Confidence Intervals for Error Rates in 1:1 Matching Tasks: Critical Statistical Analysis and Recommendations

Riccardo Fogliato[1*], Pratik Patil[2*] and Pietro Perona[1]

[1*]AWS AI, USA.
[2*]University of California, Berkeley, USA.

*Corresponding author(s). E-mail(s): fogliato@amazon.com; pratikpatil@berkeley.edu;
Contributing authors: peronapp@amazon.com;

Abstract

Matching algorithms predict relationships between items in a collection. For example, in 1:1 face verification, a matching algorithm predicts whether two face images depict the same person. Accurately assessing the uncertainty of the error rates of such algorithms can be challenging when test data are dependent and error rates are low, two aspects that have been often overlooked in the literature. In this work, we review methods for constructing confidence intervals for error rates in 1:1 matching tasks. We derive and examine the statistical properties of these methods, demonstrating how coverage and interval width vary with sample size, error rates, and degree of data dependence with experiments on synthetic and real-world datasets. Based on our findings, we provide recommendations for best practices for constructing confidence intervals for error rates in 1:1 matching tasks. Our code is available at github.com/awslabs/cis-matching-tasks.

Keywords: matching tasks, confidence intervals, false match/non-match rate, false acceptance/rejection rate

Contents

1 Introduction

Accurately measuring system accuracy is essential for responsible design and deployment of automated systems [26]. Accurate measurements aid in identifying suitable use cases for a system, guiding engineers towards enhancements, and helping stakeholders comprehend the system's strengths and limitations. Nevertheless, the value of accuracy measurements is limited without considering their statistical uncertainty. While the methodology for computing confidence intervals for classification tasks on independent data in well-established [8], it remains problematic for *matching tasks*.

To construct confidence intervals for the accuracy of algorithms used in 1:1 matching tasks, using standard Wald intervals based on the Gaussian approximation of the maximum likelihood estimator may appear to be a viable approach [11, 49]. However, this approach is problematic for the following two reasons:

(M1) *Low error rates.* When the 1:1 matching algorithm is highly accurate, as is the case, for instance, in face recognition (FR) systems [23], error rates are close to zero, which makes the Gaussian approximation inaccurate. Consequently, confidence intervals based on this approximation may significantly under-cover the true error rates.

(M2) *Sample dependence.* When test sets are relatively small the pair-wise samples used in matching tasks may include the same item multiple times, e.g. the same face photograph may be used in multiple comparisons. This means that the samples are correlated. Therefore, using Wald intervals with variance estimated under the independence assumption is not suitable for this scenario.

Our study focuses precisely on these issues. It is worth mentioning that we are not the first to consider low error rates and sample dependence. Bootstrap procedures have been proposed in the FR literature to address sample dependence and have been widely used in empirical studies [7, 36, 39, 51]. However, the development of these methods was based on heuristic arguments, and there has been limited discussion regarding their statistical guarantees, such as their frequentist coverage. For instance, it is well known that

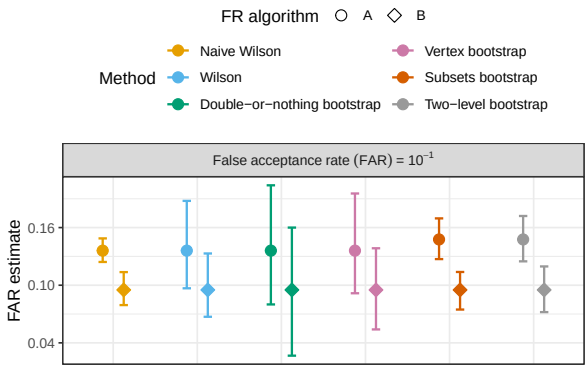

Fig. 1 Different methods for constructing confidence intervals can lead to different conclusions due to miscoverage. Six methods for computing estimates and corresponding 95% confidence intervals on synthetic data for the false accept rate (FAR) of two 1:1 matching algorithms (A and B) that have underlying equal accuracy ($\texttt{FAR} = 10^{-1}$). The data contains 50 groups, with 5 images each, and all pairwise comparisons are considered in the estimation of the error metric (details in Section 6). Dots and bars correspond to error estimates and corresponding confidence intervals. The naive Wilson, subsets bootstrap, and two-level bootstrap intervals may lead the practitioner to *erroneously* conclude that Algorithm A has inferior performance compared to Algorithm B – while in our simulation they are equivalent. In our analysis and experiments we find that only Wilson intervals achieve nominal coverage in the presence of low error rates (M1) and sample dependence (M2). Double-or-nothing and vertex bootstrap intervals also work well in settings characterized only by (M2).

confidence intervals based on bootstrap resampling can fail to achieve nominal coverage in various settings, meaning that the probability of the true parameter being contained in the intervals is lower than the desired rate. This issue is prominent when accuracy/error metrics are close to the parameter boundary (e.g., false acceptance rate is close to 0) or when sample sizes are small. These issues are particularly pertinent in intersectional analyses within bias assessments [3], where the number of images available for certain demographic subgroups is often limited.

Different methods will yield confidence intervals with different widths. Figure 1 shows such an example. Which is right? Without a thorough understanding of the statistical properties of the various methods, it is difficult to determine which method is most appropriate for a particular setting. It is unclear whether and when the constructed interval achieves the desired nominal coverage. In light of these considerations, our

investigation is guided by the following fundamental question: *Which methods should be used to construct confidence intervals for error metrics in 1:1 matching tasks?* We investigate methods with the primary aim of addressing the issues mentioned in (M1) and (M2).

Besides exploring analytically the properties of different methods, we carry out a thorough experimental investigation as well. We use both synthetic data and data coming from real-life applications. Amongst many options, we chose face verification, an important and sensitive application of Computer Vision [23, 37, 38, 48]. Thus, we present a critical examination and analysis of methodologies for constructing confidence intervals for error rates in 1:1 matching tasks, and use face verification as a representative test application. Our findings are applicable across all 1:1 matching tasks, encompassing 1:1 speaker, fingerprint, and iris recognition, among others.

Our theoretical analysis and empirical investigation reveal that, although there is no "one-size-fits-all" solution, only certain methods consistently achieve coverage that is close to nominal. Some concrete examples are illustrated in Figure 2. From the figure, we observe that in the case of the `FRR` (false rejection rate), all nonparametric bootstrap methods significantly under-cover when `FRR` is close to the parameter boundary, while parametric Wilson intervals that assume data independence under-cover for large values of `FRR`. In the case of the `FAR` (false acceptance rate), the subsets and two-level bootstrap techniques fail to achieve nominal coverage at any level of the error metric, while the *naive Wilson* interval, where one neglects to account for data dependence, shrinks with growing `FAR`. The remaining three methods are more promising: Wilson interval that accounts for data dependence (which we will refer to simply as *Wilson* hereafter) always achieves nominal coverage, while the vertex and double-or-nothing bootstraps cover at the right level when the true error metrics are large. See Section 4 for a description of the aforementioned methods.

Summary of contributions

Our main contributions are as follows:

1. *Methods review.* We provide a review of two classes of methods for constructing confidence intervals for matching tasks, one based on parametric assumptions, and the other on nonparametric, resampling-based methods. The reviewed methods include the Wilson intervals without (naive version) and with variance adjusted for data dependence, subsets, two-level, vertex, and double-or-nothing bootstraps.
2. *Theoretical analysis.* We present a theoretical analysis of the reviewed methods with a focus on intervals for error rates that are computed at a fixed threshold. Our analysis includes statistical guarantees for coverage of the intervals and their width.
3. *Empirical evaluation.* To compare the properties of confidence intervals for error rates at a fixed thresholds as well as of pointwise intervals for the ROC generated by the reviewed methods, we conduct experiments on both synthetic and real-world datasets, namely on the MORPH dataset [42].
4. *Software library.* Our code and python package `cimat` are available at github.com/awslabs/cis-matching-tasks.
5. *A recommendation.* Based on these findings, we recommend (R1) using Wilson intervals with variance adjusted for data dependence to address (M1) and (M2), or utilizing vertex and double-or-nothing bootstrap intervals in settings that exhibit only (M2).

Paper outline

In Section 2, we provide an overview of related work on confidence interval for clustered data. In Section 3, we describe the problem setup. We focus on the balanced setting, where each individual present in the data has an equal number of images. In Section 4, we describe the statistical properties of the methods in the balanced setting. In Section 5, we present extensions of these methods, including estimation in the unbalanced setting (where the number of instances can vary across individuals), strategies for constructing pointwise confidence intervals for the receiver operating characteristics (ROC) curve. In Section 6, we provide numerical evaluation of different methods. In Section 7, we discuss merits and

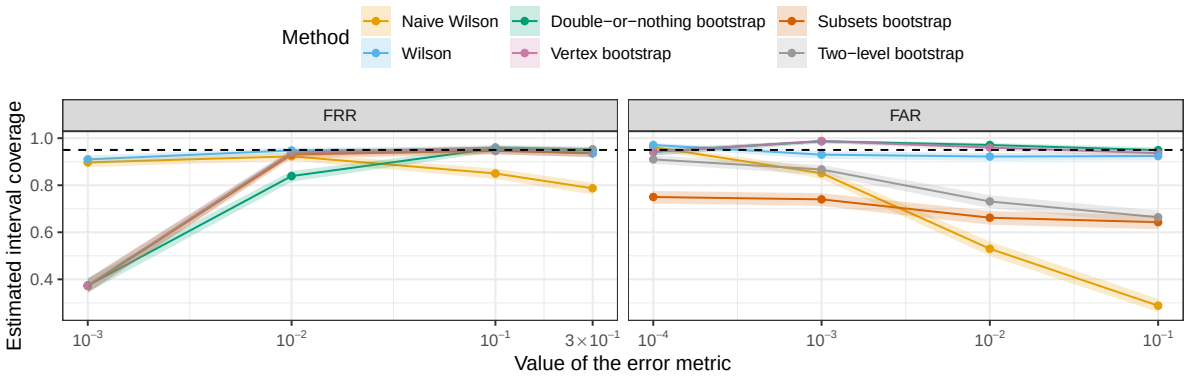

Fig. 2 Estimated interval coverage of 95% confidence intervals for FRR and FAR on synthetic data in settings characterized by (M1) and (M2). The data contains 50 identities and 5 instances (e.g., images) per identity. Lines and shaded regions indicate estimated coverage and corresponding 95% confidence intervals respectively, for each method, across data replications. The dashed line indicates nominal coverage (95%). The experimental setup is described in Section 6. Only Wilson's method (blue lines) guarantees accurate coverage across all experimental conditions.

pitfalls of the different methods, as well as directions for future work. The Appendix to the paper contains proofs of the theoretical results stated in the main paper, additional numerical experiments, and other miscellaneous details.

2 Related work

Our primary objective is to construct confidence intervals that achieve nominal coverage robustly for parameters close to the boundary of the parameter space (M1) by handling dependent samples (M2). The issue raised in (M1) has been extensively studied by statisticians, while the issue mentioned in (M2) arises in the analysis of data such as time series, networks, surveys, dyadic, and panel data, among others. Consequently, confidence interval construction methods that handle sample dependence have been developed in Economics, Statistics, and the social sciences. In this section, we briefly review existing parametric and nonparametric methods proposed in these fields, as well as those introduced by the FR community.

Parametric methods

The most commonly used parametric confidence intervals under data independence are Wald intervals, which rely on the asymptotic normality of the maximum likelihood estimator. However, this assumption may not hold in finite samples and thus this type of interval may fail to achieve nominal coverage [8]. One prominent example is the one of the binomial proportion (e.g., error rates of classification tasks under data independence) being equal to the sample size, for which Wald-type intervals as well as bootstrap-based intervals are degenerate. For this reason, in this setting the Wilson [50], Agresti-Coull [1], and Jeffreys intervals are preferred. When the independence assumption is violated, one can account for the dependency structure in the data in the estimation of the Gaussian variance [32]. Variance estimation methods have been explored in the context of dyadic data [9, 18, 45]. These approaches are discussed in Section 4.

Nonparametric resampling methods in FR

Nonparametric resampling methods offer an alternative approach to constructing confidence intervals that does not rely on the asymptotic normality assumption of the target statistic. In the context of 1:1 face verification, Bolle et al. [7] propose the subsets bootstrap, which consists of resampling at the level of the identities: If an identity is sampled, every comparison in the data that involves that identity is included in the bootstrap sample. However, as we demonstrate in Section 4, the dependence structure between the bootstrapped and original datasets may differ significantly, resulting in under-coverage of FAR intervals. In an attempt to address this issue, Poh et al. [39] propose a two-level bootstrap where resampling occurs at both the identity and

individual score levels. In the case of FRR intervals, these two techniques are standard choices of block bootstraps, which have been discussed in the statistical literature [15, 20] and are widely used in practice. However, we found that none of the articles in the FR literature we reviewed discuss the statistical properties of these procedures, except for the recent work by Conti and Clémençon [13]. They propose resampling individual images and deriving confidence intervals from the resulting bootstrap distribution, which needs to be recentered around the metric value on the original dataset. They argue that, asymptotically, this distribution converges to the law of the target statistic.

Nonparametric resampling methods in Statistics and Economics

There are a number of nonparametric bootstrap and subsampling techniques available for conducting inference on dyadic data [4, 5, 10, 14, 22, 30, 31, 45]. See Graham [21] for a comprehensive review. The majority of these approaches are intended to offer asymptotic guarantees, where the distribution of the conditional data mean follows a Gaussian distribution under specific circumstances. As a result, these methods mainly focus on the degree to which the bootstrap distribution approximates the first two moments of the underlying distribution of the target statistic. One method that has been widely employed in the social sciences is the vertex bootstrap proposed by Snijders et al. [45]. The procedure proposed by Conti and Clémençon [13] is similar in spirit, the main difference being that only the former swaps comparisons between the same image with a random sample taken from the set of comparisons present in the original data. In our work, we investigate both the vertex bootstrap and a related method, the double-or-nothing bootstrap, which has been studied in the context of exchangeable arrays, such as dyadic data [14, 35]. Both of these methods have the desirable property that, asymptotically, the first two moments of the bootstrap distribution match those of the distribution of the error metric estimators that we consider in this work. This is not the case for the subsets and two-level bootstraps.

Parametric bootstrap methods

An alternative to nonparametric resampling methods comes in the form of parametric bootstrapping. For example, Mitra et al. [33] fit a generalized linear mixed model and obtain credible intervals by sampling from the model's posterior predictive distribution. However, while mixed models are capable of handling network data dependency and have been widely studied [24, 25], fitting these models on large datasets, such as those found in face recognition applications, can be challenging. For this reason, we exclude this type of inference from our analysis. An alternative Bayesian approach is to model the distributions of scores. This is done, e.g., by Chouldechova et al. [12], although their focus is on semi- and unsupervised estimation.

3 Problem setup

In this section, we describe the problem setup and introduce our notation. We consider a set of G different identities (we use the word "identity" commonly used in FR, where it means "a specific person") denoted by $\mathcal{G}$. Each identity $i \in \mathcal{G}$ has M_i instances (e.g., face images) that are represented by embeddings $X_{(i,1)}, \ldots, X_{(i,M_i)} \in \mathbb{R}^d$. For example, these embeddings are generated by a FR model and may be normalized. We assume that the embeddings follow a common probability law Q on $\mathbb{R}^d$ for all identities $i \in \mathcal{G}$ and all $1 \leqslant k \leqslant M_i$. Furthermore, if we consider a pair of instances k and l belonging to identities i and j from $\mathcal{G}$, we assume that the embeddings $X_{(i,k)}$ and $X_{(j,l)}$ are independent when $i \neq j$.

We will focus on binary classification tasks, where the goal is to classify a pair of instances as belonging to the same identity (i.e., "genuine") or different identities (i.e., "impostor"). This classification is done based on the distance (e.g., Euclidean distance) between the embeddings and a threshold $t \in \mathbb{R}$. Specifically, the pair of instances is classified as genuine when when $d(X_{(i,k)}, X_{(j,l)}) < t$, and as impostor when $d(X_{(i,k)}, X_{(j,l)}) \geqslant t$ for some distance function d. For an identity i, when $k \neq l$, let $Y_{(i,k),(i,l)} = \mathbb{1}\{d(X_{(i,k)}, X_{(i,l)}) \geqslant t\} \sim \text{Bernoulli}(\texttt{FRR})$, where FRR is termed False Non-Match Rate (FNMR) or False Rejection Rate (FRR). For different identities $i \neq j$, let $Y_{(i,k),(j,l)} = \mathbb{1}\{d(X_{(i,k)}, X_{(j,l)}) < t\} \sim$

Bernoulli(FAR), where FAR is termed False Match Rate (FMR) or False Acceptance Rate (FAR). We are interested in estimating the parameters FRR and FAR from the sample.[1][2]

Apart from estimating these parameters, we wish to construct confidence intervals for them. There are two properties of the intervals one generally cares about: one is coverage and the other is length. Our primary focus is to construct intervals with valid nominal coverage. Formally, given a nominal coverage level of $1-\alpha$ for some $\alpha \in (0,1)$, our goal is to construct confidence (rather than credible) intervals, denoted by $\mathcal{I}_{\texttt{FRR}}$ and $\mathcal{I}_{\texttt{FAR}}$, respectively, for the two metrics FRR and FAR that satisfy the following frequentist coverage guarantees: $\mathbb{P}_Q(\texttt{FRR} \in \mathcal{I}_{\texttt{FRR}}) \geqslant 1-\alpha$ and $\mathbb{P}_Q(\texttt{FAR} \in \mathcal{I}_{\texttt{FAR}}) \geqslant 1-\alpha$. Among intervals with the correct coverage, shorter intervals are preferable. In practice, however, it may be difficult to achieve the exact coverage guarantee, and thus, one might have to settle for an approximate guarantee.

We next consider estimators for FRR and FAR in the form of empirical averages for the balanced setting, where $M_i = M$ for all $i \in \mathcal{G}$. These point estimators lead to the confidence intervals that we describe in Section 4. The estimation in the unbalanced setting, where the number of instances M_i can vary across identities, is described in Section 5.1.

