Are Your Fairness Metrics Accurate? A Semi-Supervised Approach to Improving Fairness Estimates Under Sample Selection Bias

M. Clara De Paolis Kaluza*
depaoliskaluza.m@northeastern.edu
Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts, USA

Thulasi Tholeti*
t.tholeti@northeastern.edu
Institute for Experiential AI
Northeastern University
Boston, Massachusetts, USA

Yile Chen
yilechen@uw.edu
Department of Biomedical Informatics and Medical Education
University of Washington
Seattle, Washington, USA

Ricardo Baeza-Yates
rbaeza@acm.org
AI Institute
Barcelona Supercomputing Center
Barcelona, Catalonia, Spain

Predrag Radivojac
predrag@northeastern.edu
Khoury College of Computer Sciences
Northeastern University
Boston, Massachusetts, USA

Shantanu Jain†
sh.jain@northeastern.edu
Institute for Experiential AI
Northeastern University
Boston, Massachusetts, USA

Abstract

A key challenge impeding the widespread deployment of machine learning is overcoming the impact of statistical biases in the data. Models trained on unrepresentative data can perform worse than anticipated and differentially affect cross-sections of the population. Therefore, evaluating and vetting models based on an appropriate notion of fairness is often indispensable, making accurate estimation of fairness metrics a critical step to safeguard against deployment of unfair algorithms. It is often assumed that a fairness metric computed from the observed data is accurate. However, in presence of selection bias, also referred to as distributional shifts, fairness metric estimates too can have systematic application-specific errors.

In this work, we demonstrate this phenomenon and, relying on access to an unbiased unlabeled data, derive a semi-supervised approach to mitigate estimation errors emerging from the biased labeled data. Specifically, we introduce a novel selection bias model called "sub-class-conditional invariance" (SCC-invariance), that offers a flexible framework to effectively capture distributional shifts in the real-world data, particularly compared to traditional models such as label shift and covariate shift. Assuming a finite Gaussian mixture form for each class-conditional distribution, we then derive an Expectation-Maximization algorithm to estimate model parameters and correction weights necessary for computing unbiased estimates. We focus on three widely used fairness metrics—equal opportunity, predictive equality, and predictive parity—and demonstrate the effectiveness of our approach in improving their estimates on synthetic data. Finally, we apply our bias mitigation approach to clinical genetics and study the fairness of pathogenicity predictors across ancestral groups.

*Both authors contributed equally to this work.
†Corresponding author

KDD '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1454-2/2025/08
https://doi.org/10.1145/3711896.3736846

CCS Concepts

• **Computing methodologies** → **Semi-supervised learning settings**; • **Applied computing** → *Computational biology*; Law, social and behavioral sciences.

Keywords

Fairness; sample selection bias; bias mitigation; importance sampling; gaussian mixture model; expectation maximization

ACM Reference Format:
M. Clara De Paolis Kaluza, Thulasi Tholeti, Yile Chen, Ricardo Baeza-Yates, Predrag Radivojac, and Shantanu Jain. 2025. Are Your Fairness Metrics Accurate? A Semi-Supervised Approach to Improving Fairness Estimates Under Sample Selection Bias . In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '25), August 3–7, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3711896.3736846

KDD Availability Link:
The source code of this paper has been made publicly available at https://doi.org/10.5281/zenodo.15572033.

1 Introduction

In the era of artificial intelligence (AI), public and private institutions are increasingly leveraging machine learning algorithms to automate decisions that profoundly impact individuals' lives. These decisions span high-stakes applications in criminal justice, hiring, lending, medicine, advertising, recommendation systems, and exposure to personalized content on social media [3, 17, 29, 31], shaping access to resources, opportunities, and rights. In spite of the immense incentives of AI in cost-savings terms and consistency of data-driven decisions, some of these advantages are not realized in practice. A machine learning algorithm is only as good as the data it is trained on and the fairness embedded in the optimization function [1]. A seemingly neutral algorithm, when trained or evaluated on unrepresentative data, may replicate or exacerbate the problems of data collection, posing a significant challenge in eliciting fair decisions from AI systems and often undermining public trust in technology.

Various theoretical and practical approaches have been explored to improve algorithmic fairness. Fairness metrics such as demographic parity, equal opportunity and predictive equality have been derived to quantify different notions of fairness, appropriate in different contexts. These metrics are often incorporated during (in-processing) or after (post-processing) model training to enhance its fairness properties [29, 32, 36]. More importantly, fairness metrics play a crucial role in evaluating and vetting machine learning models prior to deployment in the real world. They also act as the final safeguard to ensure that models meet ethical principles and legal regulations, mitigate bias, and promote fair outcomes.

