

$$\alpha = \left(\frac{\mathbb{E}[P](1-\mathbb{E}[P])}{\mathbb{V}[P]} - 1\right)\mathbb{E}[P], \quad \beta = \left(\frac{\mathbb{E}[P](1-\mathbb{E}[P])}{\mathbb{V}[P]} - 1\right)(1-\mathbb{E}[P]). \tag{48}$$

Suppose we impose that $\mathbb{E}[P] = p$ for some $p \in (0,1)$ and let $\mathbb{V}[P] = v$. We thus have from (48) that

$$\alpha = \left(\frac{p(1-p)}{v} - 1\right)p, \quad \beta = \left(\frac{p(1-p)}{v} - 1\right)(1-p). \tag{49}$$

A $\mathrm{Beta}(\alpha,\beta)$ distribution is well-formed when both $\alpha,\beta > 0$. Substitution into (49) then specifies that P follows a valid $\mathrm{Beta}(\alpha,\beta)$ distribution when

$$\left(\frac{p(1-p)}{v} - 1\right) > 0 \iff p(1-p) > v.$$

By definition, $v_* = \min_i\{p_i(1-p_i)\}$ and so we have that the distributions $P_i \sim \text{Beta}(\alpha_i, \beta_i)$ are all well formed when $\mathbb{V}[P_i] < v_*$.

It is well-known that the distribution $Y \sim \mathcal{B}(P)$, where $P \sim \text{Beta}(\alpha, \beta)$ is identical to $Y \sim \mathcal{B}(\frac{\alpha}{\alpha+\beta})$. We also have that $\mathbb{E}[P] = \frac{\alpha}{\alpha+\beta}$. By assumption, we impose that $\mathbb{E}[P_i] = p_i$ for all $i \in [n]$, giving that $\frac{\alpha_i}{\alpha_i+\beta_i} = p_i$. Thus, $Y_i \sim \mathcal{B}(p_i)$ which is precisely the setup of Theorem 3. □

Remark 12. *Theorem 4 is a relaxation of Theorem 3 because for $P_i \sim \text{Beta}(\alpha_i, \beta_i)$ such that $\mathbb{E}[P_i] = p_i$, in the limit $\mathbb{V}[P_i] \to 0$, we have that $P_i \sim \mathbb{1}\{P_i = p_i\}$.*

B.7 Estimators, Mean and Covariance for ROC-AUC

The probability query corresponding to ROC-AUC is given in (3) as

$$P(\hat{y}(X) \geq \hat{y}(X') \mid Y = 1,\ Y' = 0).$$

We can use the definition of conditional expectation to rewrite (3) as

$$\frac{P(\hat{y}(X) \geq \hat{y}(X'),\ Y = 1,\ Y' = 0)}{P(Y = 1,\ Y' = 0)}. \tag{50}$$

Suppose $n \in \mathbb{N}$ and we have an index set $[n] := \{1, \ldots, n\}$ for evaluation set $D_n = \{\langle x_i, y_i, m_i\rangle\}_{i\in[n]}$. Assuming briefly the vanilla case where $m_i = 0$ for all $i \in [n]$, we can substitute in the usual estimators into (50) for the ratio like so

$$\widehat{Q}_{\text{roc-auc},n} = \frac{\widehat{P}(\hat{y}(X) \geq \hat{y}(X'),\ Y = 1,\ Y' = 0)}{\widehat{P}(Y = 1,\ Y' = 0)} = \frac{\sum_{i\in[n]}\sum_{j\in[n]} \mathbb{1}\{\hat{y}(x_i) \geq \hat{y}(x_j)\}\mathbb{1}\{y_i = 1\}\mathbb{1}\{y_j = 0\}}{\sum_{i\in[n]}\sum_{j\in[n]} \mathbb{1}\{y_i = 1\}\mathbb{1}\{y_j = 0\}} \tag{51}$$

Now we allow some missing labels $m_i = 1 \Leftrightarrow y_i = \texttt{NA}$. We have a partition into known $\mathcal{K} \subset [n]$ and unknown $\overline{\mathcal{K}} \subset [n]$ indices, that is, $\mathcal{K} \cap \overline{\mathcal{K}} = \emptyset$, $\mathcal{K} \cup \overline{\mathcal{K}} = [n]$. Define the random variables Y_i for $i \in [n]$ such that

$$Y_i \sim \begin{cases} \mathcal{B}(p_i), & i \in \overline{\mathcal{K}}; \\ \mathbb{1}\{Y_i = y_i\}, & i \in \mathcal{K}, \end{cases} \tag{52}$$

and $\mathcal{B}(p)$ is a Bernoulli distribution with parameter $p \in (0, 1)$ and the $y_i \in \{0, 1\}$ are constants for $i \in \mathcal{K}$.

The estimator for ROC-AUC when labels are missing now becomes

$$\widehat{Q}^{(S)}_{\text{roc-auc},n} = \frac{\sum_{i\in[n]}\sum_{j\in[n]} \mathbb{1}\{\hat{y}(x_i) \geq \hat{y}(x_j)\}Y_i(1-Y_j)}{\sum_{i\in[n]}\sum_{j\in[n]} Y_i(1-Y_j)} \tag{53}$$

We shall require Lemma 13 for the mean and covariance of $\widehat{Q}^{(S)}_{\text{roc-auc},n}$.

Lemma 13. *Let $\xi_{i,j}, \zeta_{i,j} \in \mathbb{R}$ for all $i, j \in [n]$. Then*

$$\mathbb{E}\left[\sum_{i,j\in[n]} \xi_{i,j}Y_i(1-Y_j)\right] = \sum_{i,j\in[n]} \xi_{i,j}\omega_{i,j},$$

where

$$\omega_{i,j} =: \mathbb{E}[Y_i(1-Y_j)] = \begin{cases} y_i(1-y_j), & (i \neq j) \wedge (i, j \in \mathcal{K}); \\ p_i(1-y_j), & (i \neq j) \wedge (i \in \overline{\mathcal{K}},\ j \in \mathcal{K}); \\ y_i(1-p_j), & (i \neq j) \wedge (i \in \mathcal{K},\ j \in \overline{\mathcal{K}}); \\ p_i(1-p_j), & (i \neq j) \wedge (i, j \in \overline{\mathcal{K}}); \\ 0, & \textit{otherwise.} \end{cases}$$