Balanced setting

Consider a sample where each of the G identities available has M instances. One can define natural empirical estimators of FRR and FAR for identities $i, j \in \mathcal{G}$ as follows:

$$\overline{Y}_{ij} = \begin{cases} \displaystyle\sum_{k=1}^{M} \sum_{\substack{l=1,\\ l\neq k}}^{M} \frac{Y_{(i,k),(i,l)}}{M(M-1)} & \text{when } i = j, \\ \displaystyle\sum_{k,l=1}^{M} \frac{Y_{(i,k),(j,l)}}{M^2} & \text{when } i \neq j. \end{cases} \tag{1}$$

Here, the estimator $\overline{Y}_{ij}$ measures the error metrics at the level of each identity. Estimators of FRR and FAR for the entire sample are then given by

$$\widehat{\texttt{FRR}} = \sum_{i=1}^{G} \frac{\overline{Y}_{ii}}{G}, \ \widehat{\texttt{FAR}} = \sum_{i=1}^{G} \sum_{\substack{j=1,\\ j\neq i}}^{G} \frac{\overline{Y}_{ij}}{G(G-1)}, \tag{2}$$

respectively. The type of confidence intervals for FRR and FAR that we consider are based on these estimators. It is easy to see that $\widehat{\texttt{FRR}}$ and $\widehat{\texttt{FAR}}$ are unbiased estimators of FRR and FAR respectively, that is $\text{Bias}(\widehat{\texttt{FRR}}) = \mathbb{E}[\widehat{\texttt{FRR}}] - \texttt{FRR} = 0$ and $\text{Bias}(\widehat{\texttt{FAR}}) = \mathbb{E}[\widehat{\texttt{FAR}}] - \texttt{FAR} = 0$. In addition, we have:

$$\text{Var}(\widehat{\texttt{FRR}}) = \frac{1}{G}\text{Var}(\overline{Y}_{11}), \tag{3}$$

$$\text{Var}(\widehat{\texttt{FAR}}) = \frac{2}{G(G-1)}\text{Var}(\overline{Y}_{12}) + \frac{4(G-2)}{G(G-1)}\text{Cov}(\overline{Y}_{12}, \overline{Y}_{13}). \tag{4}$$

The variances will be of key interest throughout our discussion of the validity of the confidence intervals. While $\text{Var}(\widehat{\texttt{FRR}})$ corresponds to the variance of $\widehat{\texttt{FRR}}$ across identities, we observe that $\text{Var}(\widehat{\texttt{FAR}})$ will coincide with variance under data independence only when $\text{Cov}(\overline{Y}_{12}, \overline{Y}_{13}) = 0$. Thus, in general, the covariance terms will need to be accounted for in the construction of the confidence intervals.

Our asymptotic analysis in Section 4 will focus on the setting where G grows while M remains fixed. This is motivated by the observation that in FR applications the number of unseen identities is generally larger than the number of face images per identity. This kind of asymptotic analysis is also typical in prior studies on inference using clustered data [10, 20].

4 Methods description

In this section, we describe parametric (Section 4.1) and nonparametric resampling-based methods (Section 4.2) for constructing confidence intervals for error rates in matching tasks with binary model predictions. Our focus will be on the balanced setting. Methods extensions, including confidence intervals in the unbalanced setting, pointwise intervals for the receiver operating characteristic (ROC) curve, as well as protocol design strategies, can be found

[1] The framework can be adapted to include conditioning on predefined attributes of identities, such as when it's already known which demographic groups certain identities belong to.

[2] The framework can also apply to other losses such as cross-entropy. The principles and methods reviewed, including the bootstrap techniques, can be adapted or directly employed.

in Section 5. We will defer all the proofs of the theoretical statements to Appendix A.

4.1 Parametric methods

Parametric methods for constructing confidence intervals rely on assumptions made about the distribution of the target statistic. In Section 2, we have mentioned that Wald intervals are typically used for constructing intervals for statistics that are asymptotically normal under data independence, while other methods such as Wilson, Agresti-Coull, and Jeffreys have been explored specifically for binomial proportions. To derive intervals that have good coverage in the presence of data dependence, we need to characterize the asymptotic behavior of $\sqrt{G}(\widehat{\texttt{FRR}} - \texttt{FRR})$ and $\sqrt{G}(\widehat{\texttt{FAR}} - \texttt{FAR})$. The following proposition establishes a set of conditions under which these statistics are asymptotically normal.

Proposition 1 (Normality of scaled error rates)**.** *Assume that* $\lim_{G\to\infty} \mathrm{Var}(\sqrt{G}\widehat{\texttt{FRR}}) = c_{\texttt{FRR}}$ *and* $\lim_{G\to\infty} \mathrm{Var}(\sqrt{G}\widehat{\texttt{FAR}}) = c_{\texttt{FAR}}$ *for some positive constants* $c_{\texttt{FRR}}, c_{\texttt{FAR}}$*. Then, as* $G \to \infty$, $\sqrt{G}(\widehat{\texttt{FRR}} - \texttt{FRR}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, \mathrm{Var}(\overline{Y}_{11}))$ *and* $\sqrt{G}(\widehat{\texttt{FAR}} - \texttt{FAR}) \xrightarrow{\mathrm{d}} \mathcal{N}(0, 4\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13}))$.

Since identity-level observations are assumed to be independent, the convergence in distribution of $\sqrt{G}\texttt{FRR}$ follows from an application of the central limit theorem. The case of the FAR follows from Proposition 3.2 in Tabord-Meehan [46]. The result in the proposition motivates the construction of confidence intervals based on the limiting distribution.

Construction of confidence intervals

The use of confidence intervals for binomial proportions in the presence of dependent data was first proposed by Miao and Gastwirth [32]. For instance, Wald intervals in this setting take the form $\mathcal{I}_{\texttt{FRR}} = [\widehat{\texttt{FRR}} \pm z_{1-\alpha/2}\sqrt{\mathrm{Var}(\overline{Y}_{11})/G}]$ and $\mathcal{I}_{\texttt{FAR}} = [\widehat{\texttt{FAR}} \pm z_{1-\alpha/2}\sqrt{\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})/G}]$, where $z_{1-\alpha/2}$ corresponds to the $(1-\alpha/2)$-th quantile of the standard normal. From Proposition 1, it then follows that the intervals have the correct asymptotic coverage. In practice, Wilson intervals are preferred as they achieve good coverage even in the presence of small sample sizes [8]. The $1-\alpha$ *Wilson* confidence interval for FAR, which assumes data dependence, is given by

$$\mathcal{I}_{\texttt{FAR}} = \left[\frac{\widehat{\texttt{FAR}}\widehat{N}^*_{\texttt{FAR}} + \frac{1}{2}z^2_{1-\alpha/2}}{\widehat{N}^*_{\texttt{FAR}} + z^2_{1-\alpha/2}} \pm \frac{z_{1-\alpha/2}\sqrt{\widehat{N}^*_{\texttt{FAR}}}}{\widehat{N}^*_{\texttt{FAR}} + z^2_{1-\alpha/2}} \cdot \sqrt{\widehat{\texttt{FAR}}(1-\widehat{\texttt{FAR}}) + z^2_{1-\alpha/2}/(4\widehat{N}^*_{\texttt{FAR}})} \right], \quad (5)$$

where $\widehat{N}^*_{\texttt{FAR}} = \max\{\widehat{\texttt{FAR}}(1-\widehat{\texttt{FAR}})/\mathrm{Var}(\widehat{\texttt{FAR}}), \lfloor G/2 \rfloor\}$. The *naive Wilson* confidence interval, which assumes data independence, for FAR uses $\widehat{N}^*_{\texttt{FAR}} = G(G-1)M^2/2$. The Wilson interval for FRR is obtained by replacing $\widehat{\texttt{FAR}}$ with $\widehat{\texttt{FRR}}$ and $N^*_{\texttt{FAR}}$ with $N^*_{\texttt{FRR}} = \max\{\widehat{\texttt{FRR}}(1-\widehat{\texttt{FRR}})/\mathrm{Var}(\widehat{\texttt{FRR}}), G\}$, while its naive version employs $\widehat{N}^*_{\texttt{FRR}} = GM(M-1)/2$. If Proposition 1 holds, then the Wilson intervals will have the nominal coverage. As a side remark, it is worth mentioning that the 95% Wilson interval (5) bears resemblance to a Wald interval that is calculated on a dataset with two successes and two failures appended. For more details, see Agresti and Coull [1].

It should be noted that the construction of these intervals relies on having knowledge of $\mathrm{Var}(\widehat{\texttt{FRR}})$ and $\mathrm{Var}(\widehat{\texttt{FAR}})$. However, thanks to Slutsky's theorem, by replacing these variances with their consistent estimators, it is possible to use a modified version of Proposition 1 and certify the coverage of the resulting intervals. In the following discussion, we will focus on constructing consistent estimators of $\mathrm{Var}(\overline{Y}_{11})$ and $\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$.

Estimation of $\mathrm{Var}(\widehat{\texttt{FRR}})$ *and* $\mathrm{Var}(\widehat{\texttt{FAR}})$

The variances in Proposition 1 can be estimated using the following plug-in estimators:

$$\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FRR}}) = \frac{1}{G}\sum_{i=1}^{G}(\overline{Y}_{11} - \widehat{\texttt{FRR}})^2, \quad (6)$$

$$\widehat{\mathrm{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}) = \frac{1}{G(G-1)(G-2)} \cdot \sum_{i=1}^{G}\sum_{\substack{j=1,\\ j\neq i}}^{G}\sum_{\substack{k=1\\ k\neq j,i}}^{G}(\overline{Y}_{ij} - \widehat{\texttt{FAR}})(\overline{Y}_{ik} - \widehat{\texttt{FAR}}). \quad (7)$$

The estimator in (6) is the standard variance estimator under data independence. The estimator in (7) is employed for $\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$. However, in

finite samples, when $\mathrm{Var}(\overline{Y}_{12}) \gg \mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$, the individual variance terms in (4) may dominate. In that case, we may want to employ the following estimator of $\mathrm{Var}(\overline{Y}_{12})$:

$$\widehat{\mathrm{Var}}(\overline{Y}_{12}) = \frac{1}{G(G-1)} \sum_{i=1}^{G} \sum_{\substack{j=1 \\ j \neq i}}^{G} (\overline{Y}_{ij} - \widehat{\texttt{FAR}})^2, \quad (8)$$

and then plug in the estimators above into the variance expression (4). That is, we use:

$$\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FAR}}) = \frac{2}{G-1}\widehat{\mathrm{Var}}(\overline{Y}_{12}) + \frac{4(G-2)}{G-1}\widehat{\mathrm{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}). \quad (9)$$

This estimator is a special case of the robust variance estimator proposed by Fafchamps and Gubert [18] in the context of dyadic regression. The following proposition states the convergence in probability of these estimators to the target parameters.

Proposition 2 (Consistency of plug-in variance estimators)**.** *Consider the variance estimators* $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FRR}})$ *and* $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FAR}})$ *defined in* (6) *and* (9)*, respectively. Then, as* $G \to \infty$, $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FRR}}) \xrightarrow{\mathrm{p}} \mathrm{Var}(\overline{Y}_{11})$*, and* $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FAR}}) \xrightarrow{\mathrm{p}} 4\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$.

An alternative way to estimate $\mathrm{Var}(\sqrt{G}\widehat{\texttt{FAR}})$ is by using the following jackknife estimator:

$$\widehat{\mathrm{Var}}_{JK}(\sqrt{G}\widehat{\texttt{FAR}}) = \frac{(G-2)^2}{G} \cdot \sum_{i=1}^{G} (\widehat{\texttt{FAR}}_{-i} - \widehat{\texttt{FAR}})^2 - 2\frac{\widehat{\mathrm{Var}}(\overline{Y}_{12})}{G-1}. \quad (10)$$

Here, we have defined $\widehat{\texttt{FAR}}_{-i} = (G-1)^{-1}(G-2)^{-1} \sum_{j=1}^{G} \sum_{k=1, k\neq j}^{G} \overline{Y}_{jk} \mathbb{1}(\{j \neq i\} \cap \{k \neq i\})$. It turns out that the plug-in and jackknife estimators produce exactly the estimates. This is formalized in the following proposition.

Proposition 3 (Equivalence of plug-in and jackknife variance estimators)**.** *Consider the estimators* $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FAR}})$ *and* $\widehat{\mathrm{Var}}_{JK}(\sqrt{G}\widehat{\texttt{FAR}})$ *defined in equations* (9) *and* (10)*, respectively. It holds that* $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\texttt{FAR}}) = \widehat{\mathrm{Var}}_{JK}(\sqrt{G}\widehat{\texttt{FAR}})$.

The equivalence between the two estimators follows from the work of Graham [21]. The implementations of the two methods have similar computational costs, as they both involve $O(G^3)$ operations. Moreover, the estimators can be rewritten by using a multiway clustering decomposition, as outlined in Proposition 2 of Aronow et al. [2]. The implementation of this method is available in existing software packages in both R [53] and Python [43].

4.2 Resampling-based methods

We now consider an alternate and popular class of methods, confidence intervals constructed by bootstrap resampling. Bootstrap confidence intervals employ the so-called bootstrap distribution of the statistic of interest, which is obtained by resampling with replacement from the original data, the statistic of interest. In the percentile bootstrap method, the interval is based on the percentiles of this distribution [17]. For instance, let $\{\widehat{\texttt{FAR}}_b^*\}_{b=1}^B$ be the bootstrap distribution and let $\widehat{\texttt{FAR}}_b^*$ be the FAR estimated on the b-th bootstrap sample (i.e., resampled dataset). A $1-\alpha$ confidence interval for FAR is given by $[\widehat{\texttt{FAR}}^*_{(\lfloor B(\alpha/2) \rfloor)}, \widehat{\texttt{FAR}}^*_{(\lceil B(1-\alpha/2) \rceil)}]$. Below we discuss nonparametric resampling techniques that can be used to obtain the bootstrap distribution. The asymptotic coverage properties of the intervals constructed via the bootstrap depend on the mean and variance of the bootstrap distribution, hence we focus on these properties. In our subsequent discussion, we denote with $\mathbb{E}^*$ and Var^* the expectation and variance conditional on the original sample. Table 1 summarizes the statistical properties of the bootstraps that will be reviewed in the current section.

Table 1 Overview of asymptotic bias of the variance of bootstrapped error rates for the methods reviewed. All bootstrapped estimators have unbiased first moments.

Bootstrap	$\mathrm{Bias}(\mathrm{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}_b^*))$	$\mathrm{Bias}(\mathrm{Var}^*(\sqrt{G}\widehat{\texttt{FAR}}_b^*))$
Subsets	≈ 0	< 0
Two-level	≈ 0	< 0
Vertex	≈ 0	≈ 0
Double-or -nothing	≈ 0	≈ 0

Subsets bootstrap

At a fundamental level, the naive bootstrap resamples individual comparisons at the level of either identities or instances (of identities). However, ignoring the dependence structure present in the data can lead to significant undercoverage, as seen in the naive Wilson method. The subsets bootstrap [6] attempts to modify the conventional bootstrap by incorporating some of the dependency into the resampling process. Specifically, this bootstrap involves resampling with replacement at the identity level G times in each iteration. If the i-th identity is drawn in the b-th repetition, then all comparisons involving that identity are included in the bootstrap sample. More precisely, let the multinomial vector $(W_1, \ldots, W_G) \sim \text{Multinomial}(\mathcal{G}, (G^{-1}, \ldots, G^{-1}))$ such that $\sum_{i=1}^{G} W_i = G$. Then, we calculate:

$$\widehat{\texttt{FRR}}_b^* = \sum_{i=1}^{G} \frac{W_i \overline{Y}_{ii}}{G}, \; \widehat{\texttt{FAR}}_b^* = \frac{\sum_{\substack{i,j=1 \\ j \neq i}}^{G} W_i \overline{Y}_{ij}}{G(G-1)}. \quad (11)$$

By including all observations corresponding to a resampled identity, this bootstrap should better approximate the true distribution of $\widehat{\texttt{FRR}}$ and $\widehat{\texttt{FAR}}$ than the conventional bootstrap. Unfortunately, in a balanced setting, this procedure will underestimate the variance of $\widehat{\texttt{FAR}}$, as the following proposition demonstrates.

Proposition 4 (Bias of subsets bootstrap estimators)**.** *For the subsets bootstrap, we have* $\text{Bias}(\widehat{\texttt{FRR}}_b^*) = 0$*, and* $\text{Bias}(\widehat{\texttt{FAR}}_b^*) = 0$*. In addition, we have* $\text{Bias}(\text{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}_b^*)) = -\text{Var}(\widehat{\texttt{FRR}})$*, and* $\text{Bias}(\text{Var}^*(\sqrt{G}\widehat{\texttt{FAR}}_b^*)) = -\text{Var}(\widehat{\texttt{FAR}}) - \{\text{Var}(\overline{Y}_{12}) + 3(G-2)\text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})\}/(G-1)$.

Deriving the unbiasedness of the $\texttt{FRR}_b^*$ variance is straightforward while proving the same for $\texttt{FAR}_b^*$ requires more intricate analysis. The proposition shows that either taking bootstrap samples of size $G-1$ or rescaling the bootstrap metrics by $\sqrt{G(G-1)^{-1}}$ provide estimates whose distribution has unbiased variance for $\texttt{FRR}$. However, for $\texttt{FAR}$, there is a significant negative bias if $\text{Cov}(\overline{Y}_{12}, \overline{Y}_{13}) > 0$.

Two-level bootstrap

The two-level bootstrap is an attempt to address the undercoverage issue of the subsets bootstrap by employing two stages of resampling [39]. In the first stage, we use the subsets bootstrap, while in the second stage we employ a naive bootstrap to resample with replacement the instance comparisons belonging to the data subsets obtained in the first stage. In other words, after drawing $(W_1, \ldots, W_G) \sim \text{Multinomial}(\mathcal{G}, (G^{-1}, \ldots, G^{-1}))$, we compute

$$\widehat{\texttt{FRR}}_b^* = \frac{\sum_{i=1}^{G} W_i \overline{Y}_{ii}^*}{G}, \; \widehat{\texttt{FAR}}_b^* = \frac{\sum_{\substack{i,j=1 \\ j \neq i}}^{G} W_i \overline{Y}_{ij}^*}{G(G-1)}. \quad (12)$$

Here, $\overline{Y}_{ii}^*$ and $\sum_{j=1, j\neq i}^{G} \overline{Y}_{ij}^*$ are obtained by applying a naive bootstrap on each resampled data subset. By applying the law of total variance, we can show that $\text{Bias}(\text{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}_b^*)) = -\text{Var}(\widehat{\texttt{FRR}}) + (1 - 2/[M(M-1)])\text{Var}(Y_{(1,1),(1,2)}) + O(M^{-3})$. This implies that, after rescaling the estimates, the bootstrap may produce excess variation in $\texttt{FRR}$ computations. The derivation of the $\texttt{FAR}$ variance follows a similar strategy.

Vertex bootstrap

An alternative resampling procedure is the vertex bootstrap, which is commonly used for inference on networks in the social sciences [45]. This method involves resampling with replacement at the level of the identities G times, and then considering all comparisons between the resampled identities. In case of the $\texttt{FRR}$, this method is equivalent to the subsets bootstrap. For $\texttt{FAR}$ computations, the comparisons between instances belonging to the same identity are swapped with $\widehat{\texttt{FAR}}$; note that the original version of this bootstrap swaps it with a random sample from all comparisons. That is, for the b-th bootstrap sample, we take $(W_1, \ldots, W_G) \sim \text{Multinomial}(\mathcal{G}, (G^{-1}, \ldots, G^{-1}))$ and obtain $\widehat{\texttt{FRR}}_b^*$ as in expression (11), while for $\widehat{\texttt{FAR}}_b^*$, we use:

$$\widehat{\texttt{FAR}}_b^* = \sum_{i,j=1}^{G} W_i \Bigg[\frac{(W_i - 1)\widehat{\texttt{FAR}}\mathbb{1}(i=j)}{G(G-1)} + \frac{W_j \overline{Y}_{ij} \mathbb{1}(i \neq j)}{G(G-1)} \Bigg]. \quad (13)$$

Proposition 5 (Bias of vertex bootstrap estimators)**.** *For the vertex bootstrap, we have* $\mathrm{Bias}(\widehat{\texttt{FAR}}_b^*) = 0$*, and* $\mathrm{Bias}(\mathrm{Var}^*(\sqrt{G}\widehat{\texttt{FAR}}_b^*)) = [4(G-2)/G^3 + O(G^{-3})]\mathrm{Var}(\overline{Y}_{12}) - [28/[G(G-1)] + O(G^{-3})]\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$.