Most fairness metrics are estimated from labeled data. Such estimates are unbiased only if the distribution of the labeled data accurately reflects the target population on which the model will be applied. However, the idiosyncrasies in the generation or collection of the labels often leads to a distributional shift, broadly referred to as sample selection bias; e.g., a healthcare dataset that includes mostly data from urban hospitals might not represent rural healthcare needs. Selection bias could come in different forms such as sampling bias (over-representation or under-representation of certain groups or features in the population), self-selection bias (when the individuals who provide data differ systematically from those who do not), temporal bias (distribution changes over time), observer bias (emerging from subjective interpretation of data collectors/labelers), attrition bias (certain groups are systematically more likely to drop out or be excluded during data collection), survival bias (when only some types of cases are selected), etc.

Computing a fairness metric from labeled data under selection bias can give a statistically biased estimate, with significant systematic errors that do not vanish with increasing dataset sizes. When the estimate of a fairness metric is significantly off, a model not satisfying fairness constraints could be incorrectly deemed fit for deployment, and, conversely, a model that satisfies a fairness constraint could be deemed unfit. Mitigating such errors is imperative to accurately evaluating and vetting machine learning models.

In this work, we derive an approach to mitigate statistical bias on fairness metrics estimated from labeled data under selection bias. To this end, we rely on access to an unlabeled and unbiased dataset, often available in many applications, to infer bias mitigation parameters. Access to such data is commonly assumed in the domain adaptation literature and is referred to as the target domain or the population of interest where the model will be deployed [15, 44]. Such data is readily available in hiring, lending or genetics, as considered in this paper (Section 7). As correcting for arbitrary distribution shifts in the labeled data is an intractable problem, correcting for selection bias requires making explicit assumptions over the nature of the bias. Such assumptions are specified as a selection bias model that constrains how the relationship among the input features x and the target variable y changes due to selection bias [16, 42, 49]. An ideal selection bias model should be flexible enough to capture complex distribution shifts and yet lead to tractable algorithms for bias correction.[1]

Covariate-shift [43] and label-shift [15, 24, 40, 47] are two dominant selection bias models. However, they make strong assumptions on the form of the distribution shift in the biased data. Covariate-shift assumes that the probability of y given x ($p(y|x)$) remains unaltered due to selection bias whereas label-shift assumes that the probability of x given y ($p(x|y)$) is unaltered. Though they give tractable algorithms for bias correction of predictors, these assumptions do not capture the distribution shifts for most real-world applications, limiting their practical appeal. Further, a sparse joint-shift approach was developed to incorporate both covariate-shift and label-shift models within a limited number of features [2]. Recently, more flexible selection bias models of *Gaussian component-based mixing bias* (MB-GC) and *partition-projected class-conditional invariance* (PCC-invariance) have been derived as more flexible alternatives to covariate-shift and label-shift and have been demonstrated as effective tools to detect and mitigate the effect of bias in predictors [5, 50]. A salient aspect of these models is that they account for shift, in both $p(x|y)$ and $p(y|x)$, simultaneously and yet lead to a tractable algorithm for bias mitigation. MB-GC is based on a Gaussian mixture model with a shared component structure to account for selection bias, but it does not account for groups. PCC-invariance accounts for groups, but assumes the feature space to be composed of non-overlapping regions (partitions) that are over- and under-represented due to selection bias. The utility of these models has not been explored in the context of algorithmic fairness.

We derive a generalization of these models, *sub class-conditional invariance* (SCC-invariance), which assumes that the selection bias and differences among groups emerges from over and under-representation of overlapping regions in the feature space—an intuitive and flexible model for real-world data. Next, making a parametric Gaussian mixture assumption, similar to MB-GC, we derive two multi-sample EM algorithms to estimate model parameters. Lastly, we derive two importance sampling based approaches that incorporate the estimated parameters to give unbiased estimates of a given fairness metric. A key aspect of one of our approaches is that it can provide unbiased fairness metric estimates even without group information in the labeled data. This makes it applicable in scenarios where group information is withheld for privacy, policy, or ethical reasons. We demonstrate the effectiveness of our approach on three fairness metrics—equal opportunity, predictive equality, and predictive parity [32]—using synthetic data. We apply our approach to clinical genetics of rare Mendelian diseases to assess the fairness of widely used pathogenicity prediction tools across diverse populations.