Moreover,

$$\mathrm{Cov}\left[\sum_{i,j\in[n]}\xi_{i,j}Y_i(1-Y_j),\sum_{i,j\in[n]}\zeta_{i,j}Y_i(1-Y_j)\right]=\sum_{i,j,k,l\in[n]}\xi_{i,j}\zeta_{k,l}\cdot\left(\nu_{i,k}-\Gamma_{i,j,k}-\Gamma_{k,l,i}+\eta_{i,j,k,l}\right),$$

where

$$\nu_{i,j}=:\mathbb{E}[Y_iY_j]-\mathbb{E}[Y_i]\mathbb{E}[Y_j]=\begin{cases}p_i(1-p_i), & (i=j)\wedge(i,j\in\overline{\mathcal{K}});\\ 0, & \textit{otherwise}\end{cases}$$

and

$$\Gamma_{i,j,k}=:\mathbb{E}[Y_iY_jY_k]-\mathbb{E}[Y_iY_j]\mathbb{E}[Y_k]=\begin{cases}p_i(1-p_i), & (i=j=k)\wedge(i,j,k\in\overline{\mathcal{K}});\\ p_ip_j(1-p_i), & (i=k\neq j)\wedge(i,j,k\in\overline{\mathcal{K}});\\ p_ip_j(1-p_j), & (k=j\neq i)\wedge(i,j,k\in\overline{\mathcal{K}});\\ y_ip_k(1-p_k), & (j=k\neq i)\wedge(i\in\mathcal{K},\ j,k\in\overline{\mathcal{K}});\\ y_jp_k(1-p_k), & (i=k\neq j)\wedge(j\in\mathcal{K},\ i,k\in\overline{\mathcal{K}});\\ 0, & \textit{otherwise}\end{cases}$$

and

$$\eta_{i,j,k,l}=:\mathbb{E}[Y_iY_jY_kY_l]-\mathbb{E}[Y_iY_j]\mathbb{E}[Y_kY_l]=\begin{cases}p_i(1-p_i), & (i=j=k=l)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_l(1-p_i), & (i=j=k\neq l)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_k(1-p_i), & (i=j=l\neq k)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_j(1-p_i), & (i=k=l\neq j)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_j(1-p_j), & (j=k=l\neq i)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_j(1-p_ip_j), & (i=k\neq j=l)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_j(1-p_ip_j), & (i=l\neq j=k)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_jp_l(1-p_i), & (i=k\neq j\neq l\neq i)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_jp_k(1-p_j), & (j=l\neq i\neq k\neq j)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_jp_k(1-p_i), & (i=l\neq j\neq k\neq i)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ p_ip_jp_l(1-p_j), & (j=k\neq i\neq l\neq j)\wedge(i,j,k,l\in\overline{\mathcal{K}});\\ y_i\Gamma_{k,l,j}, & i\in\mathcal{K};\\ y_j\Gamma_{k,l,i}, & j\in\mathcal{K};\\ y_k\Gamma_{i,j,l}, & k\in\mathcal{K};\\ y_l\Gamma_{i,j,k}, & l\in\mathcal{K};\\ 0, & \textit{otherwise.}\end{cases}$$

Proof. Follows readily from the following: linearity of expectation; covariance of linear combinations; the fact that $\mathbb{E}[Y]=p$, $\mathbb{V}[Y]=p(1-p)$ for $Y\sim\mathcal{B}(p)$; $\mathrm{Cov}[U,V]=0$ for independent r.v.s U,V.

□

Remark 14 (ROC-AUC Estimator)**.** *We can substitute $\xi_{i,j}=\mathbb{1}\{\hat{y}(x_i)\geq\hat{y}(x_j)\}$ and $\zeta_{i,j}=1$ into Lemma 13 to recover the mean and covariance matrix of the numerator and denominator of the estimator $\widehat{Q}^{(S)}_{roc\text{-}auc,n}$. The mean and variance of the ratio of Gaussians computed using Theorem 2 can then be readily used with PEMI-Gauss algorithm.*

We note that computing the covariance matrix is intractable for large n, since there are $O(n^4)$ components in the sum. In practise, for $n>120$ we use PEMI with $B=10000$ and empirically compute $\hat{\mu}$ and $\widehat{\sigma^2}$ from the returned distribution for use within PEMI-Gauss.

C Supplementary Empirical Results

All experiments were conducted using an AWS EC2 `c5a.12xlarge` instance and datasets were downloaded using the OpenML API (Feurer et al., 2021).

C.1 MCAR Regime

C.1.1 Missingness Fraction $p_m = 0.1$

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		(0.2688, 0.311)	(0.2879, 0.3371)	(0.2691, 0.312)	(0.2843, 0.3351)	(0.7648, 0.8359)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.3908, 0.5)	(0.5333, 0.6333)	(0.6667, 0.7333)	(0.575, 0.6667)	(0.6083, 0.6917)
	p calibrated	**(0.1083, 0.175)**	**(0.1083, 0.1833)**	**(0.0825, 0.15)**	**(0.1, 0.175)**	**(0.1, 0.175)**
	$p = N_+/N$	(0.5917, 0.675)	(0.5417, 0.6417)	(0.625, 0.7083)	(0.575, 0.6583)	(0.6, 0.6833)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.425, 0.5333)	(0.625, 0.725)	(0.7167, 0.8)	(0.6667, 0.7583)	(0.7, 0.7833)
	p calibrated	**(0.0833, 0.1583)**	**(0.075, 0.1417)**	**(0.1083, 0.2167)**	**(0.0667, 0.125)**	**(0.0833, 0.1833)**
	$p = N_+/N$	(0.625, 0.7083)	(0.6417, 0.7333)	(0.65, 0.75)	(0.5917, 0.6833)	(0.6583, 0.7508)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.4933, 0.6067)	(0.6632, 0.7567)	(0.765, 0.8467)	(0.705, 0.7917)	(0.715, 0.795)
	p calibrated	**(0.115, 0.21)**	**(0.065, 0.1617)**	(0.1583, 0.2683)	**(0.0667, 0.1633)**	**(0.1217, 0.2317)**
	$p = N_+/N$	(0.66, 0.75)	(0.6717, 0.7633)	(0.7, 0.7917)	(0.6617, 0.7483)	(0.7017, 0.7967)
PEMI-Gauss	$p = \frac{1}{2}$	(0.3987, 0.5108)	(0.6587, 0.7536)	(0.7289, 0.8289)	(0.6893, 0.781)	(0.7299, 0.8065)
	p calibrated	**(0.0666, 0.1171)**	**(0.0606, 0.1423)**	**(0.0922, 0.2058)**	**(0.0714, 0.1521)**	**(0.1499, 0.2583)**
	$p = N_+/N$	(0.5865, 0.6821)	(0.6403, 0.7288)	(0.6522, 0.7477)	(0.6525, 0.7424)	(0.6937, 0.7853)