By comparing Proposition 4 and Proposition 5, one observes that in large samples, the expected FAR bootstrap variance of the vertex bootstrap is closer to the true variance than the subsets bootstrap. However, in finite samples, the vertex bootstrap overestimates the variances of the individual observations. As we will observe in the experiments in Section 6, this behavior can cause the bootstrap to achieve a coverage rate that is higher than nominal coverage.

Double-or-nothing bootstrap

The double-or-nothing bootstrap has been proposed in the context of separately exchangeable arrays [35], and it is a natural approach for analyzing matching tasks. In each iteration of this bootstrap procedure, we sample weights $W_i \overset{\text{iid}}{\sim} \mathrm{Uniform}\{0, 2\}$ for each identity $i \in \mathcal{G}$, and then we compute the estimates $\widehat{\texttt{FRR}}_b^*$ and $\widehat{\texttt{FAR}}_b^*$ as follows:

$$\begin{aligned} \widehat{\texttt{FRR}}_b^* &= \frac{\sum_{i=1}^G W_i \overline{Y}_{ii}}{\sum_{i=1}^G W_i}, \\ \widehat{\texttt{FAR}}_b^* &= \frac{\sum_{\substack{i,j=1 \\ j\neq i}}^G W_i W_j \overline{Y}_{ij}}{\sum_{\substack{i,j=1 \\ j\neq i}}^G W_i W_j}. \end{aligned} \tag{14}$$

Proposition 6 (Bias of double-or-nothing bootstrap estimators)**.** *For the double-or-nothing bootstrap, we have* $\mathrm{Bias}(\widehat{\texttt{FRR}}_b^*) = 0$*, and* $\mathrm{Bias}(\widehat{\texttt{FAR}}_b^*) = 0$*. In addition, we have* $\mathrm{Bias}(\mathrm{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}_b^*)) = -\mathrm{Var}(\widehat{\texttt{FRR}})$ *and* $\mathrm{Bias}(\mathrm{Var}^*(\sqrt{G}\widehat{\texttt{FAR}}_b^*)) = [G-1]^{-1}\{-(4G-2)/G\mathrm{Var}(\widehat{\texttt{FAR}}) + 4\mathrm{Var}(\overline{Y}_{12})\}$.

Thus, the FRR estimates obtained through subsets, vertex, and double-or-nothing bootstraps share similar properties. However, when computing FAR, this procedure, like the vertex bootstrap, tends to overestimate the variances of the individual identity-level comparisons.

5 Practical considerations

In Section 4, we described the construction of confidence intervals in the simplified balanced setting and analyzed their properties. In this section, we will provide an overview of practical considerations related to the implementation of these methods. Specifically, we will discuss how the reviewed methods can be extended and applied in the unbalanced setting. We will then address the construction of pointwise confidence intervals for the ROC curve. In Appendix B, we also cover the design of the protocols for the estimation of error rates and the associated uncertainty levels on large datasets. In Appendix C.5, we illustrate a power analysis.

5.1 Handling unbalanced datasets

In many datasets, the number of instances (say, face images) varies across identities (say, different people). We now demonstrate how the proposed methods can be applied to construct confidence intervals in this setting. Let $M_i \in \mathbb{N}$ denote a random variable representing to the finite number of instances belonging to the i-th identity. We assume that $M_i \overset{\text{i.i.d.}}{\sim} L$ for $i \in \mathcal{G}$ and some probability law L on natural numbers $\mathbb{N}$. We consider the following natural estimators of the error metrics:

$$\widehat{\texttt{FRR}} = \frac{\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}}{\sum_{i=1}^G \widetilde{M}_i}, \tag{15}$$

$$\widehat{\texttt{FAR}} = \frac{\sum_{i=1}^G \sum_{j=1, j\neq i}^G M_i M_j \overline{Y}_{ij}}{\sum_{i=1}^G \sum_{j=1, j\neq i}^G M_i M_j}, \tag{16}$$

where we have defined $\widetilde{M}_i = M_i(M_i - 1)$. Unlike in the balanced setting, these estimators are only unbiased as $G \to \infty$. The expressions of their variances are slightly more involved compared to (3) and (4), and will be discussed below.

Parametric methods

The construction of the parametric-based confidence intervals of the form (5) in the balanced setting can be easily extended to the unbalanced setting once estimators of $\mathrm{Var}(\sqrt{G}\widehat{\texttt{FRR}})$ and $\mathrm{Var}(\sqrt{G}\widehat{\texttt{FAR}})$ are available. To obtain an estimator for $\mathrm{Var}(\sqrt{G}\widehat{\texttt{FRR}})$, we can apply the Delta method

that yields:

$$\mathrm{Var}(\sqrt{G}\widehat{\texttt{FRR}}) = \frac{\mathrm{Var}(\widetilde{M}_1\overline{Y}_{11})}{\mathbb{E}[\widetilde{M}_1]^2} - 2\frac{\mathbb{E}[\widetilde{M}_1\overline{Y}_{11}]\mathrm{Cov}(\widetilde{M}_1\overline{Y}_{11},\widetilde{M}_1)}{\mathbb{E}[\widetilde{M}_1]^3} + \frac{\mathbb{E}[\widetilde{M}_1\overline{Y}_{11}]^2\mathrm{Var}(\widetilde{M}_1)}{\mathbb{E}[\widetilde{M}_1]^4}. \quad (17)$$

Note that when the number of instances per identity is constant, the expression in (17) reduces to $\mathrm{Var}(\overline{Y}_{11})$, which corresponds to the variance of $\widehat{\texttt{FRR}}$ in (3) in the balanced setting. Unlike the balanced setting, however, using plug-in estimators for the various terms in (17) may result in negative estimates of $\mathrm{Var}(\sqrt{G}\widehat{\texttt{FRR}})$. One way around this is to rely on a different plug-in estimator. For this derivation, we assume that M_i is independent of $Y_{(j,k),(l,p)}$ for any of the indices (even $i = j$ or $i = l$), i.e., the number of instances available for each identity is independent of whether the model classification is correct.[3] Then, appealing to the Delta method and this independence assumption, we obtain

$$\mathrm{Var}(\sqrt{G}\widehat{\texttt{FRR}}) = \frac{\mathbb{E}[\widetilde{M}_1^2(\overline{Y}_{11} - \texttt{FAR})^2]}{\mathbb{E}[\widetilde{M}_1]^2}. \quad (18)$$

The expression in (18) provides a simple way to estimate the FRR variance when the independence assumption holds.

To derive the plug-in estimators for $\mathrm{Var}(\sqrt{G}\widehat{\texttt{FAR}})$, we use similar arguments. Under the independence assumption described above, the Delta method yields

$$\mathrm{Var}(\sqrt{G}\widehat{\texttt{FAR}}) = \frac{2}{G-1}\frac{\mathbb{E}\left[M_1^2M_2^2(\overline{Y}_{12} - \texttt{FAR})^2\right]}{\mathbb{E}[M_1]^4} + \frac{4(G-2)}{(G-1)}\frac{\mathbb{E}\left[M_1^2M_2M_3(\overline{Y}_{12}\overline{Y}_{13} - \mathbb{E}[\overline{Y}_{12}\overline{Y}_{13}])\right]}{\mathbb{E}[M_1]^4}.$$

Once we have the estimators $\widehat{\mathrm{Var}}(\widehat{\texttt{FRR}})$ and $\widehat{\mathrm{Var}}(\widehat{\texttt{FAR}})$ for $\mathrm{Var}(\widehat{\texttt{FRR}})$ and $\mathrm{Var}(\widehat{\texttt{FAR}})$ respectively, we can construct Wilson confidence intervals using the recipe described in Section 4.1.

Resampling-based methods

Adapting the resampling-based methods for interval construction in the unbalanced setting is rather straightforward, similarly to the confidence intervals based on parametric methods. Since we have assumed that the M_i's are i.i.d., the methods will operate in the same way as in the balanced setting, with the only exception being that the metric computations follow (15). In other words, the resampling is performed at the identity level, regardless of the number of instances M_i for each identity. According to the following proposition, the subsets, vertex, and double-or-nothing bootstrap variances asymptotically converge to the target parameter in the case of the FRR.

Proposition 7 (Consistency of bootstrap estimators for FRR)**.** *Under the unbalanced setting, as* $G \to \infty$, $\mathrm{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}_b^*) - \mathrm{Var}(\sqrt{G}\widehat{\texttt{FRR}}) \xrightarrow{p} 0$, *where* $\widehat{\texttt{FRR}}_b^*$ *is the* FRR *estimate of the b-th subsets, vertex, or double-or-nothing bootstrap sample.*

This result also indicates that the bootstrap methods may be a suitable alternative for estimating $\mathrm{Var}(\widehat{\texttt{FRR}})$ instead of relying on the previously described plug-in estimator. Note that we have not talked about FAR in Proposition 7. Proving the consistency for FAR is a more intricate task as it involves computing the variance of non-independent terms. Thus, we do not pursue it in this paper.

Lastly, it is worth making a note of the scenario where the number of instances available for each identity is fixed instead of being random. In this situation, the variance computations undergo slight modifications. For instance, when computing the variance for the FRR, we have $\mathrm{Var}(\widehat{\texttt{FRR}}) = \sum_{i=1}^G \widetilde{M}_i^2\mathrm{Var}(\overline{Y}_i \mid \widetilde{M}_i)/(\sum_{i=1}^G \widetilde{M}_i^2)$, where $\widetilde{M}_i$ is a fixed quantity. Moreover, when applying the bootstrap method in this context, it is essential to resample conditioning on $\widetilde{M}$.

5.2 Pointwise intervals for ROC curves

In this section, we focus on the construction of pointwise confidence intervals for ROC curves, i.e.,

[3] This is a simplifying assumption that may not always hold true. For instance, in datasets containing mugshots like MORPH, individuals who have been arrested more frequently could be more identifiable because their facial images are more up-to-date.

intervals for error metrics such as FRR@FAR. While there is a wide range of techniques available [19, 29], we will limit our discussion to a few strategies that have proven effective in previous work and in our own experiments.

Parametric methods

To construct $1-\alpha$ pointwise confidence intervals for the ROC, we can use the Wilson method, as well as other parametric methods such as Wald intervals, as follows. First, we first compute a $1-\alpha_{\texttt{FAR}}$ interval for FAR, and we denote the lower and upper bounds of this interval as $\widehat{\texttt{FAR}}_{\text{lb}}$ and $\widehat{\texttt{FAR}}_{\text{ub}}$. Intuitively, these intervals contain FAR with high probability. We then estimate $1-\alpha$ confidence intervals for FRR at the thresholds t that yield $\widehat{\texttt{FAR}}_{\text{lb}}$ and $\widehat{\texttt{FAR}}_{\text{ub}}$. The resulting FRR@FAR interval is given by the region between the minima and maxima of the union of the two intervals. If the Wilson method is used, all FRR intervals computed on FAR values within $[\widehat{\texttt{FAR}}_{\text{lb}}, \widehat{\texttt{FAR}}_{\text{ub}}]$ will be nested within this region. Thus, as long as $\alpha_{\texttt{FAR}}$ is small, we should expect the resulting intervals to be conservative. In practice, we have found that even large values of $\alpha_{\texttt{FAR}}$ may yield intervals whose coverage is close to nominal. Therefore, the parameter $\alpha_{\texttt{FAR}}$ should be calibrated to the specific sample to avoid severe over- or under-coverage.

Nonparametric resampling-based methods

An alternative approach is to employ the nonparametric methods, such as bootstrapping techniques, which we have discussed in the previous sections. In this approach, we first obtain several ROC curves via some bootstrapping methods, such as the double-or-nothing or the vertex bootstraps. We then used these curves to construct the confidence intervals for FRR@FAR. For example, in the vertical averaging technique [19, 29], one computes $\widehat{\texttt{FRR}}_b^*$@FAR for each curve and then via the percentile bootstrap obtains the interval $[\widehat{\texttt{FRR}}^*_{(\lfloor B\alpha/2 \rfloor)}\texttt{@FAR}, \widehat{\texttt{FRR}}^*_{(\lceil B(1-\alpha/2) \rceil)}\texttt{@FAR}]$. However, as alluded to before, the main issue with the bootstrap methods is the interval under-coverage for error metrics close to the parameter boundary. This issue can be mitigated by imposing smoothness assumptions. For instance, instead of using the empirical ROC, we can estimate the ROC curve parametrically (e.g., with the widely used binormal model) or nonparametrically with kernels instead of using its empirical estimator [28].

6 Empirical evaluation

We will first present the experiments on synthetic data in the balanced setting, followed by experiments on on the MORPH dataset in the unbalanced setting. Additional experiments and results, including those on pointwise confidence intervals for the ROC, are reported in Appendix C of the Appendix.

6.1 Experiments on synthetic data

We consider the balanced setting with G identities and $M = 5$ instances for each identity. The embedding of the k-th image in the i-th identity are defined as: $X_{i,k} = \beta_i + \epsilon_{i,k}$, where $\beta_i, \epsilon_{i,k} \in \mathbb{R}^{128}$ with $\beta_i(d) \overset{\text{iid}}{\sim}$ Exponential(1) and $\epsilon_{i,k}(d) \overset{\text{iid}}{\sim} N(0,5)$ for $1 \leqslant d \leqslant 128$. We then define

$$Y_{(i,k),(j,l)} = \begin{cases} \mathbb{1}\left(\left\|\widetilde{X}_{i,k} - \widetilde{X}_{j,l}\right\|_2 > t\right) & \text{if } i = j, \\ \mathbb{1}\left(\left\|\widetilde{X}_{i,k} - \widetilde{X}_{j,l}\right\|_2 \leqslant t\right) & \text{if } i \neq j, \end{cases}$$

where we denote by $\widetilde{X}_{i,k} = X_{i,k}/\left\|X_{i,k}\right\|_2$ (and $\|\cdot\|_2$ denotes the Euclidean norm). Here, $t > 0$, $i, j \in \mathcal{G}$, and $1 \leqslant k, l \leqslant M$, leaving $Y_{(i,k),(i,k)}$ undefined. The error metrics estimation follows the description of Section 3. The thresholds t that yield the target error metrics, which are the underlying true parameters, were computed by resampling large datasets ($G = 2 \cdot 10^2$, $M = 10$). Coverage and average width of the intervals were then estimated by repeating the described sampling process 10^2 or 10^3 times.

In Figure 3, we compare estimated and nominal interval coverage for the methods discussed in Section 4 using synthetic data with $G = 50$. We can derive three key takeaways, which we hinted when discussing Figure 2 in Section 1.

First, when FAR is far from 0 (e.g., FAR $= 10^{-2}$ in this example), the Wilson intervals, vertex, and double-or-nothing bootstrap intervals achieve coverage close to nominal coverage. In contrast, the naive Wilson, subsets, and two-level bootstrap intervals are too narrow and under-cover. Our empirical analysis confirms this finding, where we observed that only the naive Wilson intervals

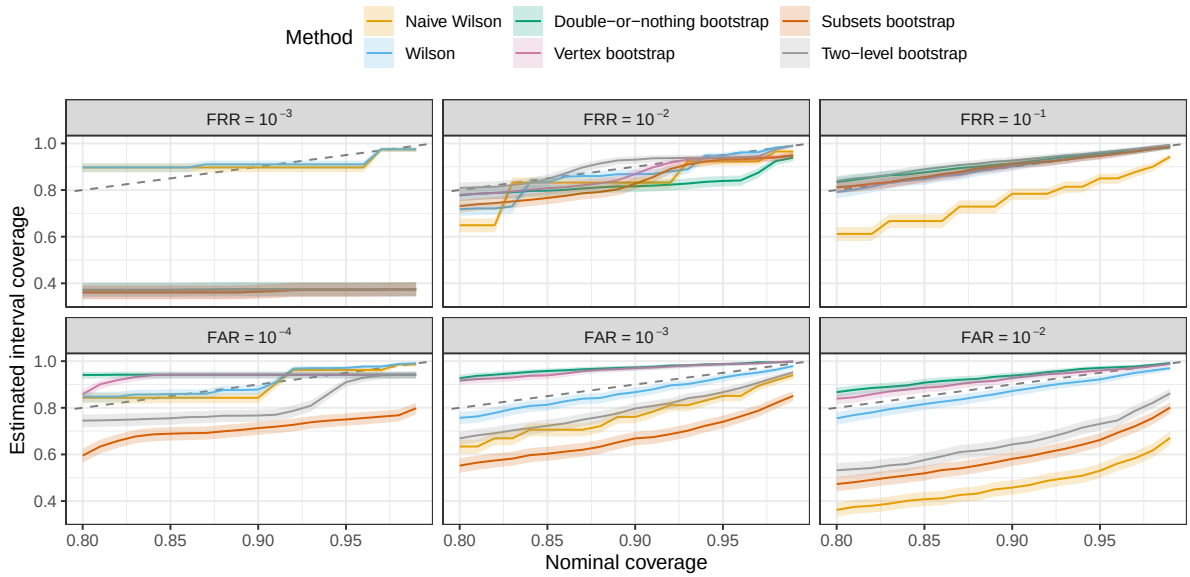

Fig. 3 Estimated interval coverage versus nominal coverage for FRR and FAR on synthetic data. Data contain $G = 50$ identities with $M = 5$ instances each. Colored lines and shaded bands indicate estimated coverage computed on 10^3 independent data replications and corresponding 95% naive Wilson intervals for the coverage respectively. Ideally, estimated coverage would coincide with nominal coverage (black dashed line).

Fig. 4 Estimated coverage of 95% confidence intervals for FAR versus sample size on synthetic data. Data contain G identities (horizontal axis) with $M = 5$ instances each.

suffer from under-coverage when FRR $= 10^{-2}$ or 10^{-1}. By contrast, the two-level bootstrap tends to slightly over-cover.

Second, when FAR $= 10^{-3}$, the vertex and double-or-nothing bootstraps overestimate the variance of the FAR and thus produce intervals that are too large to be useful. Wilson intervals achieve coverage close to the nominal level, whereas the remaining intervals under-cover.

Third, when error metrics are small, actual coverage does not scale linearly with nominal coverage for any of the methods. The use of the bootstrap is most problematic in case of FRR $= 10^{-3}$, as its distribution often results in a point mass at 0 and thus leads to the observed severe under-coverage. Although the issue is somewhat mitigated in case of larger (relatively to the sample size) error metrics such as FAR $= 10^{-4}$, the bootstrap still may not achieve nominal coverage.