2 Related Work

Fairness metrics in machine learning are broadly categorized into individual fairness and group fairness. Individual fairness, as introduced by Dwork et al. [10], emphasizes that similar individuals should be treated similarly. In contrast, group fairness [14] focuses on measuring disparities in outcomes between different demographic groups. There are multiple ways of defining fairness even within these broad categories. Some examples include equal opportunity, demographic parity and predictive parity for group fairness; Lipschitz fairness and counterfactual fairness for individual fairness. A comprehensive list of fairness metrics and their use cases can be found in [30] and [46]. Accurately estimating fairness

[1] We use the term "correction" to mean that bias can be eliminated only if the model assumptions hold; otherwise, the correction can only mitigate bias to some extent but cannot eliminate it completely.

metrics on many real-world datasets is often challenging due to factors such as small sample sizes, class or group imbalances, missing values and selection bias. Sample size and class imbalance can lead to unreliable conclusions and exaggerated disparities [20]. Additionally, fairness estimation becomes more complex in the presence of missing values, where incomplete data can introduce biases that distort fairness evaluations [13, 48]. An even greater challenge arises when protected attributes—such as race or gender—are unavailable or censored, requiring specialized techniques to infer disparities indirectly while maintaining robustness and reliability [11].

A particularly challenging scenario arises when the training data itself is subject to sample selection bias. Du and Wu [8] derive an approach to train fair classifiers robust to selection bias by employing a classification loss and a demographic parity-enhancing penalty term, both computed as a weighted average. However, their approach can only address selection bias under the restrictive covariate shift assumption. In contrast, our work shifts the focus from training fair classifiers to accurately estimating fairness metrics under sample selection bias. Our approach computes fairness metrics as an importance sampling-based weighted average but with a more flexible selection bias model. We also consider other fairness metrics (equal opportunity, predictive equality, and predictive parity), since demographic parity can be directly computed on the unbiased unlabeled data without any class labels.

Ji et al. [19] also derive an approach to improve fairness estimation when only a small set of labeled data is available by incorporating the unlabeled data to reduce the variance of the estimates, however, without considering the effect of selection bias. Their approach leverages a Bayesian framework to utilize the unlabeled data for improved fairness estimation. We use this approach as a baseline to compare our methods. Chzhen et al. [4] explore equal opportunity fairness in the presence of labeled and unlabeled data, proposing a group-dependent threshold for a Bayes classifier to achieve fairness. However, these studies assume access to unbiased labeled data and are tailored to specific fairness metrics rather than providing a general framework.

In this work, we address the problem of fairness estimation in the presence of both labeled and unlabeled data under sample selection bias. Unlike previous approaches, our correction methods are applicable to a broad range of fairness metrics, making them flexible and adaptable across different fairness criteria. By improving fairness estimation under biased data conditions, our work contributes to the development of more reliable and generalizable fairness assessments in machine learning.

3 Problem Formulation

We consider the binary classification problem to classify individuals belonging to protected groups (e.g., making hiring or lending decisions). Let $x \in \mathbb{R}^D$ contain an individual's attributes (input features), used to predict the class label $y \in \mathcal{Y} = \{-, +\}$, where + and − represent the positive and negative class, respectively. Let $g \in \mathcal{G} = \{1, 2, \ldots, G\}$ give the individual's group. Let $p(x, g, y)$ be the unknown joint distribution that governs how x, g and y appear in nature or in the target population where the classifier will be deployed. We refer to $p(x, g, y)$ as the unbiased distribution. For a given binary classifier $\hat{y}(x) \in \mathcal{Y}$, let $\mathrm{FM}(\hat{y})$ denote the true value of a fairness metric with respect to the unbiased distribution. Thus, only a statistically unbiased estimate of $\mathrm{FM}(\hat{y})$ accurately reflects a classifier's fairness when deployed in the real-world.

Let $\mathrm{FM}(\hat{y}; \mathcal{L})$ denote an estimate of $\mathrm{FM}(\hat{y})$ computed using the standard formula for the metric under consideration on the labeled dataset $\mathcal{L}$. For most fairness metrics, this involves computing an average of the classifier output over the labeled dataset or its subset. For example, equal opportunity is estimated as $1/|\mathcal{L}_1^+| \cdot \sum_{x \in \mathcal{L}_1^+} \hat{y}(x) - 1/|\mathcal{L}_2^+| \cdot \sum_{x \in \mathcal{L}_2^+} \hat{y}(x)$, where $\mathcal{L}_g^+$ is the subset of points in $\mathcal{L}$ from group g, labeled as positive. If $\mathcal{L}$ is an unbiased sample from $p(x, g, y)$, $\mathrm{FM}(\hat{y}; \mathcal{L})$ gives an unbiased estimate of $\mathrm{FM}(\hat{y})$. However, if $\mathcal{L}$ suffers from selection bias, i.e., is not a random sample from $p(x, g, y)$, then $\mathrm{FM}(\hat{y}; \mathcal{L})$ is not guaranteed to be an unbiased estimate of $\mathrm{FM}(\hat{y})$. The labeled data in many applications is often not representative of the unbiased distribution due to the idiosyncrasies of the labeling process, and consequently is not guaranteed to give a statistically unbiased estimate of the fairness metric.