Table 5: KS-distance of PIT for different algorithms in MCAR setting (all datasets). 90% confidence intervals. $p_m = 0.1$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		(0.1512, 0.169)	(0.1519, 0.1677)	(0.0977, 0.1224)	(0.153, 0.1682)	(0.4526, 0.4712)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.1842, 0.2467)	(0.2892, 0.335)	(0.3433, 0.375)	(0.3117, 0.3492)	(0.3317, 0.3642)
	p calibrated	**(0.0408, 0.0792)**	**(0.0425, 0.0784)**	**(0.0317, 0.0808)**	**(0.0374, 0.1067)**	**(0.0475, 0.0867)**
	$p = N_+/N$	(0.3058, 0.3475)	(0.2917, 0.3383)	(0.3283, 0.3633)	(0.3142, 0.3508)	(0.3025, 0.3458)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.2087, 0.273)	(0.3388, 0.3854)	(0.4, 0.4246)	(0.3737, 0.4084)	(0.3825, 0.4133)
	p calibrated	**(0.0346, 0.0863)**	**(0.03, 0.0679)**	**(0.0554, 0.1171)**	**(0.0229, 0.0521)**	**(0.0396, 0.0858)**
	$p = N_+/N$	(0.3554, 0.395)	(0.3512, 0.3934)	(0.3733, 0.4067)	(0.3592, 0.3925)	(0.3592, 0.4004)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.2465, 0.3173)	(0.386, 0.4301)	(0.4475, 0.4694)	(0.4154, 0.449)	(0.4224, 0.4533)
	p calibrated	(0.0615, 0.1238)	**(0.0234, 0.0734)**	**(0.0776, 0.1471)**	**(0.0255, 0.0739)**	**(0.0523, 0.1156)**
	$p = N_+/N$	(0.3915, 0.4325)	(0.3875, 0.4297)	(0.4118, 0.4456)	(0.3992, 0.4354)	(0.4076, 0.4445)
PEMI-Gauss	$p = \frac{1}{2}$	(0.1979, 0.2696)	(0.3795, 0.4268)	(0.4168, 0.4515)	(0.4088, 0.4455)	(0.4339, 0.4619)
	p calibrated	**(0.0245, 0.0529)**	**(0.0218, 0.0569)**	**(0.035, 0.0977)**	**(0.0269, 0.064)**	**(0.0591, 0.1203)**
	$p = N_+/N$	(0.3502, 0.4014)	(0.37, 0.4199)	(0.3881, 0.4304)	(0.3841, 0.4278)	(0.415, 0.4495)

Table 6: W_1 distance of PIT for different algorithms in MCAR setting (all datasets). 90% confidence intervals. $p_m = 0.1$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		**(0.0074, 0.0107)**	(0.0067, 0.0098)	(0.0055, 0.0071)	**(0.0057, 0.0085)**	(0.0931, 0.1073)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.0165, 0.0206)	(0.0409, 0.0529)	(0.0248, 0.0284)	(0.0334, 0.0417)	(0.0495, 0.0638)
	p calibrated	**(0.0071, 0.0103)**	**(0.0054, 0.0077)**	**(0.0038, 0.0056)**	**(0.0046, 0.0067)**	**(0.0052, 0.0072)**
	$p = N_+/N$	(0.0195, 0.0245)	(0.0194, 0.0236)	(0.0122, 0.0145)	(0.0194, 0.0237)	(0.0269, 0.0317)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.0163, 0.0205)	(0.0396, 0.0516)	(0.024, 0.0277)	(0.033, 0.0415)	(0.0495, 0.0638)
	p calibrated	**(0.0069, 0.0096)**	**(0.0045, 0.0063)**	**(0.0035, 0.0052)**	**(0.004, 0.0061)**	**(0.0045, 0.0063)**
	$p = N_+/N$	(0.0196, 0.0245)	(0.0202, 0.0244)	(0.0126, 0.0148)	(0.0193, 0.0233)	(0.026, 0.0308)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.0163, 0.0201)	(0.0397, 0.0516)	(0.0241, 0.0276)	(0.0332, 0.0416)	(0.0495, 0.0636)
	p calibrated	**(0.0063, 0.009)**	**(0.0041, 0.006)**	**(0.0035, 0.0049)**	**(0.0043, 0.0061)**	**(0.0044, 0.0059)**
	$p = N_+/N$	(0.0193, 0.0242)	(0.0202, 0.0244)	(0.0121, 0.0144)	(0.0194, 0.0235)	(0.0261, 0.0307)
PEMI-Gauss	$p = \frac{1}{2}$	(0.0161, 0.02)	(0.04, 0.052)	(0.0241, 0.0276)	(0.0333, 0.0416)	(0.0496, 0.0638)
	p calibrated	**(0.0063, 0.009)**	**(0.0042, 0.0061)**	**(0.0035, 0.0049)**	**(0.0043, 0.0061)**	**(0.0043, 0.0059)**
	$p = N_+/N$	(0.0192, 0.024)	(0.02, 0.0242)	(0.0121, 0.0143)	(0.0194, 0.0235)	(0.0262, 0.0309)