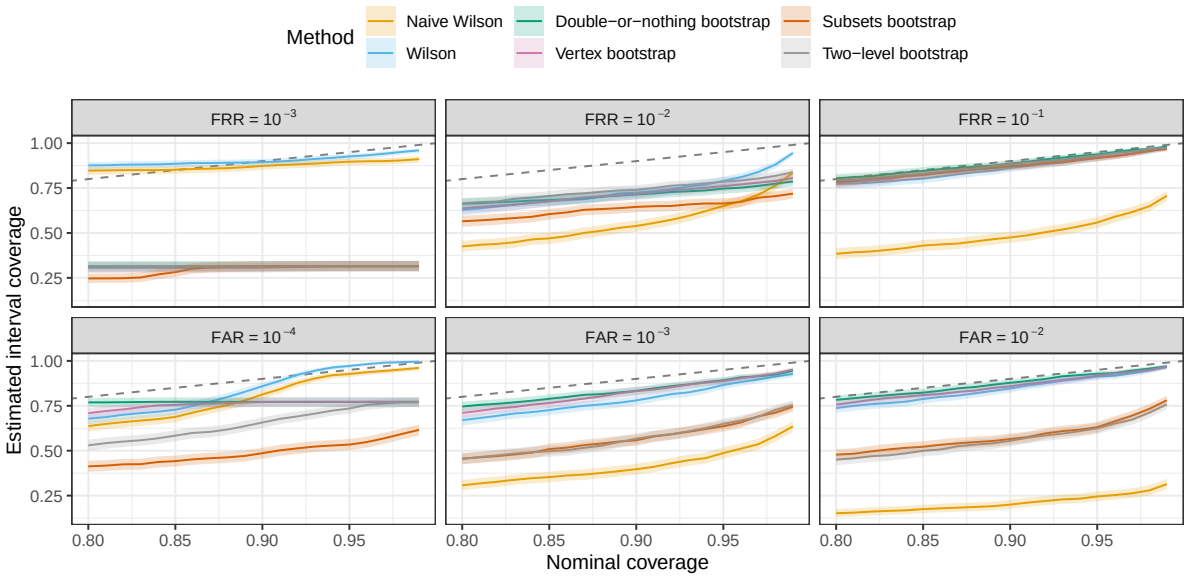

Fig. 5 Estimated interval coverage versus nominal coverage for FRR and FAR on the MORPH dataset. Samples were generated by resampling $G = 50$ identities from the original dataset without replacement.

Figure 2 and Figure 3 provide additional insights into the relationship between the two terms in (4), specifically $\mathrm{Var}(\overline{Y}_{12})$ and $\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$, respectively. Notably, as the FAR moves away from the parameter boundary, the ratio $\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})/\mathrm{Var}(\overline{Y}_{12})$ also increases. This phenomenon is linked to a more pronounced under-coverage of the naive Wilson intervals and a less pronounced over-coverage of the vertex and double-or-nothing bootstrap intervals.

More generally, the findings from our study highlight the crucial relationship between the sample size and the magnitude of the target error metric. Figure 4 provides additional insights by depicting how the coverage of 95% confidence intervals for FAR varies for $5 \leqslant G \leqslant 100$. We observe that the coverage of naive Wilson intervals decreases with an increasing sample size, as the covariance terms become the leading factor in $\mathrm{Var}(\widehat{\mathrm{FAR}})$. Wilson intervals always cover approximately at the right level. In the case of the vertex and double-or-nothing bootstraps, they under-cover when G is small and tend to over-cover for larger values of G, as can be observed for $G \geqslant 50$ (corresponding to 31k distinct comparisons) and $G \geqslant 20$ (5k comparisons) in case of $\mathrm{FAR} = 10^{-4}$ and $\mathrm{FAR} = 10^{-3}$, respectively. However, in line with our theoretical analysis, these intervals eventually achieve nominal coverage as G keeps increasing, as demonstrated by the case of $\mathrm{FAR} = 10^{-4}$.

6.2 Experiments on the MORPH dataset

The MORPH dataset [42] licensed for commercial use comprises approximately 400k mugshot images of 65k distinct individuals. As the number of images available for each individual varies, our estimation challenge is in the unbalanced setting. To expedite computations, we limited the number of face images per individual to 10. Using `dlib`'s face recognition model through `DeepFace` [27, 44], we extracted the 128-dimensional embeddings for the face images. We then split the data in half and a large random sample of images from one half was employed to estimate the thresholds that yield the target FRR and FAR using the Euclidean norm of the differences between the embeddings in the verification task. Construction of the confidence intervals was performed on the other half of the data. For this step, we generated datasets by randomly resampling without replacement G identities and considering all pairwise comparisons between images corresponding to those identities. Estimation of error metrics and interval construction followed the method descriptions in Section 5.1.

Here, we will only focus on the methods that have produced the most promising results on synthetic data. Therefore, we exclude the subsets and two-level bootstraps but retain the naive Wilson as a baseline. Figure 5 illustrates how the estimated coverage of confidence intervals for FAR and FRR of these methods vary with nominal coverage on the MORPH dataset with $G = 50$, where different identities can have different numbers of images. The behavior of the intervals somewhat mirrors our observations on synthetic data. More specifically, when the error metrics are close to zero ($\texttt{FAR} = 10^{-4}$ and $\texttt{FRR} = 10^{-3}$), the double-or-nothing and vertex bootstrap intervals significantly under-cover, while the Wilson intervals perform better in this regard, although their actual coverage does not scale linearly with nominal coverage. For larger error metrics, the naive Wilson intervals are too narrow. In the case of FRR, all intervals under-cover when $\texttt{FRR} = 10^{-2}$, while coverage is close to nominal when $\texttt{FRR} = 10^{-1}$. For FAR, all intervals tend to cover approximately at the nominal level when $\texttt{FAR} = 10^{-3}$ or 10^{-2}.

Despite our theoretical guarantees on Wilson's method, we notice that in some cases there is undercoverage. This may happen for two reasons. First, our analysis is based on the assumption that observations are independent across identities or individuals. This assumption may not hold. For example, in scenarios where the same background, the same camera, or the same lighting are used for all mugshots taken only on a particular day, sequential dependence across identities may occur. Second, estimating the true values of the metrics of interest is challenging when using limited amounts of data. As a result, the estimated interval coverage in the figures may not fully mirror the actual coverage in relation to the real value of the metric depicted in the figures.

7 Conclusions and recommendations

We aimed to provide guidelines for practitioners on how to compute confidence intervals for their experimental results. To this end, we explored the popular methods for constructing confidence intervals for error metrics in 1:1 matching tasks and evaluated their properties empirically and theoretically. Based on our findings:

(R1) We recommend the use of Wilson intervals with adjusted variance. They generally achieve coverage close to the nominal level. For large error metrics relative to sample size, the vertex and double-or-nothing bootstrap methods can be considered as good alternatives.

(R2) We strongly advise against using naive Wilson intervals, subsets, and two-level bootstrap techniques. They fail to achieve nominal coverage and may lead to incorrect inferences.

Our recommendations are especially relevant when test datasets are small-to-medium size, where all pairwise comparisons between instances are used in the computation of error metrics. On massive datasets non-overlapping sample pairs may be used, and data dependence may play a lesser role in the estimation of error metrics.

Our study is limited to 1:1 matching tasks. Computing confidence intervals for 1:N matching tasks is left open and will be the focus of future work. Concepts and insights presented here will likely serve as a useful starting point towards that goal.

Acknowledgments

The authors would like to thank Mathew Monfort and Yifan Xing for the insightful discussions and valuable feedback on the paper. The anonymous reviewers and the associate editor are also gratefully acknowledged for their constructive feedback that helped improve the clarity of the paper.

References

[1] Agresti, A. and Coull, B. A. (1998). Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126.

[2] Aronow, P. M., Samii, C., and Assenova, V. A. (2015). Cluster–robust variance estimation for dyadic data. *Political Analysis*, 23(4):564–577.

[3] Balakrishnan, G., Xiong, Y., Xia, W., and Perona, P. (2020). Towards causal benchmarking of bias in face analysis algorithms. In *European Conference on Computer Vision*, pages 547–563.

[4] Bhattacharyya, S. and Bickel, P. J. (2015). Subsampling bootstrap of count features of networks. *The Annals of Statistics*, 43(6):2384–2411.

[5] Bickel, P. J., Chen, A., and Levina, E. (2011). The method of moments and degree distributions for network models. *The Annals of Statistics*, 39(5):2280–2301.

[6] Bolle, R. M., Pankanti, S., and Ratha, N. K. (2000). Evaluation techniques for biometrics-based authentication systems (FRR). In *International Conference on Pattern Recognition*, pages 831–837.

[7] Bolle, R. M., Ratha, N. K., and Pankanti, S. (2004). Error analysis of pattern recognition systems—the subsets bootstrap. *Computer Vision and Image Understanding*, 93(1):1–33.

[8] Brown, L. D., Cai, T. T., and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2):101–133.

[9] Cameron, A. C., Gelbach, J. B., and Miller, D. L. (2011). Robust inference with multiway clustering. *Journal of Business and Economic Statistics*, 29(2):238–249.

[10] Cameron, A. C. and Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372.

[11] Casella, G. and Berger, R. L. (2021). *Statistical Inference*. Cengage Learning.

[12] Chouldechova, A., Deng, S., Wang, Y., Xia, W., and Perona, P. (2022). Unsupervised and semi-supervised bias benchmarking in face recognition. In *European Conference on Computer Vision*, pages 289–306.

[13] Conti, J.-R. and Clémençon, S. (2022). Assessing performance and fairness metrics in face recognition-bootstrap methods. *arXiv preprint arXiv:2211.07245*.

[14] Davezies, L., D'Haultfœuille, X., and Guyonvarch, Y. (2021). Empirical process results for exchangeable arrays. *The Annals of Statistics*, 49(2):845–862.

[15] Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and Their Application*. Cambridge University Press.

[16] Desplanques, B., Thienpondt, J., and Demuynck, K. (2020). Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *arXiv preprint arXiv:2005.07143*.

[17] DiCiccio, T. J. and Efron, B. (1996). Bootstrap confidence intervals. *Statistical Science*, 11(3):189–228.

[18] Fafchamps, M. and Gubert, F. (2007). Risk sharing and network formation. *American Economic Review*, 97(2):75–79.

[19] Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning*, 31(1):1–38.

[20] Field, C. A. and Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(3):369–390.

[21] Graham, B. S. (2020). Network data. In *Handbook of Econometrics*, volume 7, pages 111–218. Elsevier.

[22] Green, A. and Shalizi, C. R. (2022). Bootstrapping exchangeable random graphs. *Electronic Journal of Statistics*, 16(1):1058–1095.

[23] Grother, P., Ngan, M., and Hanaoka, K. (2019). *Face recognition vendor test (FVRT): Part 3, demographic effects*. National Institute of Standards and Technology Gaithersburg, MD.

[24] Hoff, P. (2021). Additive and multiplicative effects network models. *Statistical Science*, 36(1):34–50.

[25] Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098.

[26] Kearns, M. and Roth, A. (2019). *The Ethical Algorithm: The Science of Socially Aware*

Algorithm Design. Oxford University Press.

[27] King, D. E. (2009). Dlib-ml: A machine learning toolkit. *Journal of Machine Learning Research*, 10:1755–1758.

[28] Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data*. Chapman and Hall/CRC.

[29] Macskassy, S., Provost, F., and Rosset, S. (2005). Pointwise ROC confidence bounds: An empirical evaluation. In *International Conference on Machine Learning*.

[30] McCullagh, P. (2000). Resampling and exchangeable arrays. *Bernoulli*, pages 285–301.

[31] Menzel, K. (2021). Bootstrap with cluster-dependence in two or more dimensions. *Econometrica*, 89(5):2143–2188.

[32] Miao, W. and Gastwirth, J. L. (2004). The effect of dependence on confidence intervals for a population proportion. *The American Statistician*, 58(2):124–130.

[33] Mitra, S., Savvides, M., and Brockwell, A. (2007). Statistical performance evaluation of biometric authentication systems using random effects models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(4):517–530.

[34] Ni, J., Li, J., and McAuley, J. (2019). Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 188–197.

[35] Owen, A. B. and Eckles, D. (2012). Bootstrapping data arrays of arbitrary order. *The Annals of Applied Statistics*, 6(3):895–927.

[36] Phillips, P. J., Flynn, P. J., Bowyer, K. W., Bruegge, R. W. V., Grother, P. J., Quinn, G. W., and Pruitt, M. (2011). Distinguishing identical twins by face recognition. In *International Conference on Automatic Face and Gesture Recognition*, pages 185–192.

[37] Phillips, P. J., Grother, P., Micheals, R., Blackburn, D. M., Tabassi, E., and Bone, M. (2003). Face recognition vendor test 2002. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*.

[38] Phillips, P. J., Yates, A. N., Hu, Y., Hahn, C. A., Noyes, E., Jackson, K., Cavazos, J. G., Jeckeln, G., Ranjan, R., Sankaranarayanan, S., et al. (2018). Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences*, 115(24):6171–6176.

[39] Poh, N., Martin, A., and Bengio, S. (2007). Performance generalization in biometric authentication using joint user-specific and sample bootstraps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):492–498.

[40] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

[41] Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., Chou, J.-C., Yeh, S.-L., Fu, S.-W., Liao, C.-F., Rastorgueva, E., Grondin, F., Aris, W., Na, H., Gao, Y., Mori, R. D., and Bengio, Y. (2021). SpeechBrain: A general-purpose speech toolkit. arXiv:2106.04624.

[42] Ricanek, K. and Tesafaye, T. (2006). MORPH: A longitudinal image database of normal adult age-progression. In *International Conference on Automatic Face and Gesture Recognition*, pages 341–345.

[43] Seabold, S. and Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the 9th Python in Science Conference*.

[44] Serengil, S. I. and Ozpinar, A. (2020). Lightface: A hybrid deep face recognition framework.

In *Innovations in Intelligent Systems and Applications Conference*, pages 23–27.

[45] Snijders, T. A., Borgatti, S. P., et al. (1999). Non-parametric standard errors and tests for network statistics. *Connections*, 22(2):161–170.

[46] Tabord-Meehan, M. (2019). Inference with dyadic data: Asymptotic behavior of the dyadic-robust t-statistic. *Journal of Business and Economic Statistics*, 37(4):671–680.

[47] Van Horn, G., Cole, E., Beery, S., Wilber, K., Belongie, S., and Mac Aodha, O. (2021). Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893.

[48] Vangara, K., King, M. C., Albiero, V., and Bowyer, K. (2019). Characterizing the variability in face recognition accuracy relative to race. In *Conference on Computer Vision and Pattern Recognition Workshops*.

[49] Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer.

[50] Wilson, E. B. (1927). Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212.

[51] Wu, J. C., Martin, A. F., Greenberg, C. S., and Kacker, R. N. (2016). The impact of data dependence on speaker recognition evaluation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(1):5–18.

[52] Xiao, S., Liu, Z., Zhang, P., and Muennighoff, N. (2023). C-pack: Packaged resources to advance general chinese embedding.

[53] Zeileis, A., Köll, S., and Graham, N. (2020). Various versatile variances: An object-oriented implementation of clustered covariances in R. *Journal of Statistical Software*, 95:1–36.

Supplementary Material for "Confidence Intervals for Error Rates in 1:1 Matching Tasks: Critical Statistical Analysis and Recommendations"

This document acts as a supplement to the paper "Confidence Intervals for Error Rates in 1:1 Matching Tasks: Critical Statistical Analysis and Recommendations." The supplement is organized as follows.

(A) In Appendix A, we provide proof of all the theoretical claims in the main paper.
  (1) Appendix A.1 contains proofs for parametric methods in Section 4.1.
  (2) Appendix A.2 contains proofs for resampling-based methods in Section 4.2
  (3) Appendix A.3 contains proofs for unbalanced datasets in Section 5.1.

(B) In Appendix B, we describe protocol design strategies (i.e., sampling) for the estimation of error rates and their associated uncertainty on large datasets.

(C) In Appendix C, we provide additional experiments, supplementing those in Sections 1 and 6.
  (1) Appendix C.1 provides illustrations comparing confidence interval widths.
  (2) Appendix C.2 examines variance estimation accuracy against sample size.
  (3) Appendix C.3 contains experiments on pointwise intervals for the ROC.
  (4) Appendix C.4 contains additional experiments on text, image, and audio data.
  (5) Appendix C.5 contains power analyses for 1:1 matching tasks.

Appendix A Proofs of theoretical results

A.1 Proofs for parametric methods in Section 4.1

A.1.1 Proof of Proposition 1 (normality of scaled error rates)

As explained in the main paper, because identity-level observations are assumed to be independent, the case of FRR in Proposition 1 follows from applying the central limit theorem. The case of the FAR follows from Proposition 3.2 in Tabord-Meehan [46].

A.1.2 Proof of Proposition 2 (consistency of plug-in variance estimators)

The convergence in probability of $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\mathtt{FRR}})$ to $\mathrm{Var}(\overline{Y}_{11})$ simply follows from an application of the weak law of large numbers. In the following, we will show the convergence in probability of $\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\mathtt{FAR}}) - (G-1)^{-1}(2\mathrm{Var}(\overline{Y}_{12}) + 4(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13}))$ to 0. We begin by recalling the estimator:

$$\widehat{\mathrm{Var}}(\sqrt{G}\widehat{\mathtt{FAR}}) = \frac{2}{G-1}\left[\widehat{\mathrm{Var}}(\overline{Y}_{12}) + 2(G-2)\widehat{\mathrm{Cov}}(\overline{Y}_{12}, \overline{Y}_{13})\right], \tag{A1}$$

where the components $\widehat{\mathrm{Var}}(\overline{Y}_{12})$ and $\widehat{\mathrm{Cov}}(\overline{Y}_{12}, \overline{Y}_{13})$ are defined as:

$$\widehat{\mathrm{Var}}(\overline{Y}_{12}) = \frac{1}{G(G-1)}\sum_{i=1}^{G}\sum_{j=1, j\neq i}^{G}(\overline{Y}_{ij} - \widehat{\mathtt{FAR}})^2, \tag{A2}$$

$$\widehat{\mathrm{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}) = \frac{1}{G(G-1)(G-2)}\sum_{i=1}^{G}\sum_{\substack{j=1\\ j\neq i}}^{G}\sum_{\substack{k=1\\ k\neq i,j}}^{G}(\overline{Y}_{ij} - \widehat{\mathtt{FAR}})(\overline{Y}_{ik} - \widehat{\mathtt{FAR}}). \tag{A3}$$

We want to show that, as $G \to \infty$, $\widehat{\text{Var}}(\overline{Y}_{12}) \xrightarrow{\text{p}} \text{Var}(\overline{Y}_{12})$, and $\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}) \xrightarrow{\text{p}} \text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$. If these conditions are verified, then $\widehat{\text{Var}}(\sqrt{G}\widehat{\texttt{FAR}}) - \text{Var}(\sqrt{G}\widehat{\texttt{FAR}}) \xrightarrow{\text{p}} 0$ by Slutsky's theorem.

Consistency of $\widehat{\text{Var}}(\overline{Y}_{12})$

By Chebyshev's inequality, we have

$$\mathbb{P}(|\widehat{\text{Var}}(\overline{Y}_{12}) - \text{Var}(\overline{Y}_{12})| \geqslant t) \leqslant \frac{\mathbb{E}\left[\left(\widehat{\text{Var}}(\overline{Y}_{12}) - \text{Var}(\overline{Y}_{12})\right)^2\right]}{t^2}, \tag{A4}$$

for any $t > 0$. We will now bound the numerator of (A4). Decompose the numerator into:

$$\mathbb{E}\left[\left(\widehat{\text{Var}}(\overline{Y}_{12})) - \text{Var}(\overline{Y}_{12})\right)^2\right] = \underbrace{\text{Var}(\widehat{\text{Var}}(\overline{Y}_{12}))}_{\textbf{Term 1}} + \underbrace{\left(\mathbb{E}\left[\widehat{\text{Var}}(\overline{Y}_{12})\right] - \text{Var}(\overline{Y}_{12})\right)^2}_{\textbf{Term 2}}. \tag{A5}$$

We will show below that both the two terms on the right-hand side of (A5) are $O(G^{-1})$.

Term 1 The first term in (A5) is equal to

$$\text{Var}(\widehat{\text{Var}}(\overline{Y}_{12})) = \frac{1}{G(G-1)}\Big[2\text{Var}((\overline{Y}_{12} - \widehat{\texttt{FAR}})^2) + 4(G-2)\text{Cov}((\overline{Y}_{12} - \widehat{\texttt{FAR}})^2, (\overline{Y}_{13} - \widehat{\texttt{FAR}})^2) + (G-2)(G-3)\text{Cov}((\overline{Y}_{12} - \widehat{\texttt{FAR}})^2, (\overline{Y}_{34} - \widehat{\texttt{FAR}})^2)\Big]. \tag{A6}$$

It is easy to see that all terms are $O(G^{-1})$ or of smaller order.

Term 2 The second term on the right-hand side of (A5) is equal to

$$\left[\mathbb{E}\left[\widehat{\text{Var}}(\overline{Y}_{12})\right] - \text{Var}(\overline{Y}_{12})\right]^2 = -\left[\frac{\text{Var}(\overline{Y}_{12}) + 4(G-2)\text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)}\right]^2 = O(G^{-2}). \tag{A7}$$

The consistency of $\widehat{\text{Var}}(\overline{Y}_{12})$ then follows by combining the results in (A6) and (A7) with the inequality in (A4).

Consistency of $\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13})$

By Chebyshev's inequality,

$$\mathbb{P}(|\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}) - \text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})| \geqslant t) \leqslant \frac{\mathbb{E}\left[\left(\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}) - \text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})\right)^2\right]}{t^2}, \tag{A8}$$

for any $t > 0$. We now proceed to bound the numerator of (A8). Note that

$$\mathbb{E}\left[\left(\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}) - \text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})\right)^2\right] = \underbrace{\text{Var}(\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}))}_{\textbf{Term 3}} + \underbrace{\left(\mathbb{E}\left[\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13})\right] - \text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})\right)^2}_{\textbf{Term 4}}. \tag{A9}$$

To complete the proof, we will show below that each of the two terms on the right-hand side of (A9) is $O(G^{-1})$.

Term 3 We start with the first term, the variance of the covariance estimator. We can rewrite

$$\mathrm{Var}(\widehat{\mathrm{Cov}}(\overline{Y}_{12},\overline{Y}_{13})) = \sum_{i=1}^{G}\sum_{\substack{j=1\\j\neq i}}^{G}\sum_{\substack{k=1\\k\neq i,j}}^{G}\sum_{l=1}^{G}\sum_{\substack{m=1\\m\neq l}}^{G}\sum_{\substack{n=1\\n\neq l,m}}^{G}\frac{\mathrm{Cov}\left\{(\overline{Y}_{ij}-\widehat{\mathtt{FAR}})(\overline{Y}_{ik}-\widehat{\mathtt{FAR}}),(\overline{Y}_{lm}-\widehat{\mathtt{FAR}})(\overline{Y}_{ln}-\widehat{\mathtt{FAR}})\right\}}{G^2(G-1)^2(G-2)^2}.$$

In order to show that it converges to 0, we need to prove that the number of nonzero covariance terms is of the order smaller than G^6.

- Terms involving $\mathrm{Cov}(\overline{Y}_{ij}\overline{Y}_{ik},\overline{Y}_{lm}\overline{Y}_{ln})$: These terms will be zero when all indices are different, that is in $G!/(G-6)!$ cases. Thus, $[G(G-1)(G-2)]^2 - G!/(G-6)! = O(G^5)$ of the terms in the sum above will be nonzero.
- Terms involving $\mathrm{Cov}(\overline{Y}_{ij}\overline{Y}_{ik},\overline{Y}_{lm}\widehat{\mathtt{FAR}})$: We have

$$\begin{aligned}
&\mathrm{Cov}(\overline{Y}_{ij}\overline{Y}_{ik},\overline{Y}_{lm}\widehat{\mathtt{FAR}})\\
&= \frac{1}{G(G-1)}\mathrm{Cov}\left\{\overline{Y}_{ij}\overline{Y}_{ik}, 2\overline{Y}_{lm}^2 + 4\sum_{\substack{n=1\\n\neq l,m}}^{G}\overline{Y}_{lm}\overline{Y}_{ln} + \sum_{\substack{n=1\\n\neq l,m}}^{G}\sum_{\substack{p=1\\p\neq l,m,n}}^{G}\overline{Y}_{lm}\overline{Y}_{np}\right\}\\
&= \frac{\mathrm{Cov}(\overline{Y}_{ij}\overline{Y}_{ik},2\overline{Y}_{lm}^2)}{G(G-1)} + 4\sum_{\substack{n=1\\n\neq l,m}}^{G}\frac{\mathrm{Cov}\left(\overline{Y}_{ij}\overline{Y}_{ik},\overline{Y}_{lm}\overline{Y}_{ln}\right)}{G(G-1)} + \sum_{\substack{n=1\\n\neq l,m}}^{G}\sum_{\substack{p=1\\p\neq l,m,n}}^{G}\frac{\mathrm{Cov}(\overline{Y}_{ij}\overline{Y}_{ik},\overline{Y}_{lm}\overline{Y}_{np})}{G(G-1)}.
\end{aligned}$$

The first term will be nonzero when $\overline{Y}_{ij}\overline{Y}_{ik}$ and $\overline{Y}_{lm}^2$ share any of the indices, hence

$$\begin{aligned}
&\frac{1}{G(G-1)}\sum_{i=1}^{G}\sum_{\substack{j=1\\j\neq i}}^{G}\sum_{\substack{k=1\\k\neq i,j}}^{G}\sum_{l=1}^{G}\sum_{\substack{m=1\\m\neq l}}^{G}\sum_{\substack{n=1\\n\neq l,m}}^{G}\mathrm{Cov}(\overline{Y}_{ij}\overline{Y}_{ik},2\overline{Y}_{lm}^2)\\
&= \frac{G(G-1)(G-2)^2}{G(G-1)}\sum_{l=1}^{G}\sum_{\substack{m=1\\m\neq l}}^{G}\mathrm{Cov}(\overline{Y}_{12}\overline{Y}_{13},2\overline{Y}_{lm}^2),
\end{aligned}$$

which is $O(G^3)$. The second term is $O(G^4)$, while the third term is $O(G^5)$.

- Terms involving $\mathrm{Cov}(\widehat{\mathtt{FAR}}^2,\widehat{\mathtt{FAR}}^2)$: We have

$$\begin{aligned}
&\mathrm{Cov}(\widehat{\mathtt{FAR}}^2,\widehat{\mathtt{FAR}}^2)\\
&= \frac{1}{G^2(G-1)^2}\sum_{i=1}^{G}\sum_{\substack{j=1\\j\neq i}}^{G}\mathrm{Cov}\left\{2\overline{Y}_{ij}^2 + 4\sum_{\substack{k=1\\k\neq i,j}}^{G}\overline{Y}_{ij}\overline{Y}_{ik} + \sum_{\substack{k=1\\k\neq i,j}}^{G}\sum_{\substack{l=1\\l\neq i,j,k}}^{G}\overline{Y}_{ij}\overline{Y}_{kl},\widehat{\mathtt{FAR}}^2\right\}\\
&= \frac{1}{G(G-1)}\mathrm{Cov}\left\{2\overline{Y}_{12}^2 + 4\sum_{\substack{k=1\\k\neq 1,2}}^{G}\overline{Y}_{12}\overline{Y}_{1k} + \sum_{\substack{k=1\\k\neq 1,2}}^{G}\sum_{\substack{l=1\\l\neq 1,2,k}}^{G}\overline{Y}_{12}\overline{Y}_{kl},\widehat{\mathtt{FAR}}^2\right\}\\
&= \frac{1}{G(G-1)}\mathrm{Cov}\left\{2\overline{Y}_{12}^2 + 4(G-2)\overline{Y}_{12}\overline{Y}_{13} + (G-2)(G-3)\overline{Y}_{12}\overline{Y}_{34},\widehat{\mathtt{FAR}}^2\right\}.
\end{aligned}$$

The leading term in this expression is

$$\frac{(G-2)(G-3)}{G^3(G-1)^3}\sum_{i=1}^{G}\sum_{\substack{j=1\\ j\neq i}}^{G}\sum_{\substack{k=1\\ k\neq i,j}}^{G}\sum_{\substack{l=1\\ l\neq i,j,k}}^{G}\mathrm{Cov}(\overline{Y}_{12}\overline{Y}_{34},\overline{Y}_{ij}\overline{Y}_{kl})=O(G^{-1}).$$

- Terms involving $\mathrm{Cov}(\overline{Y}_{ij}\overline{Y}_{ik},\widehat{\mathtt{FAR}}^2)$ and $\mathrm{Cov}(\overline{Y}_{ij}\widehat{\mathtt{FAR}},\widehat{\mathtt{FAR}}^2)$: These terms are handled in a similar manner and their proofs are omitted.

Thus, we have thus shown that

$$\mathrm{Var}(\widehat{\mathrm{Cov}}(\overline{Y}_{12},\overline{Y}_{13}))=O(G^5/G^6)=O(G^{-1}). \tag{A10}$$

Term 4 We now turn to the second term, which is the squared bias. We have

$$\mathbb{E}\left[\widehat{\mathrm{Cov}}(\overline{Y}_{12},\overline{Y}_{13})\right]=\left[1-\frac{4(G-2)}{G(G-1)}\right]\mathbb{E}[\overline{Y}_{12}\overline{Y}_{13}]-\frac{2}{G(G-1)}\mathbb{E}[\overline{Y}_{12}^2]-\mathtt{FAR}^2\frac{(G-2)(G-3)}{G(G-1)}.$$

It follows that the bias is given by

$$\mathbb{E}\left[\widehat{\mathrm{Cov}}(\overline{Y}_{12},\overline{Y}_{13})\right]-\mathrm{Cov}(\overline{Y}_{12},\overline{Y}_{13})=-\frac{4(G-2)}{G(G-1)}\mathbb{E}[\overline{Y}_{12}\overline{Y}_{13}]-\frac{2}{G(G-1)}\mathbb{E}[\overline{Y}_{12}^2]-\frac{2(2G-3)}{G(G-1)}\mathtt{FAR}^2.$$

Thus, we have

$$\left(\mathbb{E}\widehat{\mathrm{Cov}}(\overline{Y}_{12},\overline{Y}_{13})-\mathrm{Cov}(\overline{Y}_{12},\overline{Y}_{13})\right)^2=O(G^{-2}), \tag{A11}$$

which goes to 0 as $G\to\infty$.

Putting (A10) and (A11) together, along with (A8), the result then follows. This completes the proof of the consistency of $\widehat{\mathrm{Cov}}(\overline{Y}_{12},\overline{Y}_{13})$.

A.1.3 Proof of Proposition 3 (equivalence of plug-in and jackknife variance estimators)

Recall the form of the jackknife variance estimator of $\mathrm{Var}(\sqrt{G}\widehat{\mathtt{FAR}})$:

$$\widehat{\mathrm{Var}}_{JK}(\sqrt{G}\widehat{\mathtt{FAR}})=\frac{(G-2)^2}{G}\sum_{i=1}^{G}(\widehat{\mathtt{FAR}}_{-i}-\widehat{\mathtt{FAR}})^2-2\frac{\widehat{\mathrm{Var}}(\overline{Y}_{12})}{G-1}, \tag{A12}$$

where $\widehat{\mathtt{FAR}}_{-i}$ is defined as

$$\widehat{\mathtt{FAR}}_{-i}=\frac{1}{(G-1)(G-2)}\sum_{j=1}^{G}\sum_{\substack{k=1,\\ k\neq j}}^{G}\overline{Y}_{jk}\mathbb{1}(\{j\neq i\}\cap\{k\neq i\}). \tag{A13}$$

Recall also the estimator for $\mathrm{Var}(\overline{Y}_{12})$:

$$\widehat{\mathrm{Var}}(\overline{Y}_{12})=\frac{1}{G(G-1)}\sum_{i=1}^{G}\sum_{\substack{j=1,\\ j\neq i}}^{G}(\overline{Y}_{ij}-\widehat{\mathtt{FAR}})^2. \tag{A14}$$

Through a series of algebraic manipulations, we will show that after substituting for (A13) and (A14), the expression (A12) simplifies to plug-in estimator from (9).

Towards that end, we start by expanding the sum in the first term on the right-hand side of (A12):

$$\begin{aligned}
&\sum_{i=1}^{G}(\widehat{\texttt{FAR}}_{-i} - \widehat{\texttt{FAR}})^2 \\
&= \sum_{i=1}^{G}\left(\frac{\sum_{k=1}^{G}\sum_{\substack{l=1\\l\neq k}}^{G}\overline{Y}_{kz} - 2\sum_{\substack{j=1\\j\neq i}}^{G}\overline{Y}_{ij}}{(G-1)(G-2)} - \widehat{\texttt{FAR}}\right)^2 \\
&= \frac{4}{(G-2)^2}\sum_{i=1}^{G}\left(\sum_{\substack{j=1\\j\neq i}}^{G}\frac{\overline{Y}_{ij}}{G-1} - \widehat{\texttt{FAR}}\right)^2 \\
&= \frac{4}{(G-2)^2(G-1)^2}\sum_{i=1}^{G}\left[\sum_{\substack{j=1\\j\neq i}}^{G}(\overline{Y}_{ij} - \widehat{\texttt{FAR}})^2 + \sum_{\substack{j=1\\j\neq i}}^{G}\sum_{\substack{k=1\\k\neq i,j}}^{G}(\overline{Y}_{ij} - \widehat{\texttt{FAR}})(\overline{Y}_{ik} - \widehat{\texttt{FAR}})\right].
\end{aligned} \tag{A15}$$

Moving the appropriate factor across and subtracting the second term on the right-hand side of (A12), we arrive at

$$\begin{aligned}
&\frac{(G-2)^2}{G}\sum_{i=1}^{G}(\widehat{\texttt{FAR}}_{-i} - \widehat{\texttt{FAR}})^2 - 2\frac{\widehat{\text{Var}}(\overline{Y}_{12})}{G-1} \\
&= \frac{4\sum_{i=1}^{G}\sum_{\substack{j=1\\j\neq i}}^{G}(\overline{Y}_{ij} - \widehat{\texttt{FAR}})^2}{G(G-1)^2} + \frac{4\sum_{i=1}^{G}\sum_{\substack{j=1\\j\neq i}}^{G}\sum_{\substack{k=1\\k\neq i,j}}^{G}(\overline{Y}_{ij} - \widehat{\texttt{FAR}})(\overline{Y}_{ik} - \widehat{\texttt{FAR}})}{G(G-1)^2} - 2\frac{\widehat{\text{Var}}(\overline{Y}_{12})}{G-1} \\
&= \frac{2\widehat{\text{Var}}(\overline{Y}_{12}) + 4\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13})(G-2)}{G-1} + 2\frac{\widehat{\text{Var}}(\overline{Y}_{12})}{G-1} - 2\frac{\widehat{\text{Var}}(\overline{Y}_{12})}{G-1} \\
&= \frac{2}{G-1}\widehat{\text{Var}}(\overline{Y}_{12}) + \frac{4(G-2)}{G-1}\widehat{\text{Cov}}(\overline{Y}_{12}, \overline{Y}_{13}),
\end{aligned} \tag{A16}$$

Noting that the (A16) matches with (9). we have that $\widehat{\text{Var}}_{JK}(\sqrt{G}\widehat{\texttt{FAR}}) = \widehat{\text{Var}}(\sqrt{G}\widehat{\texttt{FAR}})$, as claimed.

A.2 Proofs for resampling-based methods in Section 4.2

A.2.1 Proof of Proposition 4 (bias of subsets bootstrap estimators)

Recall that $\widehat{\texttt{FRR}}_b^*$ and $\widehat{\texttt{FAR}}_b^*$ indicate the FRR and FAR estimates respectively based on the b-th bootstrap sample. The proofs for various statements in the proposition are separated below.

- Showing that $\text{Bias}(\widehat{\texttt{FRR}}_b^*) = 0$ and $\text{Bias}(\text{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}^*)) = -\text{Var}(\widehat{\texttt{FRR}})$ is straightforward.
  - For $\text{Bias}(\widehat{\texttt{FRR}}_b^*)$, it is easy to see that

$$\mathbb{E}[\mathbb{E}^*[\widehat{\texttt{FRR}}_b^*]] = \frac{1}{G}\sum_{i=1}^{G}\mathbb{E}[\mathbb{E}^*[W_i]\overline{Y}_{ii}] = \frac{1}{G}\mathbb{E}[\overline{Y}_{ii}] = \texttt{FRR}.$$

    Hence, $\text{Bias}(\widehat{\texttt{FRR}}_b^*) = 0$.

– Towards computing $\text{Bias}(\sqrt{G}\text{Var}^*(\widehat{\texttt{FRR}}_b^*))$, observe that

$$\begin{aligned}
\mathbb{E}[\text{Var}^*[\widehat{\texttt{FRR}}_b^*]] &= \frac{1}{G^2}\sum_{i=1}^{G}\mathbb{E}\left\{\text{Var}^*(W_i)\overline{Y}_{ii}^2 + \text{Cov}^*(\overline{W}_i, \overline{W}_k)\sum_{\substack{k=1\\k\neq i}}^{G}\overline{Y}_{ii}\overline{Y}_{kk}\right\} \\
&= \frac{1}{G^2}\sum_{i=1}^{G}\mathbb{E}\left\{\frac{G-1}{G}\overline{Y}_{ii}^2 - \frac{1}{G}\sum_{\substack{k=1\\k\neq i}}^{G}\overline{Y}_{ii}\overline{Y}_{kk}\right\} \\
&= \frac{G-1}{G^2}\mathbb{E}[\overline{Y}_{11}^2] - \frac{G-1}{G^2}\texttt{FRR} \\
&= \frac{G-1}{G}\text{Var}(\widehat{\texttt{FRR}}).
\end{aligned}$$

Thus, we have $\text{Bias}(\text{Var}^*(\sqrt{G}\widehat{\texttt{FRR}})) = (G-1)\text{Var}(\widehat{\texttt{FRR}}) - G\text{Var}(\widehat{\texttt{FRR}}) = -\text{Var}(\widehat{\texttt{FAR}})$, as claimed.