Table 7: MAE of centre of evaluation metric predictive distributions (all datasets), MCAR setting. 90% confidence intervals. $p_m = 0.1$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		**(0.0134, 0.0198)**	(0.0125, 0.0183)	**(0.0082, 0.0104)**	**(0.0102, 0.0167)**	(0.1086, 0.1236)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.0227, 0.0283)	(0.0631, 0.0749)	(0.0289, 0.0323)	(0.0474, 0.0557)	(0.0741, 0.0909)
	p calibrated	**(0.0122, 0.019)**	**(0.0095, 0.0134)**	**(0.0075, 0.0105)**	**(0.0084, 0.0122)**	**(0.009, 0.0117)**
	$p = N_+/N$	(0.0276, 0.0345)	(0.026, 0.0298)	(0.0154, 0.0182)	(0.0264, 0.0312)	(0.0334, 0.038)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.0224, 0.0283)	(0.0623, 0.0739)	(0.0283, 0.0316)	(0.047, 0.0551)	(0.0736, 0.0906)
	p calibrated	**(0.0116, 0.0166)**	**(0.008, 0.0113)**	**(0.0069, 0.0097)**	**(0.0077, 0.0128)**	**(0.0082, 0.0107)**
	$p = N_+/N$	(0.0276, 0.034)	(0.0268, 0.0307)	(0.0156, 0.0182)	(0.0258, 0.03)	(0.0325, 0.0368)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.0217, 0.0267)	(0.0622, 0.074)	(0.0281, 0.0314)	(0.0471, 0.0552)	(0.0735, 0.0903)
	p calibrated	**(0.0109, 0.0157)**	**(0.0076, 0.0113)**	**(0.0065, 0.0087)**	**(0.0074, 0.011)**	**(0.0076, 0.0098)**
	$p = N_+/N$	(0.0274, 0.0342)	(0.0269, 0.0307)	(0.0151, 0.0175)	(0.0259, 0.0304)	(0.0322, 0.0364)
PEMI-Gauss	$p = \frac{1}{2}$	(0.0216, 0.0266)	(0.0624, 0.0742)	(0.0282, 0.0314)	(0.047, 0.0552)	(0.0736, 0.0903)
	p calibrated	**(0.0109, 0.0156)**	**(0.0076, 0.0113)**	**(0.0065, 0.0086)**	**(0.0073, 0.0111)**	**(0.0075, 0.0097)**
	$p = N_+/N$	(0.0271, 0.0336)	(0.0267, 0.0305)	(0.0151, 0.0176)	(0.0259, 0.0304)	(0.0323, 0.0365)

Table 8: RMSE of centre of evaluation metric predictive distributions (all datasets), MCAR setting. 90% confidence intervals. $p_m = 0.1$. Best results in **bold**.

C.1.2 Missingness Fraction $p_m = 0.2$

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		(0.2705, 0.3091)	(0.2337, 0.285)	(0.2174, 0.2737)	(0.2336, 0.2883)	(0.7302, 0.795)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.3667, 0.4833)	(0.6, 0.6917)	(0.7333, 0.7833)	(0.675, 0.7417)	(0.6833, 0.75)
	p calibrated	**(0.1083, 0.1833)**	**(0.125, 0.2092)**	**(0.0833, 0.1583)**	**(0.1333, 0.2167)**	**(0.1158, 0.175)**
	$p = N_+/N$	(0.5667, 0.65)	(0.5833, 0.675)	(0.6083, 0.6917)	(0.6, 0.6833)	(0.6833, 0.75)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.4167, 0.525)	(0.65, 0.7425)	(0.7917, 0.8583)	(0.75, 0.8167)	(0.7667, 0.8258)
	p calibrated	**(0.05, 0.1167)**	**(0.075, 0.1333)**	**(0.0833, 0.175)**	**(0.0917, 0.175)**	**(0.1, 0.1917)**
	$p = N_+/N$	(0.6667, 0.7333)	(0.65, 0.7333)	(0.65, 0.7333)	(0.6833, 0.7508)	(0.6917, 0.7833)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.445, 0.5567)	(0.6998, 0.7883)	(0.84, 0.8833)	(0.7783, 0.85)	(0.835, 0.8917)
	p calibrated	**(0.055, 0.115)**	**(0.0617, 0.1483)**	**(0.1167, 0.215)**	**(0.075, 0.1335)**	**(0.125, 0.2217)**
	$p = N_+/N$	(0.7133, 0.7867)	(0.6967, 0.775)	(0.7317, 0.7933)	(0.705, 0.7767)	(0.77, 0.8317)
PEMI-Gauss	$p = \frac{1}{2}$	(0.4137, 0.5303)	(0.7024, 0.7901)	(0.8298, 0.8798)	(0.7833, 0.8509)	(0.8471, 0.893)
	p calibrated	**(0.0877, 0.181)**	**(0.0592, 0.1378)**	**(0.0907, 0.1893)**	**(0.0771, 0.1408)**	**(0.1458, 0.2458)**
	$p = N_+/N$	(0.6369, 0.7239)	(0.6886, 0.7589)	(0.6742, 0.7412)	(0.6922, 0.7675)	(0.7795, 0.8435)

Table 9: KS-distance of PIT for different algorithms in MCAR setting (all datasets). 90% confidence intervals. $p_m = 0.2$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		(0.1385, 0.1573)	(0.1201, 0.1406)	(0.0895, 0.1265)	(0.1293, 0.1502)	(0.4432, 0.4644)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.1657, 0.2308)	(0.325, 0.3618)	(0.375, 0.3917)	(0.355, 0.3792)	(0.3625, 0.3833)
	p calibrated	(0.0508, 0.1192)	**(0.0542, 0.1208)**	**(0.0325, 0.0684)**	(0.0658, 0.1292)	**(0.0417, 0.09)**
	$p = N_+/N$	(0.3283, 0.3583)	(0.3258, 0.3592)	(0.3367, 0.3658)	(0.33, 0.3625)	(0.3583, 0.3817)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.1867, 0.2554)	(0.3808, 0.41)	(0.4125, 0.4375)	(0.4117, 0.4321)	(0.4125, 0.4371)
	p calibrated	**(0.0212, 0.058)**	**(0.0254, 0.0571)**	**(0.0354, 0.0888)**	**(0.0341, 0.0838)**	**(0.0408, 0.0971)**
	$p = N_+/N$	(0.3817, 0.41)	(0.3763, 0.4104)	(0.3842, 0.4129)	(0.3983, 0.42)	(0.3958, 0.4208)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.232, 0.3021)	(0.4202, 0.4513)	(0.4703, 0.4815)	(0.4527, 0.471)	(0.4638, 0.479)
	p calibrated	**(0.0197, 0.0469)**	**(0.0234, 0.0713)**	**(0.0615, 0.1291)**	**(0.0263, 0.0553)**	**(0.0666, 0.1315)**
	$p = N_+/N$	(0.433, 0.4556)	(0.419, 0.4491)	(0.4393, 0.4591)	(0.4328, 0.4558)	(0.4442, 0.4679)
PEMI-Gauss	$p = \frac{1}{2}$	(0.1972, 0.2702)	(0.4252, 0.4555)	(0.4627, 0.4807)	(0.4529, 0.4723)	(0.4696, 0.4843)
	p calibrated	**(0.0344, 0.0925)**	**(0.0202, 0.0634)**	**(0.0383, 0.1006)**	**(0.029, 0.0577)**	**(0.0816, 0.1453)**
	$p = N_+/N$	(0.4089, 0.4389)	(0.4151, 0.4485)	(0.4233, 0.4503)	(0.4313, 0.4564)	(0.4529, 0.4732)