- Obtaining expressions for $\text{Bias}(\widehat{\texttt{FAR}}_b^*)$ and $\text{Bias}(\text{Var}^*(\sqrt{G}\widehat{\texttt{FAR}}_b^*))$ is slightly more involved.
  – For $\text{Bias}(\widehat{\texttt{FAR}}_b^*)$, note that

$$\mathbb{E}[\mathbb{E}^*[\widehat{\texttt{FAR}}_b^*]] = \frac{1}{G(G-1)}\sum_{i=1}^{G}\mathbb{E}\left[\mathbb{E}^*[W_i]\sum_{\substack{j=1\\j\neq i}}^{G}\overline{Y}_{ij}\right] = \frac{1}{G(G-1)}\sum_{i=1}^{G}\sum_{\substack{j=1\\j\neq i}}^{G}\mathbb{E}\left[\overline{Y}_{ij}\right] = \texttt{FAR}.$$

Hence, $\text{Bias}(\widehat{\texttt{FAR}}_b^*) = 0$.

– For $\text{Var}^*(\widehat{\texttt{FAR}}_b^*)$, observe that

$$\begin{aligned}
&\mathbb{E}[\text{Var}^*[\widehat{\texttt{FAR}}_b^*]] \\
&= \frac{1}{G^2}\sum_{i=1}^{G}\mathbb{E}\left\{\frac{\left(\sum_{\substack{j=1\\j\neq i}}^{G}\overline{Y}_{ij}\right)^2}{(G-1)^2}\text{Var}^*\left(W_i\right) + \sum_{\substack{k=1\\k\neq i}}^{G}\frac{\sum_{\substack{j=1\\j\neq i}}^{G}\overline{Y}_{ij}\sum_{\substack{l=1\\l\neq k}}^{G}\overline{Y}_{kl}}{(G-1)^2}\text{Cov}^*\left(W_i, W_k\right)\right\} \\
&= \frac{1}{G^2}\sum_{i=1}^{G}\mathbb{E}\left\{\frac{\left(\sum_{\substack{j=1\\j\neq i}}^{G}\overline{Y}_{ij}\right)^2}{(G-1)^2}\frac{G-1}{G} - \sum_{\substack{k=1\\k\neq i}}^{G}\frac{\sum_{\substack{j=1\\j\neq i}}^{G}\overline{Y}_{ij}\sum_{\substack{l=1\\l\neq k}}^{G}\overline{Y}_{kl}}{(G-1)^2}\frac{1}{G}\right\}.
\end{aligned}$$

We thus have

$$\begin{aligned}
&\mathbb{E}[\text{Var}^*[\widehat{\texttt{FAR}}_b^*]] \\
&= \frac{1}{G^2}\sum_{i=1}^{G}\mathbb{E}\left\{\frac{\left(\sum_{\substack{j=1\\j\neq i}}^{G}\overline{Y}_{ij}\right)^2}{(G-1)^2} - \widehat{\texttt{FAR}}^2\right\} \\
&= \frac{1}{G}\mathbb{E}\left\{\frac{\overline{Y}_{12}^2 + (G-2)\overline{Y}_{12}\overline{Y}_{13}}{G-1} - \widehat{\texttt{FAR}}^2\right\}
\end{aligned}$$

$$= \frac{1}{G}\mathbb{E}\Bigg\{ \frac{\overline{Y}_{12}^2 + (G-2)\overline{Y}_{12}\overline{Y}_{13}}{G-1} - \left[\frac{2\overline{Y}_{12}^2 + 4(G-2)\overline{Y}_{12}\overline{Y}_{13} + (G-2)(G-3)\overline{Y}_{12}\overline{Y}_{34}}{G(G-1)} \right] \Bigg\}. \tag{A17}$$

We can rewrite the first of the two terms in (A17) as

$$\frac{1}{G}\mathbb{E}\left\{ \frac{\overline{Y}_{12}^2 + (G-2)\overline{Y}_{12}\overline{Y}_{13}}{G-1} \right\} = \mathrm{Var}(\widehat{\texttt{FAR}}) - \frac{\mathrm{Var}(\overline{Y}_{12}) + 3(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)} + \frac{\texttt{FAR}^2}{G}, \tag{A18}$$

and the second as

$$\frac{1}{G}\mathbb{E}\left\{ \frac{2\overline{Y}_{12}^2 + 4(G-2)\overline{Y}_{12}\overline{Y}_{13} + (G-2)(G-3)\overline{Y}_{12}\overline{Y}_{34}}{G(G-1)} \right\} = \frac{\mathrm{Var}(\widehat{\texttt{FAR}})}{G} + \frac{\texttt{FAR}^2}{G}. \tag{A19}$$

Thus, combining (A18) and (A19) with (A17), we obtain

$$\mathbb{E}[\mathrm{Var}^*[\widehat{\texttt{FAR}}_b^*]] = \frac{G-1}{G}\mathrm{Var}(\widehat{\texttt{FAR}}) - \frac{\mathrm{Var}(\overline{Y}_{12}) + 3(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)}.$$

Therefore, we have

$$\begin{aligned}\mathrm{Bias}(\mathrm{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}_b^*)) &= (G-1)\mathrm{Var}(\widehat{\texttt{FAR}}) - \frac{\mathrm{Var}(\overline{Y}_{12}) + 3(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{(G-1)} - G\mathrm{Var}((\widehat{\texttt{FAR}}) \\ &= -\mathrm{Var}(\widehat{\texttt{FAR}}) - \frac{\mathrm{Var}(\overline{Y}_{12}) + 3(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{(G-1)},\end{aligned}$$

as promised.

This completes the bias derivations for subsets bootstrap estimators.

A.2.2 Proof of Proposition 5 (bias of vertex bootstrap estimators)

Recall from (13) the expression for $\widehat{\texttt{FAR}}_b^*$, the estimator for FAR based on the b-th bootstrap sample using vertex bootstrap:

$$\widehat{\texttt{FAR}}_b^* = \sum_{i,j=1}^{G} W_i \left[\frac{(W_i - 1)\widehat{\texttt{FAR}}\mathbb{1}(i=j)}{G(G-1)} + \frac{W_j \overline{Y}_{ij}\mathbb{1}(i \neq j)}{G(G-1)} \right].$$

- We start with deriving $\mathrm{Bias}(\widehat{\texttt{FAR}}_b^*)$. Note that

$$\mathbb{E}\left\{\mathbb{E}^*[\widehat{\texttt{FAR}}_b^*]\right\} = \frac{1}{G(G-1)} \sum_{i,j=1}^{G} \mathbb{E}\left\{\mathbb{E}^*\left[W_i(W_i-1)\mathbb{1}(i=j)\right]\widehat{\texttt{FAR}} + \overline{Y}_{ij}\mathbb{E}^*\left[W_i W_j \mathbb{1}(i \neq j)\right]\right\}$$

$$
\begin{aligned}
&= \frac{1}{G(G-1)} \sum_{i=1}^{G} \mathbb{E}^* \left[\widehat{\texttt{FAR}} \frac{G-1}{G} + \sum_{\substack{j=1 \\ j \neq i}}^{G} \overline{Y}_{ij} \frac{G-1}{G} \right] \\
&= \frac{1}{G(G-1)} \left[G\texttt{FAR} \frac{G-1}{G} + G(G-1)\texttt{FAR} \frac{G-1}{G} \right] \\
&= \texttt{FAR}.
\end{aligned}
$$

Thus, $\text{Bias}(\widehat{\texttt{FAR}}_b^*) = 0$.

- We next turn to deriving $\text{Bias}(\text{Var}^*(\widehat{\texttt{FAR}}_b^*))$. Let $\overline{Y}_{ij}^*$ denote the observation corresponding to the i-th and j-th identities in the b-th bootstrap sample, where the subscript b is omitted. It is easy to see that

$$
\begin{aligned}
&\mathbb{E}\left\{ \text{Var}^*(\widehat{\texttt{FAR}}_b^*) \right\} \\
&\quad = \frac{1}{G(G-1)} \mathbb{E}\Big\{ 2\text{Var}^*(\overline{Y}_{12}^*) + 4(G-2)\text{Cov}^*\left(\overline{Y}_{12}^*, \overline{Y}_{13}^* \right) + (G-2)(G-3)\text{Cov}^*\left(\overline{Y}_{12}^*, \overline{Y}_{34}^* \right) \Big\}.
\end{aligned}
$$

For the variance term $\text{Var}^*(\overline{Y}_{12}^*)$, we have

$$
\begin{aligned}
\mathbb{E}\left\{ \text{Var}^*(\overline{Y}_{12}^*) \right\} &= \mathbb{E}\left\{ \mathbb{E}^*\left[\overline{Y}_{12}^{2*} \right] - \mathbb{E}^*\left[\overline{Y}_{12}^* \right]^2 \right\} \\
&= \frac{G-1}{G} \mathbb{E}\left[\overline{Y}_{12}^2 \right] + \frac{1}{G} \mathbb{E}\left[\widehat{\texttt{FAR}}^2 \right] - \mathbb{E}\left[\widehat{\texttt{FAR}}^2 \right] \\
&= \frac{G-1}{G} \left\{ \mathbb{E}[\overline{Y}_{12}^2] - \mathbb{E}[\widehat{\texttt{FAR}}^2] \right\} \\
&= \frac{G-1}{G} \left\{ \text{Var}(\overline{Y}_{12}) - \text{Var}(\widehat{\texttt{FAR}}) \right\}. \qquad \text{(A20)}
\end{aligned}
$$

For the covariance term $\text{Cov}^*\left(\overline{Y}_{12}^*, \overline{Y}_{34}^* \right)$, we can show that

$$
\begin{aligned}
\mathbb{E}\left\{ \text{Cov}^*(\overline{Y}_{12}^*, \overline{Y}_{13}^*) \right\} &= \mathbb{E}\left\{ \mathbb{E}^*[\overline{Y}_{12}^* \overline{Y}_{13}^*] - \mathbb{E}^*[\overline{Y}_{12}^*] \mathbb{E}^*[\overline{Y}_{13}^*] \right\} \\
&= \mathbb{E}\left\{ \frac{2G-1}{G^2} \widehat{\texttt{FAR}}^2 + \frac{(G-1)^2}{G^2} \left(\frac{1}{G-1} \overline{Y}_{12}^2 + \frac{G-2}{G-1} \overline{Y}_{12} \overline{Y}_{13} \right) - \widehat{\texttt{FAR}}^2 \right\} \\
&= \mathbb{E}\left\{ \frac{(G-1)^2}{G^2} \left(\frac{1}{G-1} \overline{Y}_{12}^2 + \frac{G-2}{G-1} \overline{Y}_{12} \overline{Y}_{13} \right) - \frac{(G-1)^2}{G^2} \widehat{\texttt{FAR}}^2 \right\} \\
&= \frac{(G-1)^2}{G} \frac{1}{G} \mathbb{E}\left\{ \frac{\overline{Y}_{12}^2}{G-1} + \frac{(G-2)\overline{Y}_{12}\overline{Y}_{13}}{G-1} - \widehat{\texttt{FAR}}^2 \right\}.
\end{aligned}
$$

By following the same derivation as in (A17), we can further show that

$$
\mathbb{E}\left\{ \text{Cov}^*(\overline{Y}_{12}^*, \overline{Y}_{13}^*) \right\} = \frac{(G-1)^2}{G} \left\{ \frac{G-1}{G} \text{Var}(\widehat{\texttt{FAR}}) - \frac{\text{Var}(\overline{Y}_{12}) + 3(G-2)\text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)} \right\}. \qquad \text{(A21)}
$$

Thus, combining (A20) and (A21), together with the fact that $\mathrm{Cov}(\overline{Y}^*_{12}, \overline{Y}^*_{34}) = 0$ (by independence), yields

$$
\begin{aligned}
&\mathbb{E}\left\{\mathrm{Var}^*(\widehat{\mathtt{FAR}}^*_b)\right\} \\
&= \frac{1}{G(G-1)}\left\{2\frac{G-1}{G}\left[\mathrm{Var}(\overline{Y}_{12}) - \mathrm{Var}(\widehat{\mathtt{FAR}})\right]\right. \\
&\quad \left. + \frac{4(G-1)^2(G-2)}{G}\left[\frac{G-1}{G}\mathrm{Var}(\widehat{\mathtt{FAR}}) - \frac{\mathrm{Var}(\overline{Y}_{12}) + 3(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)}\right]\right\} \\
&= \frac{2}{G^2}\left[\mathrm{Var}(\overline{Y}_{12}) - \mathrm{Var}(\widehat{\mathtt{FAR}})\right] \\
&\quad + \frac{4(G-1)(G-2)}{G^2}\left[\frac{G-1}{G}\mathrm{Var}(\widehat{\mathtt{FAR}}) - \frac{\mathrm{Var}(\overline{Y}_{12}) + 3(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)}\right] \\
&= \frac{2}{G^2}\left[\mathrm{Var}(\overline{Y}_{12}) - \mathrm{Var}(\widehat{\mathtt{FAR}})\right] \\
&\quad + \frac{4(G-1)(G-2)}{G^2}\left[\frac{\mathrm{Var}(\overline{Y}_{12}) + (G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)} - \frac{\mathrm{Var}(\widehat{\mathtt{FAR}})}{G}\right] \\
&= \mathrm{Var}(\widehat{\mathtt{FAR}}) - \frac{2\mathrm{Var}(\overline{Y}_{12})}{G^2(G-1)} - \frac{2\mathrm{Var}(\widehat{\mathtt{FAR}})}{G^2} \\
&\quad - \frac{4(3G-2)(G-2)}{G^3(G-1)}\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13}) + \frac{4(G-2)}{G^3}\mathrm{Var}(\overline{Y}_{12}) - \frac{4(G-1)(G-2)}{G^3}\mathrm{Var}(\widehat{\mathtt{FAR}}).
\end{aligned}
$$

We can rearrange the terms to obtain

$$
\begin{aligned}
&\mathrm{Bias}(\mathrm{Var}^*(\widehat{\mathtt{FAR}}^*_b)) \\
&\quad = \mathrm{Var}(\overline{Y}_{12})\left[\frac{4(G-2)}{G^3} + O\left(G^{-3}\right)\right] + \mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})\left[-\frac{28}{G(G-1)} + O(G^{-3})\right].
\end{aligned} \tag{A22}
$$

Up to constants, the expression in (A22) matches with the expression in the statement. This completes the derivations of the bias for the vertex bootstrap estimators.

A.2.3 Proof of Proposition 6 (bias of double-or-nothing bootstrap estimators)

Recall from (14) the expression for $\widehat{\mathtt{FRR}}^*_b$ and $\widehat{\mathtt{FAR}}^*_b$, the estimator for FRR and FAR based on the b-th bootstrap sample using double-or-nothing bootstrap:

$$
\widehat{\mathtt{FRR}}^*_b = \frac{\sum_{i=1}^G W_i \overline{Y}_{ii}}{\sum_{i=1}^G W_i}, \quad \text{and} \quad \widehat{\mathtt{FAR}}^*_b = \frac{\sum_{\substack{i,j=1 \\ j\neq i}}^G W_i W_j \overline{Y}_{ij}}{\sum_{\substack{i,j=1 \\ j\neq i}}^G W_i W_j},
$$

where $\mathbb{E}[W_i] = 1$, $\mathrm{Var}(W_i) = \tau$, and W_i is independent of W_j whenever $i \neq j$ for $i, j \in \mathcal{G}$. The double-or-nothing bootstrap falls in this framework when $\tau = 1$.

- It it straightforward to show that $\mathrm{Bias}(\widehat{\mathtt{FRR}}^*_b) = 0$. Next, we examine $\mathrm{Bias}(\mathrm{Var}^*(\widehat{\mathtt{FRR}}^*_b))$. Let $T^* = \sum_{i=1}^G W_i \overline{Y}_{ii}$ and $N^* = \sum_{i=1}^G W_i$. Through an application of the Delta method, we obtain

$$
\mathbb{E}\left\{\mathrm{Var}^*(\widehat{\mathtt{FRR}}^*_b)\right\} = \frac{1}{G^2}\mathbb{E}\left\{\mathrm{Var}^*(T^*) - 2\widehat{\mathtt{FRR}}\mathrm{Cov}^*(T^*, N^*) + \widehat{\mathtt{FRR}}^2\mathrm{Var}^*(N^*)\right\},
$$

where

$$\mathbb{E}[\mathrm{Var}^*(T^*)] = \mathbb{E}\left[\sum_{i=1}^{G} \overline{Y}_{ii}^2 \tau\right] = \tau G \mathbb{E}[\overline{Y}_{11}^2],$$
$$\mathbb{E}[\widehat{\mathtt{FRR}}\mathrm{Cov}^*(T^*, N^*)] = \mathbb{E}[\widehat{\mathtt{FRR}}^2 \mathrm{Var}^*(N^*)] = G\tau \mathbb{E}[\widehat{\mathtt{FRR}}^2] = \tau\left[\mathbb{E}[\overline{Y}_{11}^2] + \mathtt{FRR}^2(G-1)\right].$$

Hence, we have

$$\mathbb{E}[\mathrm{Var}^*(\widehat{\mathtt{FRR}}_b^*)] = \frac{G-1}{G}\tau \mathrm{Var}(\widehat{\mathtt{FRR}}).$$

Taking $\tau = 1$ yields the result.

- With respect to the FAR, let $T^* = \sum_{i=1}^{G}\sum_{\substack{j=1 \\ j\neq i}}^{G} W_i W_j \overline{Y}_{ij}$ and $N^* = \sum_{i=1}^{G}\sum_{\substack{j=1 \\ j\neq i}}^{G} W_i W_j$. Again, it is easy to see that $\mathrm{Bias}(\widehat{\mathtt{FAR}}_b^*) = 0$. An application of the Delta method yields

$$\mathbb{E}\left\{\mathrm{Var}^*(\widehat{\mathtt{FAR}}_b^*)\right\} = \frac{1}{G^2(G-1)^2}\mathbb{E}\left\{\mathrm{Var}^*(T^*) - 2\widehat{\mathtt{FAR}}\mathrm{Cov}^*(T^*, N^*) + \widehat{\mathtt{FAR}}^2 \mathrm{Var}^*(N^*)\right\},$$

where

$$\begin{aligned}
\mathbb{E}[\mathrm{Var}^*(T^*)] &= G(G-1)\mathbb{E}\left\{2\mathrm{Var}^*(\overline{Y}_{12}W_1W_2) + 4(G-2)\mathrm{Cov}^*(\overline{Y}_{12}W_1W_2, \overline{Y}_{13}W_1W_3)\right\} \\
&= G(G-1)\mathbb{E}\left\{2\overline{Y}_{12}(\mathbb{E}^*[W_1^2]\mathbb{E}^*[W_2^2] - 1) + 4(G-2)\overline{Y}_{12}\overline{Y}_{13}(\mathbb{E}^*[W_1^2] - 1)\right\} \\
&= G(G-1)[2\tau(\tau+2)\mathbb{E}[\overline{Y}_{12}^2] + 4(G-2)\tau\mathbb{E}[\overline{Y}_{12}\overline{Y}_{13}]],
\end{aligned}$$

and

$$\mathbb{E}\left[\widehat{\mathtt{FAR}}\mathrm{Cov}^*(T^*, N^*)\right] = \mathbb{E}[\widehat{\mathtt{FAR}}^2 \mathrm{Var}^*(N^*)] = G(G-1)[2\tau(\tau+2) + 4(G-2)\tau]\mathbb{E}[\widehat{\mathtt{FAR}}^2].$$

It then follows that

$$\begin{aligned}
\mathbb{E}[\mathrm{Var}^*(\widehat{\mathtt{FAR}}_b^*)] &= \frac{1}{G(G-1)}\left[2\tau(\tau+2)\mathbb{E}[\overline{Y}_{12}^2 - \widehat{\mathtt{FAR}}^2] + 4(G-2)\tau\mathbb{E}[\overline{Y}_{12}\overline{Y}_{13} - \widehat{\mathtt{FAR}}^2]\right] \\
&= \frac{1}{G(G-1)}\left[2\tau(\tau+2)\mathrm{Var}(\overline{Y}_{12}) + 4(G-2)\tau\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})\right] \\
&\quad - \frac{2\tau(\tau+2G-2)}{G(G-1)}\frac{2\mathrm{Var}(\overline{Y}_{12}) + 4(G-2)\mathrm{Cov}(\overline{Y}_{12}, \overline{Y}_{13})}{G(G-1)}.
\end{aligned}$$

Choosing $\tau = 1$ yields

$$\mathrm{Bias}(\mathrm{Var}^*(\widehat{\mathtt{FAR}}_b^*)) = -\mathrm{Var}(\widehat{\mathtt{FAR}})\frac{2(2G-1)}{G(G-1)} + \frac{4\mathrm{Var}(\overline{Y}_{12})}{G(G-1)}.$$

This completes the bias derivations for the double-or-nothing bootstrap estimators.

A.3 Proofs for the unbalanced setting in Section 5.1

A.3.1 Proof of Proposition 7 (consistency of bootstrap estimators for FRR)

We separate the proof into the consistency of subsets and vertex bootstrap, and that of double-or-nothing bootstrap below.

- **Consistency of subsets and vertex bootstrap estimators.** The resampling performed by these two bootstrap types for `FRR` computations is analogous, thus we investigate the consistency of both types altogether. By applying the Delta method, we obtain

$$\begin{aligned}
&\mathrm{Var}^*(\widehat{\mathtt{FRR}}_b^*) \\
&= \frac{\sum_{i=1}^G \widetilde{M}_i^2 \overline{Y}_{ii}^2 - \left(\frac{1}{\sqrt{G}} \sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)^2}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^2} \\
&\quad - 2\frac{\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)\left[\sum_{i=1}^G \widetilde{M}_i^2 \overline{Y}_{ii} - (\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii})(\sum_{i=1}^G \widetilde{M}_i/G)\right]}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^3} \\
&\quad + \frac{\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)^2 \left[\sum_{i=1}^G \widetilde{M}_i^2 - \left(\frac{1}{\sqrt{G}} \sum_{i=1}^G \widetilde{M}_i\right)^2\right]}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^4}.
\end{aligned}$$

  Since M_i is finite for any $i \in \mathcal{G}$, we can apply the weak law of large numbers and the continuous mapping theorem to obtain the following convergences in probability as $G \to \infty$:

$$\begin{aligned}
G\frac{\sum_{i=1}^G \widetilde{M}_i^2 \overline{Y}_{ii}^2 - \left(\frac{1}{\sqrt{G}} \sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)^2}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^2} &\xrightarrow{\mathrm{p}} \frac{\mathrm{Var}(\widetilde{M}_1 \overline{Y}_{11})}{\mathbb{E}[\widetilde{M}_1]^2}, \\
G\frac{\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)\sum_{i=1}^G \left(\widetilde{M}_i^2 \overline{Y}_{ii} - (\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii})(\sum_{i=1}^G \widetilde{M}_i/G)\right)}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^3} &\xrightarrow{\mathrm{p}} \frac{\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]\mathrm{Cov}(\widetilde{M}_1 \overline{Y}_{11}, \widetilde{M}_1)}{\mathbb{E}[\widetilde{M}_1]^3}, \\
G\frac{\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)^2 \left[\sum_{i=1}^G \widetilde{M}_i^2 - \left(\frac{1}{\sqrt{G}} \sum_{i=1}^G \widetilde{M}_i\right)^2\right]}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^4} &\xrightarrow{\mathrm{p}} \frac{\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]^2 \mathrm{Var}(\widetilde{M}_1)}{\mathbb{E}[\widetilde{M}_1]^4}.
\end{aligned}$$

  It then follows that

$$\mathrm{Var}^*(\sqrt{G}\widehat{\mathtt{FRR}}_b^*) \xrightarrow{\mathrm{p}} \frac{\mathrm{Var}(\widetilde{M}_1 \overline{Y}_{11})}{\mathbb{E}[\widetilde{M}_1]^2} - 2\frac{\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]\mathrm{Cov}(\widetilde{M}_1 \overline{Y}_{11}, \widetilde{M}_1)}{\mathbb{E}[\widetilde{M}_1]^3} + \frac{\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]^2 \mathrm{Var}(\widetilde{M}_1)}{\mathbb{E}[\widetilde{M}_1]^4}.$$

  This completes the proof for the consistency of the subsets and vertex bootstrap estimators.
- **Consistency of double-or-nothing bootstrap estimator.** Assume that $\mathbb{E}[W_i] = 1$ and $\mathrm{Var}(W_i) = \tau$. In addition, let $W_i \perp W_j$ whenever $i \neq j$ for $i, j \in \mathcal{G}$. The double-or-nothing bootstrap is obtained by taking $\tau = 1$. By applying the Delta method, we obtain

$$\mathrm{Var}^*(\widehat{\mathtt{FRR}}_b^*) = \tau\frac{\sum_{i=1}^G \widetilde{M}_i^2 \overline{Y}_{ii}^2}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^2} - 2\tau\frac{\left(\sum_{i=1}^G \widetilde{M}_i^2 \overline{Y}_{ii}\right)\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^3} + \tau\frac{\left(\sum_{i=1}^G \widetilde{M}_i^2\right)\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)^2}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^4}.$$

Since $\widetilde{M}_i$ is finite, we can apply the weak law of large numbers and the continuous mapping theorem to obtain, as $G \to \infty$,

$$\tau G \frac{\sum_{i=1}^G \widetilde{M}_i^2 \overline{Y}_{ii}^2}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^2} \xrightarrow{\text{p}} \tau \frac{\mathbb{E}[\widetilde{M}_1^2 \overline{Y}_{11}^2]}{\mathbb{E}[\widetilde{M}_1]^2},$$

$$2G\tau \frac{\left(\sum_{i=1}^G \widetilde{M}_i^2 \overline{Y}_{ii}\right)\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^3} \xrightarrow{\text{p}} 2\tau \frac{\mathbb{E}[\widetilde{M}_1^2 \overline{Y}_{11}]\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]}{\mathbb{E}[\widetilde{M}_1]^3},$$

$$\tau G \frac{\left(\sum_{i=1}^G \widetilde{M}_i^2\right)\left(\sum_{i=1}^G \widetilde{M}_i \overline{Y}_{ii}\right)^2}{\left(\sum_{i=1}^G \widetilde{M}_i\right)^4} \xrightarrow{\text{p}} \tau \frac{\mathbb{E}[\widetilde{M}_1^2]\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]^2}{\mathbb{E}[\widetilde{M}_1]^4}.$$

Putting everything together, as $G \to \infty$, we have that

$$\begin{aligned}
\text{Var}^*(\sqrt{G}\widehat{\texttt{FRR}}_b^*) &\xrightarrow{\text{p}} \tau \frac{\mathbb{E}[\widetilde{M}_1^2 \overline{Y}_{11}^2]}{\mathbb{E}[\widetilde{M}_1]^2} - 2\tau \frac{\mathbb{E}[\widetilde{M}_1^2 \overline{Y}_{11}]\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]}{\mathbb{E}[\widetilde{M}_1]^3} + \tau \frac{\mathbb{E}[\widetilde{M}_1^2]\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]^2}{\mathbb{E}[\widetilde{M}_1]^4} \\
&= \tau \frac{\text{Var}(\widetilde{M}_1 \overline{Y}_{11})}{\mathbb{E}[\widetilde{M}_1]^2} - 2\tau \frac{\text{Cov}(\widetilde{M}_1 \overline{Y}_{11}, \widetilde{M}_1)\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]}{\mathbb{E}[\widetilde{M}_1]^3} + \tau \frac{\text{Var}(\widetilde{M}_1)\mathbb{E}[\widetilde{M}_1 \overline{Y}_{11}]^2}{\mathbb{E}[\widetilde{M}_1]^4}.
\end{aligned}$$

Choosing $\tau = 1$ yields the desired result. This completes the proof of the consistency of the double-or-nothing bootstrap estimator.

Appendix B Protocol design

Many vision and audio datasets comprise hundreds of thousands of instances, making it computationally infeasible to estimate FRR and FAR on all the data. In such cases, the researcher has to decide on which instance pairs their computational resources (i.e., budget) should be spent on. Since different combinations of pairwise comparisons between instances may lead to different estimates of model accuracy, dataset designers attach protocols specifying which comparisons to consider in computations. Consequently, a natural question is then: *For a given budget, which instance pairs offer the lowest variance estimate of model accuracy?*

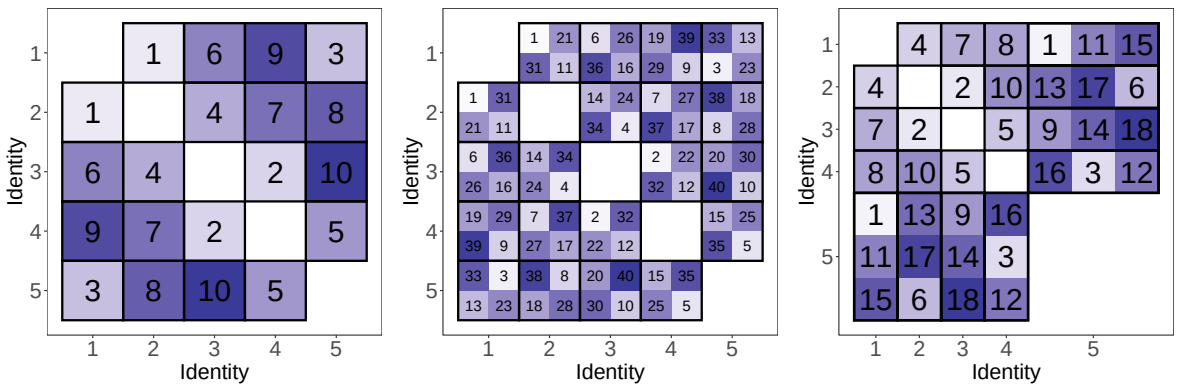

Fig. B1 Protocol design for selecting identities and instances in FAR computations. The selection is based on Algorithm 1. The number in each cell represents the iteration at which the given pair of identities and instances is chosen. Left panel: balanced setting with one instance for each of the five identities. Middle panel: balanced setting with two instances for each of the five identities. Right panel: unbalanced setting with the first four and the fifth identities have one and three instances respectively. Note that in all scenarios we iterate through all identities before selecting the same identity again. However, while in the balanced setting, we pick the combination of identities 1–2 first, in the unbalanced setting we choose identities 1–5 first as the latter has more observations.

Based on our theoretical analysis in Section 4 (and demonstrated in the empirical results in Section 6), it is clear that the dependence structure induced by the comparisons can significantly impact the coverage of the confidence intervals. This realization naturally leads to a strategy for protocol design to try to maintain the independence structure between comparisons. For simplicity, consider the computation of FAR on a sample where each identity has only one instance. Assume that a budget of $B \leqslant G(G-1)$ comparisons is available, and let $B = \sum_{i=1}^{G} \sum_{j \neq i} b_{ij}$ with $b_{ij} = b_{ji} = 1$ when $\overline{Y}_{ij}$ enters FAR computations and 0 otherwise. Minimizing the variance of the estimated FAR under budget constraints boils down to solving the following problem:

$$\underset{b_{12},\ldots,b_{G(G-1)}}{\arg\min} \sum_{\substack{i,j,k=1 \\ j\neq i \\ k\neq i,j}}^{G} b_{ij} b_{ik} \text{ s.t. } \sum_{\substack{i,j=1 \\ i\neq j}}^{G} b_{ij} = B. \tag{B23}$$

When $B \leqslant \lfloor G/2 \rfloor$, one can choose instance pairs that are independent, e.g., $\overline{Y}_{12}, \overline{Y}_{34}$, etc. When $B > \lfloor G/2 \rfloor$, the objective in (B23) is minimized when the comparisons share as few instances as possible with each other. An approach to choose the terms to include in the FAR computations is as follows: At each of the B iterations, select the observation that minimizes the objective evaluated using the allocation resulting from the previous iteration.

Algorithm 1 outlines the proposed approach for selecting the combinations of identities to be included in the FAR estimation for the balanced setting. We start by creating all possible combinations of identities

Algorithm 1 Protocol Design Strategy to Select Identities Combinations (IDs) for `FAR` Estimation in the Balanced Setting

Input: `budget` > 0, `data` $= \{$`id` : `instances`$\}$

1: Initialize `IDCombinations` to empty and `IDVisits` to priority queue for number of ID visits with IDs present in `data`
2: **while** `budget` > 0 **do**
3: Retrieve candidate IDs with the lowest number of visits from `IDVisits`
4: Sort candidate IDs in decreasing order according to the number of instances
5: Iterate through candidate IDs and find the first unused pair
6: Update `IDCombinations`, `IDVisits`, and `budget`
7: **end while**

Output: Set of ID combinations

from which we will draw the instances to be considered. At each iteration, we use a priority queue to retrieve the identity candidates with the lowest number of visits. These candidates are sorted to ensure that those with a larger number of instances are visited first, which helps minimize the number of times a given instance will be reused in the estimation. Note that in the balanced setting, the last step is not necessary. If the total budget exceeds $\lfloor G(G-1)/2 \rfloor$, we can iterate through the combinations yielded by Algorithm 1. Once the combinations of identities are available, we follow a similar strategy for selecting the pairs of instances within each pair of identities. Figure B1 describes three examples of protocols resulting from applying this strategy. For `FRR` estimation, we follow a similar idea. We first iterate through the identities starting with those with the largest number of instances. We then use Algorithm 1 to select the comparisons within each identity.

Finally, a brief note about computations of error metrics and the associated uncertainties on massive datasets under computational constraints. The proposed strategy for protocol design can be applied to handle estimation in these settings. This involves selecting a fixed number of instance pairs using the protocol design, estimating the error metrics on these pairs, and then using Wilson or bootstrap methods to obtain confidence intervals. By following this approach, one can obtain reliable estimates of error metrics and their uncertainties while minimizing computational costs.

Appendix C Additional numerical experiments

In this section, we present additional experimental details and results, supplementing those in presented in Sections 1 and 6.

C.1 Analysis of interval widths

The discussion in the main paper has focused on interval coverage and has only briefly mentioned width. In our experiments, we found that methods that yield intervals with higher coverage also generally presented larger widths, as we would expect in the case of statistics that are asymptotically normal (see Proposition 1). Figure C2 shows the relationship between estimated coverage, average width, and nominal coverage for `FAR` intervals with $G = 50$ and $M = 5$ using the setup described in Section 6.1 (see Figure 3 for estimated vs. nominal coverage). In the case of $\texttt{FAR} = 10^{-3}, 10^{-4}$, a given estimated coverage corresponds to the same interval width across all methods. This indicates that recalibrating the nominal coverage (e.g., increasing the nominal level $1-\alpha$ for the subsets or two-level bootstrap to achieve intervals with coverage $1 - \alpha_{\text{target}}$) for any of the methods will not yield intervals with the target coverage but with a smaller width.

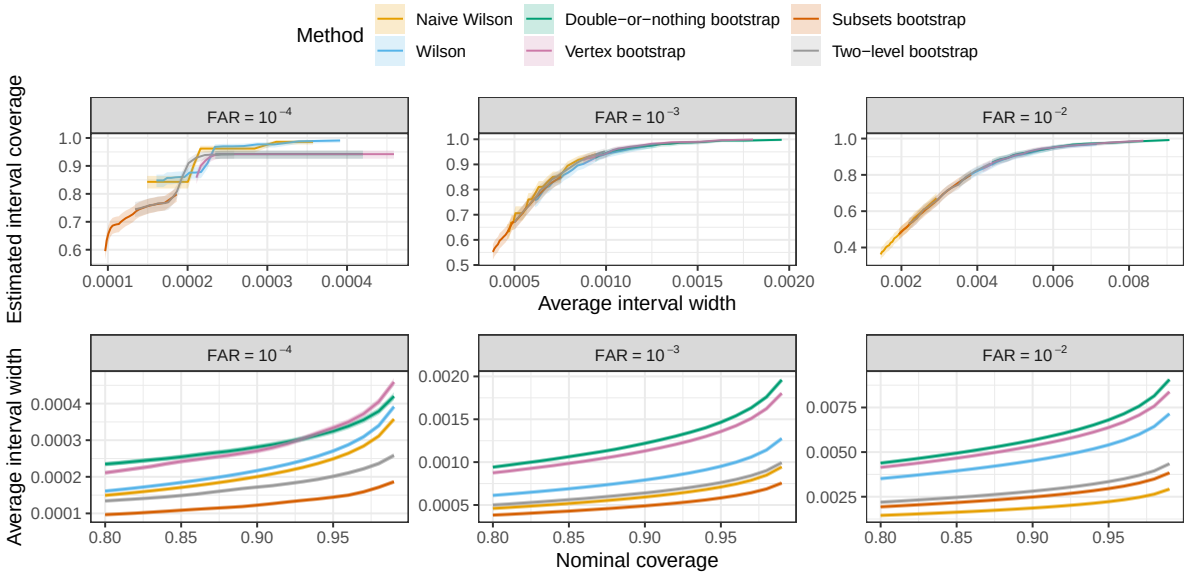

Fig. C2 Estimated coverage versus average width (top) and average width vs. nominal coverage (bottom) for `FAR` intervals on synthetic data. Data contain $G = 50$ with $M = 5$ instances each. Colored lines and shaded regions indicate estimated coverage and corresponding 95% naive Wilson or Wald confidence intervals for the different methods.