Table 10: W_1 distance of PIT for different algorithms in MCAR setting (all datasets). 90% confidence intervals. $p_m = 0.2$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		**(0.0086, 0.0119)**	(0.0095, 0.0132)	**(0.0068, 0.009)**	**(0.0071, 0.0098)**	(0.092, 0.106)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.0287, 0.0347)	(0.0678, 0.0867)	(0.047, 0.0535)	(0.0584, 0.0726)	(0.0813, 0.1003)
	p calibrated	**(0.0088, 0.0121)**	**(0.0073, 0.0101)**	**(0.0057, 0.0083)**	**(0.0055, 0.0076)**	**(0.0068, 0.0099)**
	$p = N_+/N$	(0.0359, 0.0445)	(0.0371, 0.0449)	(0.0224, 0.0267)	(0.0375, 0.0459)	(0.0509, 0.0586)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.0275, 0.0337)	(0.0681, 0.0865)	(0.0478, 0.0543)	(0.0595, 0.0736)	(0.0827, 0.1011)
	p calibrated	**(0.0095, 0.0129)**	**(0.0067, 0.0093)**	**(0.0056, 0.0079)**	**(0.0066, 0.0094)**	**(0.0066, 0.0094)**
	$p = N_+/N$	(0.0374, 0.0459)	(0.0376, 0.0455)	(0.0227, 0.0267)	(0.0378, 0.0458)	(0.0498, 0.0572)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.0273, 0.0334)	(0.0679, 0.0865)	(0.0473, 0.0535)	(0.0594, 0.0734)	(0.0823, 0.1009)
	p calibrated	**(0.0091, 0.0125)**	**(0.0064, 0.0089)**	**(0.0051, 0.0073)**	**(0.006, 0.0083)**	**(0.006, 0.0086)**
	$p = N_+/N$	(0.0364, 0.0451)	(0.0373, 0.045)	(0.0227, 0.0265)	(0.0372, 0.045)	(0.0497, 0.0571)
PEMI-Gauss	$p = \frac{1}{2}$	(0.027, 0.033)	(0.0681, 0.0867)	(0.0475, 0.0537)	(0.059, 0.0732)	(0.082, 0.1005)
	p calibrated	**(0.009, 0.0122)**	**(0.0063, 0.0087)**	**(0.0052, 0.0075)**	**(0.006, 0.0084)**	**(0.0061, 0.0086)**
	$p = N_+/N$	(0.0363, 0.0449)	(0.0373, 0.045)	(0.0226, 0.0264)	(0.0373, 0.0451)	(0.0497, 0.0572)

Table 11: MAE of centre of evaluation metric predictive distributions (all datasets), MCAR setting. 90% confidence intervals. $p_m = 0.2$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		**(0.0143, 0.02)**	(0.0162, 0.023)	**(0.0104, 0.0137)**	**(0.0121, 0.017)**	(0.1083, 0.1233)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.0374, 0.0436)	(0.1027, 0.1202)	(0.0545, 0.0605)	(0.0817, 0.0953)	(0.1113, 0.1326)
	p calibrated	**(0.0145, 0.0209)**	**(0.0127, 0.0169)**	**(0.0104, 0.0148)**	**(0.0094, 0.0133)**	**(0.0128, 0.0176)**
	$p = N_+/N$	(0.0492, 0.059)	(0.0494, 0.0567)	(0.028, 0.0334)	(0.0506, 0.0588)	(0.0605, 0.0671)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.0366, 0.043)	(0.1027, 0.1199)	(0.0549, 0.061)	(0.0825, 0.0959)	(0.1119, 0.1331)
	p calibrated	**(0.0152, 0.022)**	**(0.0116, 0.0168)**	**(0.0099, 0.0136)**	**(0.0118, 0.0161)**	**(0.0122, 0.017)**
	$p = N_+/N$	(0.0505, 0.0603)	(0.0498, 0.0571)	(0.0281, 0.0325)	(0.0501, 0.0582)	(0.0594, 0.0659)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.0363, 0.0423)	(0.1025, 0.12)	(0.0541, 0.0601)	(0.0821, 0.0956)	(0.1116, 0.1329)
	p calibrated	**(0.0149, 0.0208)**	**(0.0112, 0.0149)**	**(0.0093, 0.0127)**	**(0.0102, 0.0142)**	**(0.0111, 0.0153)**
	$p = N_+/N$	(0.0496, 0.0595)	(0.0495, 0.0566)	(0.0275, 0.0318)	(0.0492, 0.0571)	(0.0592, 0.0658)
PEMI-Gauss	$p = \frac{1}{2}$	(0.0362, 0.0421)	(0.1026, 0.12)	(0.0542, 0.0602)	(0.0819, 0.0955)	(0.1112, 0.1327)
	p calibrated	**(0.0145, 0.0202)**	**(0.0109, 0.0147)**	**(0.0094, 0.0129)**	**(0.0104, 0.0144)**	**(0.0111, 0.0153)**
	$p = N_+/N$	(0.0496, 0.0593)	(0.0495, 0.0565)	(0.0275, 0.0318)	(0.0492, 0.0572)	(0.0592, 0.0657)

Table 12: RMSE of centre of evaluation metric predictive distributions (all datasets), MCAR setting. 90% confidence intervals. $p_m = 0.2$. Best results in **bold**.

C.1.3 Missingness Fraction $p_m = 0.3$

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		(0.1759, 0.2381)	**(0.1504, 0.2124)**	(0.14, 0.1984)	(0.1478, 0.2093)	(0.7004, 0.7812)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.35, 0.4583)	(0.6667, 0.7333)	(0.5917, 0.7833)	(0.7167, 0.7667)	(0.725, 0.775)
	p calibrated	**(0.1075, 0.1833)**	**(0.15, 0.2417)**	**(0.1333, 0.2167)**	**(0.1167, 0.2)**	**(0.1167, 0.2083)**
	$p = N_+/N$	(0.625, 0.7)	(0.6583, 0.725)	(0.5833, 0.6667)	(0.6667, 0.7333)	(0.6917, 0.7583)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.475, 0.5833)	(0.7167, 0.7917)	(0.7917, 0.8833)	(0.7167, 0.7917)	(0.8075, 0.8583)
	p calibrated	**(0.0833, 0.1758)**	**(0.0917, 0.1583)**	**(0.075, 0.1333)**	**(0.0667, 0.1333)**	**(0.1242, 0.2167)**
	$p = N_+/N$	(0.6917, 0.7667)	(0.675, 0.75)	(0.625, 0.7083)	(0.6833, 0.7667)	(0.7917, 0.8583)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.4883, 0.5983)	(0.7417, 0.82)	(0.88, 0.93)	(0.8167, 0.8767)	(0.88, 0.9283)
	p calibrated	**(0.0767, 0.155)**	**(0.075, 0.1617)**	**(0.085, 0.1817)**	**(0.065, 0.1383)**	**(0.1533, 0.2483)**
	$p = N_+/N$	(0.7317, 0.8033)	(0.7465, 0.8185)	(0.6567, 0.7417)	(0.7633, 0.835)	(0.8265, 0.8917)
PEMI-Gauss	$p = \frac{1}{2}$	(0.4497, 0.5632)	(0.7517, 0.8313)	(0.8697, 0.9244)	(0.8385, 0.8935)	(0.8768, 0.9308)
	p calibrated	**(0.0923, 0.1981)**	**(0.0833, 0.1676)**	**(0.0731, 0.1491)**	**(0.0714, 0.152)**	**(0.1653, 0.2643)**
	$p = N_+/N$	(0.693, 0.773)	(0.7609, 0.8286)	(0.6405, 0.7249)	(0.7575, 0.8313)	(0.8335, 0.9001)

Table 13: KS-distance of PIT for different algorithms in MCAR setting (all datasets). 90% confidence intervals. $p_m = 0.3$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		(0.0954, 0.1231)	(0.0772, 0.1047)	**(0.0607, 0.0846)**	(0.0768, 0.1073)	(0.4246, 0.4546)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.1742, 0.2367)	(0.345, 0.375)	(0.2958, 0.3933)	(0.3675, 0.3875)	(0.3625, 0.3875)
	p calibrated	**(0.0433, 0.1117)**	**(0.0608, 0.105)**	**(0.0425, 0.0917)**	**(0.04, 0.1092)**	**(0.0625, 0.1158)**
	$p = N_+/N$	(0.3383, 0.3675)	(0.3492, 0.3767)	(0.3225, 0.3558)	(0.3383, 0.3725)	(0.3642, 0.3842)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.2279, 0.2938)	(0.4029, 0.4258)	(0.3958, 0.4446)	(0.3975, 0.425)	(0.4175, 0.4379)
	p calibrated	**(0.0383, 0.1013)**	**(0.0375, 0.0725)**	**(0.0288, 0.0667)**	**(0.025, 0.0642)**	**(0.0667, 0.1309)**
	$p = N_+/N$	(0.3917, 0.4192)	(0.3888, 0.4146)	(0.3579, 0.3946)	(0.3992, 0.4212)	(0.4125, 0.4362)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.2503, 0.3216)	(0.4345, 0.4618)	(0.4757, 0.4869)	(0.4567, 0.4751)	(0.4712, 0.486)
	p calibrated	**(0.0276, 0.0603)**	**(0.0286, 0.0759)**	**(0.0357, 0.0943)**	**(0.0229, 0.0512)**	**(0.0845, 0.1536)**
	$p = N_+/N$	(0.4333, 0.4584)	(0.4392, 0.4617)	(0.4045, 0.4372)	(0.4449, 0.4663)	(0.457, 0.478)
PEMI-Gauss	$p = \frac{1}{2}$	(0.2294, 0.3005)	(0.4397, 0.467)	(0.4749, 0.4882)	(0.4596, 0.479)	(0.4752, 0.4897)
	p calibrated	**(0.0354, 0.0942)**	**(0.031, 0.0771)**	**(0.0273, 0.0739)**	**(0.0242, 0.0519)**	**(0.0959, 0.1634)**
	$p = N_+/N$	(0.419, 0.4493)	(0.4413, 0.4666)	(0.3989, 0.4326)	(0.4442, 0.4678)	(0.4639, 0.4832)

Table 14: W_1 distance of PIT for different algorithms in MCAR setting (all datasets). 90% confidence intervals. $p_m = 0.3$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		**(0.0155, 0.0217)**	(0.0124, 0.0178)	(0.0102, 0.0137)	(0.0116, 0.0158)	(0.091, 0.1055)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.0409, 0.0492)	(0.0922, 0.1156)	(0.0694, 0.0793)	(0.0836, 0.1021)	(0.1068, 0.129)
	p calibrated	**(0.013, 0.0181)**	**(0.0085, 0.0119)**	**(0.0066, 0.0097)**	**(0.0088, 0.0128)**	**(0.0088, 0.0132)**
	$p = N_+/N$	(0.0572, 0.0699)	(0.0573, 0.0687)	(0.0334, 0.0388)	(0.0561, 0.068)	(0.0716, 0.0822)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.0399, 0.0484)	(0.091, 0.1144)	(0.07, 0.0798)	(0.0823, 0.1007)	(0.1066, 0.1292)
	p calibrated	**(0.0129, 0.0174)**	**(0.0091, 0.0131)**	**(0.0067, 0.0098)**	**(0.0078, 0.0114)**	**(0.0088, 0.0132)**
	$p = N_+/N$	(0.0571, 0.0699)	(0.0554, 0.067)	(0.0333, 0.0389)	(0.0555, 0.0673)	(0.0719, 0.0824)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.0419, 0.0504)	(0.0908, 0.1143)	(0.0701, 0.0796)	(0.083, 0.1015)	(0.1071, 0.1292)
	p calibrated	**(0.0127, 0.0169)**	**(0.0086, 0.0123)**	**(0.0063, 0.0092)**	**(0.0076, 0.0109)**	**(0.0081, 0.0122)**
	$p = N_+/N$	(0.0556, 0.0682)	(0.0547, 0.0662)	(0.0328, 0.0383)	(0.0563, 0.0682)	(0.0723, 0.0826)
PEMI-Gauss	$p = \frac{1}{2}$	(0.0414, 0.05)	(0.0912, 0.1146)	(0.07, 0.0795)	(0.0827, 0.1011)	(0.1068, 0.1292)
	p calibrated	**(0.0126, 0.0168)**	**(0.0088, 0.0124)**	**(0.0064, 0.0093)**	**(0.0075, 0.0109)**	**(0.0083, 0.0124)**
	$p = N_+/N$	(0.0561, 0.0689)	(0.0552, 0.0666)	(0.0329, 0.0384)	(0.056, 0.0679)	(0.0726, 0.0829)

Table 15: MAE of centre of evaluation metric predictive distributions (all datasets), MCAR setting. 90% confidence intervals. $p_m = 0.3$. Best results in **bold**.