C.2 Variance estimation accuracy versus sample size

One natural question is how large the sample should be to obtain an accurate estimate of the variances of `FRR` and `FAR`, and for Wilson intervals to achieve close-to-nominal coverage. As we have seen in Proposition 2, asymptotically the reviewed variance estimators converge to the true parameters. Their behavior in case of a few observations in the sample may be less clear. However, in Figure 4 we have seen that Wilson intervals achieve coverage close to nominal for any number of identities. This observation suggests that the variances of `FRR` and `FAR` are close to the true variances even when only a few identities are present in the data. This is to show in the case of `FRR`, for which one can obtain finite-sample guarantees via standard arguments. For the estimator of the `FAR` variance, the derivation of the limiting distribution is more complex due to data dependence. For this reason, we resort to simulation and in Figure C3 we show how the mean squared error (MSE) of the estimator of the `FAR` variance in (9) varies with the number of identities G present in synthetic data. The results show that the MSE of the variance estimator greatly decreases with the number of identities in the data (at rate $\sqrt{G}$, consistently with Proposition 2)), and is small even when a limited number of identities is available in the data.

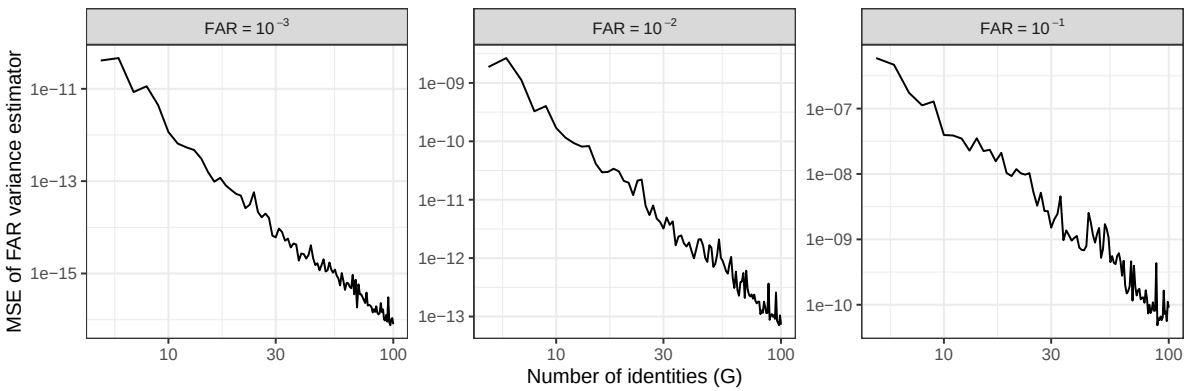

Fig. C3 Mean squared error (MSE) of `FAR` variance estimator on synthetic data. Data contain varying number of identities G with $M = 5$ instances each. The lines correspond to the MSE of `FAR` variance estimator as a function of G.

C.3 Pointwise intervals for the ROC

In this section, we evaluate the coverage of pointwise confidence intervals for the ROC on the MORPH dataset. The experimental setup follows the description of Section 6. The vertex bootstrap performs similarly to the double-or-nothing procedure and thus, for the sake of simplifying the presentation of the results, it is omitted from the discussion. Figure C4 shows estimated coverage as a function of nominal coverage for the reviewed methods at different levels of `FAR`. Consistently with the discussion of Section 5.2, we observe that Wilson intervals achieve coverage that is higher than nominal across all `FAR` levels. While the overcoverage may be expected for low values of `FRR` (e.g., see the results in Figure 3), the overestimation is present albeit it is lower for larger values of `FRR`. For low `FRR`, we also observe that the version of the double-or-nothing bootstraps that employ ROC curves smoothed using log-normal distributions to model the scores perform better than their counterparts. This is suggestive of the benefits of imposing smoothness assumptions. When `FRR` is large enough, the bootstraps perform similarly.

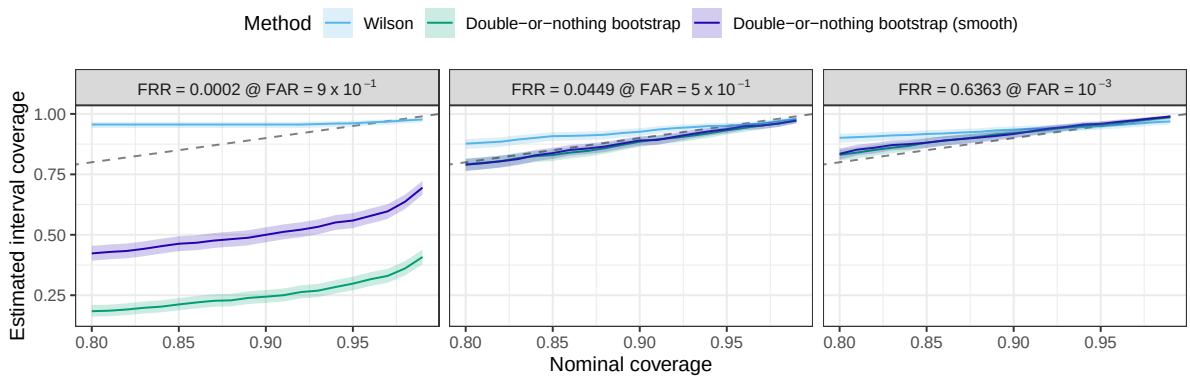

Fig. C4 Estimated coverage versus nominal coverage of confidence intervals for `FRR@FAR` on the MORPH dataset. Samples were generated by resampling $G = 50$ identities from the original dataset without replacement.

C.4 Experiments on diverse data types

Our theoretical analysis applies to any 1:1 matching task. Here we explore its properties empirically on real data, with data types and tasks beyond 1:1 face verification. In particular, we investigate the performance of our methods in the following tasks:

- **1:1 object verification.** Matching images from a randomly sampled subset of the iNat2021 dataset [47]. The iNat2021 dataset is an image collection specifically curated for species recognition, featuring over 10,000 different species. The matching task is to recognize whether the animals in two different images belong to the same species. For this purpose, we extract feature representations obtained via CLIP [40].
- **1:1 speaker verification.** We use a large dataset of voice recordings corresponding to multiple speakers. For the speaker verification task, we extract the embeddings of the audio recordings using an `ECAPA-TDNN` pre-trained model [16, 41].
- **1:1 topic verification.** We aim to detect whether two text paragraphs are related to the same topic. For this purpose, we use a subset of the Amazon review dataset [34], comprising product information and corresponding Amazon reviews. Specifically, we focus on identifying if two reviews pertain to the same product. Classification is performed using text embeddings generated by `BAAI/bge-smal-en-v1.5` [52].

For all datasets, our experimental framework follows the same setup of Section 6.2, using $G = 50$ identities. Figure C5 shows how the coverage of the confidence intervals for $\texttt{FRR} = 10^{-1}$ and $\texttt{FAR} = 10^{-2}$, constructed using the reviewed methods, varies with nominal coverage. The results are consistent with our empirical findings of Section 6: For `FRR`, all methods other than naive Wilson tend to cover approximately at the right level. For `FAR`, the Wilson intervals, as well as the vertex and double-or-nothing bootstrap intervals, achieve coverage close to nominal. The other methods severely undercover.

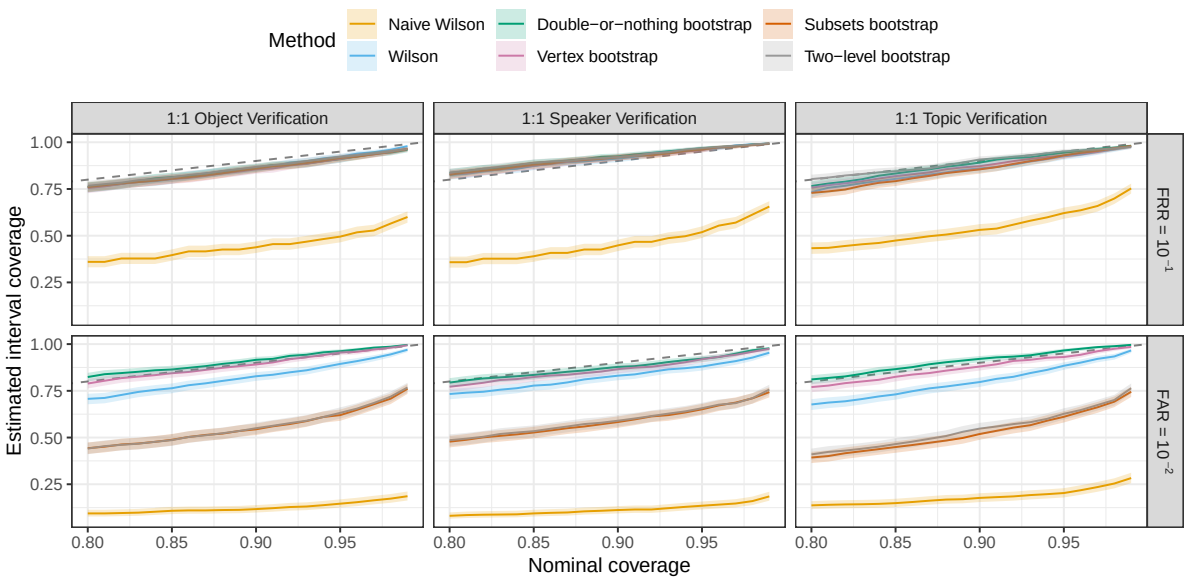

Fig. C5 Estimated interval coverage vs. nominal coverage for $\texttt{FRR} = 10^{-1}$ (top) and $\texttt{FAR} = 10^{-2}$ (bottom) on 1:1 object, speaker, and topic verification tasks respectively. Samples were generated by resampling $G = 50$ identities without replacement from the original dataset.

C.5 Power analyses for 1:1 matching tasks

Does my model meet my error rate target? Is my model's accuracy better than its competitor's? Are my confidence intervals too wide to know? If so, how much more test data do I need to collect to reach a safe conclusion? These questions may be answered through a *power analysis*. In this section, we investigate how the width of the confidence intervals for FRR and FAR varies with the number of identities G and the number of images per identity M. We also examine the statistical power of a one-sided z-test, which can be used to test whether the error rate is below or above a given target.

Experimental setup

The variance of FRR and FAR estimates maybe computed using equations (3) and (4) from the variance of the FRR estimate for one identity $\text{Var}(\overline{Y}_{11})$, the variance of the FAR estimate between the images of two identities $\text{Var}(\overline{Y}_{12})$, and the covariance between the FAR estimates of the instances of one identities with those of two other identities $\text{Cov}(\overline{Y}_{12}, \overline{Y}_{13})$, and the number of identities G. These parameters are data-dependent, and we estimate them on the Morph dataset following a setup similar to the one in Section 6.2. For the estimation, we create a subset of the dataset where each identity contains exactly $M \in \{5, 10, 15\}$ images (instead of the original $M = 5$ as in Section 6.2). We exclude identities with fewer images and randomly sample M images from those with more than M images. We then compute the variance and covariance components at different FRR and FAR values using the dataset. Finally, we construct symmetric Wald confidence intervals for FRR and FAR.

Size of confidence intervals versus sample size

The half-width of the 90% symmetric Wald confidence intervals for FRR and FAR is shown in Figure C6 as a function of G and M. First, note that, for a fixed pair of G and M, the FAR intervals are smaller than the FRR intervals. Consequently, the requirements are primarily driven by FRR. Another straightforward, albeit important, observation is that, for fixed M, *increasing the number of identities G by 10x will lead to* FRR *intervals that are approximately* $1/\sqrt{10} \approx 1/3$ *the original size.* Additionally, we generally have

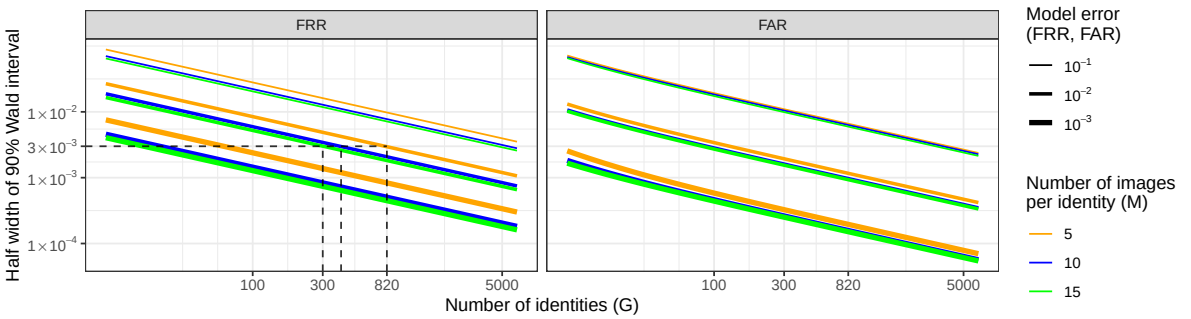

Fig. C6 Size of 90% Wald confidence interval for FRR and FAR estimated on the Morph dataset. The half-width of 90% Wald confidence intervals is shown for a given FRR or FAR level across different numbers of images per identity (M) and number of identities in the data (G). This plot can be used by practitioners to select appropriate values for G and M. First, note that FRR $\gg$ FAR and thus the numbers G and M can be chosen based on the FRR confidence interval half-width. Second, the half-width of the 90% Wald interval should be less than a certain fraction of the model's error. For example, if the model's error is 1% FRR, a practitioner may want to select a half-interval width of 3×10^{-2}. To achieve this, one can consider the intersection points of the 3×10^{-2} line with the $M = 5, 10, 15$ lines, which correspond to $G \approx 820, 400, 300$, respectively.

$\mathrm{Var}(\overline{Y}_{11}) = O(1/M)$.[4] Thus, as we see in the plot, *collecting more than* $M = 10 - 15$ *samples for each identity yields diminishing returns.*

Based on these observations and our theoretical results, we can derive the following rule of thumb for selecting the sample size required to obtain FRR intervals of a given magnitude: The size of the confidence interval will be approximately $\sqrt{\frac{\mathtt{FRR}}{10G}}$ when $M = 10$. This guideline suggests that an FRR estimate of $\mathtt{FRR} = 10^{-4}$ obtained from a dataset and $G = 10^3$ would result in confidence intervals whose half width is of magnitude 10^{-4}, which is the same magnitude as the estimate itself.

Statistical power of one-sided z-test

We also analyze the statistical power of a one-sided z-test with a significance level of 0.05. Specifically, we test the null hypothesis $H_0 : \mathtt{FRR} \leqslant$ target (e.g., target $= 10^{-3}$) against the alternative hypothesis $H_1 : \mathtt{FRR} >$ target. Our goal is to ensure that the probability of mistakenly rejecting the null hypothesis is small, i.e., $\mathbb{P}(\{H_0 \text{ is rejected}\}|\{H_0 \text{ is true}\}) \leqslant \alpha = 0.05$. The power of the test corresponds to $\mathbb{P}(\{H_0 \text{ is rejected}\}|\{H_1 \text{ is true}\})$, which is the likelihood that we would conclude that FRR is larger than the target when FRR is actually equal to x (e.g., $x = 1.5 \times$ target). Figure C7 illustrates the power of this test across a similar range of values as shown in the visualization of the confidence intervals in Figure C6. As expected, we observe that when the effect size is small, such as $\mathtt{FRR} = 1.2 \times$ target where target $\in \{10^{-2}, 10^{-3}\}$, the power remains low for most values of G and M.

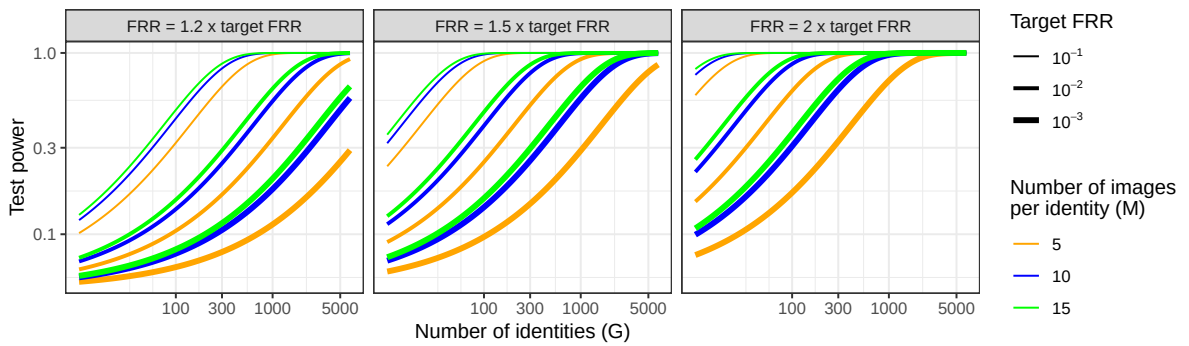

Fig. C7 Statistical power of one-sided z-test for FRR and FAR estimated on the Morph dataset. The statistical power refers to a test with a significance level of 0.05, where the null hypothesis is $H_0 : \mathtt{FRR} \leqslant$ target and the alternative hypothesis is $H_1 : \mathtt{FRR} >$ target for target $\in \{10^{-3}, 10^{-2}, 10^{-1}\}$, assuming different effect sizes of $\mathtt{FRR} = 1.2, 1.5, 2 \times$ target.

Potential limitations of the analysis

We conclude by highlighting a limitation of the power analysis we just presented. When designing a dataset collection, it is important to consider the variability of images within each identity. In the case of the Morph dataset, although the images have the same background, the appearance of the individuals can vary significantly due to differences in the time of capture (sometimes years apart). We have conducted similar computations on datasets with both ID-style photos and in-the-wild imagery, as well as using different facial recognition models, and obtained consistent results, suggesting that our findings can be generalized to a wide variety of datasets. However, it is crucial to note that our estimates are valid only

[4] This holds when the pictures of the same identity are different enough. In case of the pictures being very dissimilar (e.g., when the appearance of the individual changes substantially), then $\mathrm{Cov}(Y_{(1,1),(1,2)}, Y_{(1,1),(1,3)}) \approx 0$ and consequently $\mathrm{Var}(\overline{Y}_{11}) = O(1/M^2)$, i.e., the variance decreases more quickly. Similarly to our discussion of FAR for fig. 4, this phenomenon also occurs often when FRR is small relative to sample size. It follows our rule of thumb may be conservative for these cases. When all the pictures of the same identity are virtually identical, $\mathrm{Cov}((1,1),(1,2), Y_{(1,3),(1,4)}) \approx \mathrm{Var}(Y_{(1,1),(1,2)}) = \mathtt{FRR}(1 - \mathtt{FRR})$ (i.e., one of our key assumptions is broken), therefore $\mathrm{Var}(\overline{Y}_{11}) = \mathtt{FRR}(1 - \mathtt{FRR})$. In this setting, considering either one or all pictures from the individual in the estimator will lead to the same variance.

if study participants submit a diverse set of photos that represent the allowable variations within the specific use case. For example, in the case of passport photos, the viewpoint should always be frontal, but there can be variations in lighting and facial expression (from neutral to smiling). On the other hand, for candid or vacation pictures, there can be more variations in viewpoint, lighting, resolution, and facial expression, and it is important to include such diversity in the test photos. If the submitted photos are too similar, the estimation of `FRR` and `FAR` can suffer from large uncertainty.