		Precision	Recall	Accuracy	F_1	ROC-AUC
Bootstrap $M = 0$		**(0.0266, 0.0376)**	(0.0214, 0.0319)	**(0.0159, 0.0209)**	**(0.0189, 0.0266)**	(0.1075, 0.1226)
PEMI $B = 5$	$p = \frac{1}{2}$	(0.0536, 0.0624)	(0.1344, 0.1559)	(0.0805, 0.0897)	(0.113, 0.1304)	(0.1391, 0.1627)
	p calibrated	**(0.0227, 0.0325)**	**(0.0143, 0.02)**	**(0.0126, 0.0178)**	**(0.0165, 0.0239)**	**(0.0166, 0.025)**
	$p = N_+/N$	(0.0766, 0.0907)	(0.0749, 0.0853)	(0.0406, 0.0457)	(0.0742, 0.0858)	(0.0844, 0.0941)
PEMI $B = 10$	$p = \frac{1}{2}$	(0.0529, 0.0613)	(0.1331, 0.1548)	(0.0809, 0.0902)	(0.1118, 0.1294)	(0.1396, 0.163)
	p calibrated	**(0.0207, 0.0296)**	**(0.016, 0.0232)**	**(0.0127, 0.0174)**	**(0.0149, 0.0209)**	**(0.0175, 0.0255)**
	$p = N_+/N$	(0.0766, 0.0912)	(0.0735, 0.0841)	(0.0408, 0.0464)	(0.0736, 0.0849)	(0.0844, 0.0937)
PEMI $B = 100$	$p = \frac{1}{2}$	(0.0546, 0.0636)	(0.1333, 0.1549)	(0.0806, 0.0896)	(0.1122, 0.1299)	(0.1395, 0.1629)
	p calibrated	**(0.02, 0.0279)**	**(0.015, 0.0218)**	**(0.0117, 0.0167)**	**(0.0138, 0.019)**	**(0.0161, 0.0236)**
	$p = N_+/N$	(0.0747, 0.0895)	(0.0725, 0.0832)	(0.0402, 0.0454)	(0.0744, 0.0862)	(0.0846, 0.0937)
PEMI-Gauss	$p = \frac{1}{2}$	(0.0542, 0.0629)	(0.1335, 0.1551)	(0.0804, 0.0894)	(0.112, 0.1296)	(0.1395, 0.1629)
	p calibrated	**(0.0198, 0.0276)**	**(0.0152, 0.0215)**	**(0.0119, 0.0167)**	**(0.0139, 0.0192)**	**(0.0163, 0.024)**
	$p = N_+/N$	(0.0753, 0.0898)	(0.0729, 0.0835)	(0.0403, 0.0453)	(0.0742, 0.0859)	(0.0848, 0.0942)

Table 16: RMSE of centre of evaluation metric predictive distributions (all datasets), MCAR setting. 90% confidence intervals. $p_m = 0.3$. Best results in **bold**.

C.2 MNAR Regime

C.2.1 Missingness Fraction $p_m = 0.1$

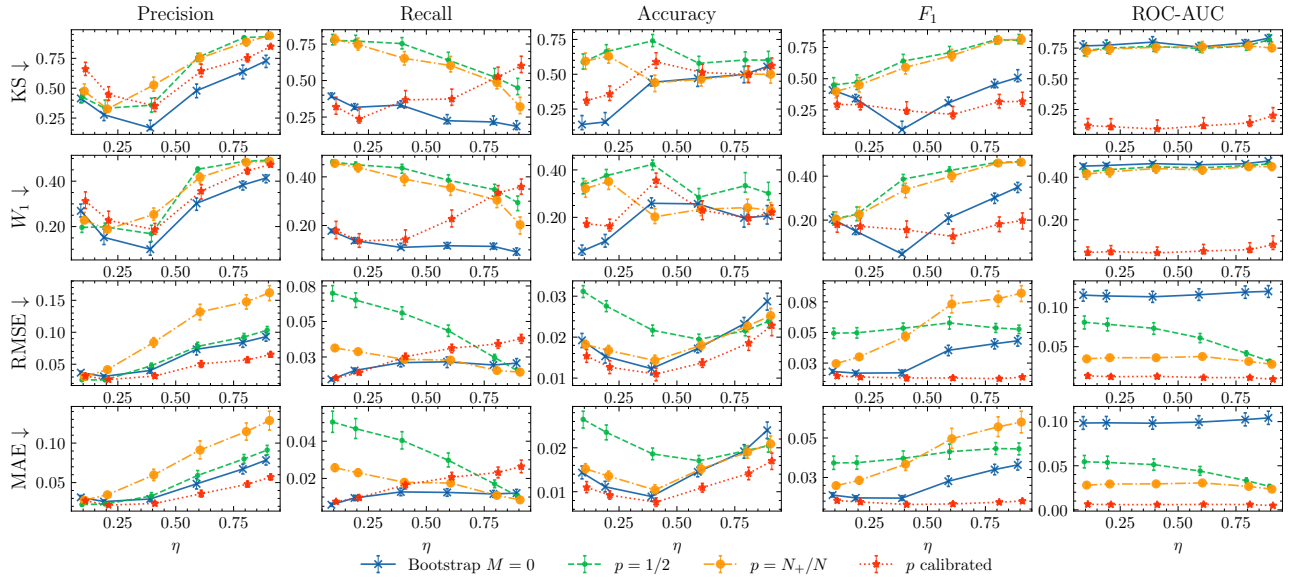

Figure 5: Effect of varying MNAR class imbalance $\eta \in (0, 1)$ on evaluation metric predictive distribution quality using PEMI-Gauss for $p_m = 0.1$. Each plot represents a metric $\widehat{Q}_n^{(S)}$. Error bars are bootstrapped confidence intervals at the $\alpha = 0.9$ level. Quality is on y-axis with η varying on the x-axis, with x-axis values slightly jittered to help separate each series visually. Each line corresponds to a predictive distribution, with p algorithms using PEMI-Gauss.

C.2.2 Missingness Fraction $p_m = 0.2$

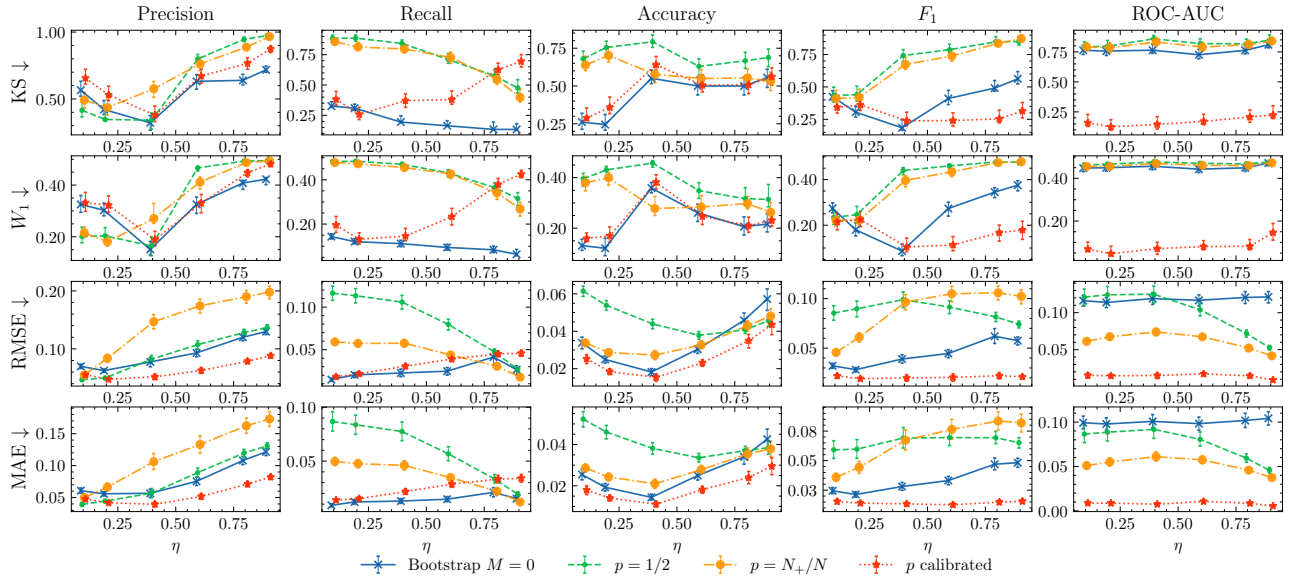

Figure 6: Effect of varying MNAR class imbalance $\eta \in (0, 1)$ on evaluation metric predictive distribution quality using PEMI-Gauss for $p_m = 0.2$. Each plot represents a metric $\widehat{Q}_n^{(S)}$. Error bars are bootstrapped confidence intervals at the $\alpha = 0.9$ level. Quality is on y-axis with η varying on the x-axis, with x-axis values slightly jittered to help separate each series visually. Each line corresponds to a predictive distribution, with p algorithms using PEMI-Gauss.

C.2.3 Missingness Fraction $p_m = 0.3$

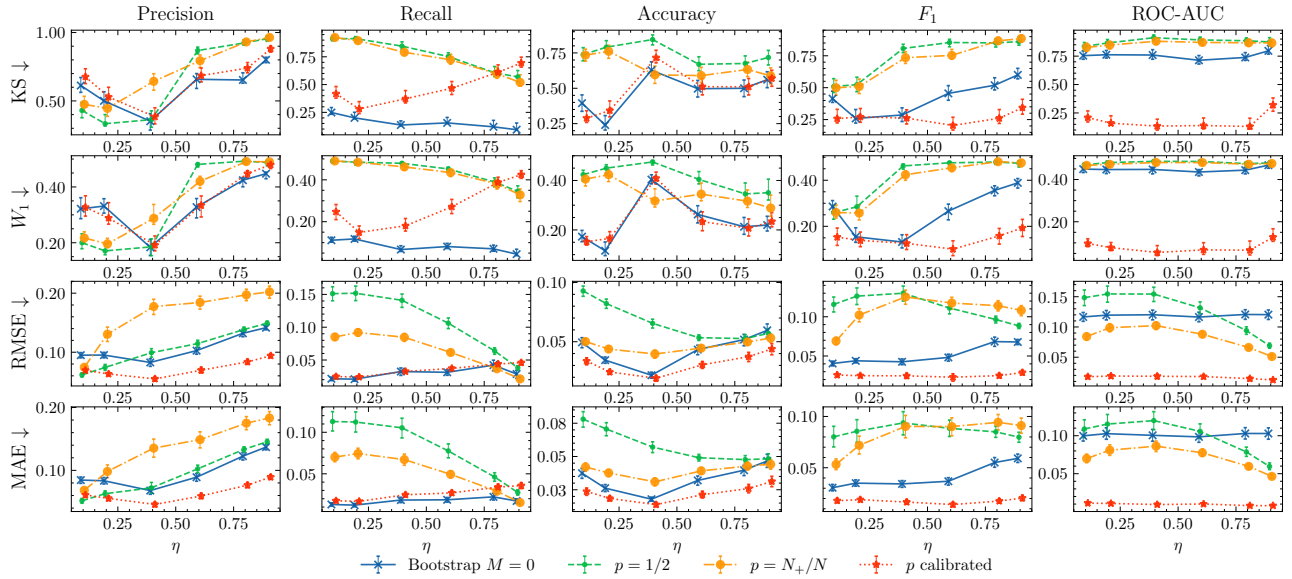

Figure 7: Effect of varying MNAR class imbalance $\eta \in (0, 1)$ on evaluation metric predictive distribution quality using PEMI-Gauss for $p_m = 0.3$. Each plot represents a metric $\widehat{Q}_n^{(S)}$. Error bars are bootstrapped confidence intervals at the $\alpha = 0.9$ level. Quality is on y-axis with η varying on the x-axis, with x-axis values slightly jittered to help separate each series visually. Each line corresponds to a predictive distribution, with p algorithms using PEMI-Gauss.