Let $\mathcal{L}$ be the labeled data sample, with or without group identities included, and $\tilde{p}(x, g, y)$ or $\tilde{p}(x, y)$, respectively, denote the statistically biased distribution from which it is drawn. Let $\mathcal{U}$ denote an unbiased and unlabeled sample drawn from $p(x, g)$. Our main objective in this work is to derive approaches to obtain fairness estimates using both $\mathcal{L}$ and $\mathcal{U}$, $\mathrm{FM}(\hat{y}; \mathcal{L}, \mathcal{U})$, such that it is a better estimate of $\mathrm{FM}(\hat{y})$ than $\mathrm{FM}(\hat{y}; \mathcal{L})$. We refer to a fairness estimate with this property as a bias mitigated estimate. To this end, we introduce a flexible selection bias model in the form of probabilistic assumptions tying biased and unbiased distributions, and derive semi-supervised approaches to obtain unbiased fairness estimates under the assumptions.

4 Selection Bias Model

If the selection bias in the labeled data arbitrarily alters the distribution of x, g and y, i.e., $\tilde{p}(x, g, y)$ is arbitrarily different from $p(x, g, y)$, correcting the statistical bias in the fairness metric estimate is not possible. Fortunately, certain probabilistic assumptions from the domain adaptation literature such as covariate shift and label shift could make the problem tractable. However, these assumptions are not flexible enough to model the selection bias in most real-world data—covariate shift assumes $\tilde{p}(y|x) = p(y|x)$ whereas label shift assumes $\tilde{p}(x|y) = p(x|y)$. Furthermore, these assumptions do not explicitly account for groups.

4.1 Sub-class-conditional invariance

In this work we introduce a flexible sample selection bias model, *sub-class-conditional invariance* (**SCC**-invariance) to improve on the limitations from existing models by accounting for changes in both $p(x|y)$ and $p(y|x)$. Decomposing the positive and negative class-conditionals as mixture distributions with K^+ and K^- components, respectively, we get

$$p(x|+) = \sum_{k \in \mathcal{K}^+} w_k p(x|+, k), \quad \text{and} \quad p(x|-) = \sum_{k \in \mathcal{K}^-} v_k p(x|-, k),$$

where $\mathcal{K}^+ = \{1, 2 \ldots K^+\}$ and $\mathcal{K}^- = \{1, 2 \ldots K^-\}$, while $p(x|+, k)$ and $p(x|-, k)$ denote the k^{th} component of the positive and negative class-conditional, respectively. Probability vectors $\boldsymbol{w} = [w_k]_{k \in \mathcal{K}^+}$

and $\boldsymbol{v} = [v_k]_{k\in\mathcal{K}^-}$ are the mixing proportions; and $w_k \geq 0$, $v_k \geq 0$, $\sum_{k\in\mathcal{K}^+} w_k = 1$ and $\sum_{k\in\mathcal{K}^-} v_k = 1$. Note that we allow the component distributions to overlap.

Under **SCC**-invariance, we assume that the class-conditional distributions in the labeled data can also be expressed as mixtures sharing the same components, however, with different mixing proportions; i.e.,

$$\tilde{p}(x|+) = \sum_{k\in\mathcal{K}^+} \tilde{w}_k p(x|+,k), \text{ and } \tilde{p}(x|-) = \sum_{k\in\mathcal{K}^-} \tilde{v}_k p(x|-,k),$$

where $\tilde{\boldsymbol{w}} = [\tilde{w}_k]_{k\in\mathcal{K}^+}$ and $\tilde{\boldsymbol{v}} = [\tilde{v}_k]_{k\in\mathcal{K}^-}$ are the mixing proportions for the biased class-conditionals. In other words, **SCC**-invariance represents a form of selection bias where certain regions (possibly overlapping) in the feature space (represented by component distributions) get undersampled whereas other regions get oversampled. Note that these assumptions are identical to *Gaussian component based mixing-bias* (MB-GC) [5] when the components are Gaussian.

To capture the distributional similarities and differences between groups and selection bias within a group, we extend the shared component mixture assumption on the group class-conditionals. Formally,

$$p(x|+,g) = \sum_{k\in\mathcal{K}^+} w_{gk} p(x|+,k) \text{ and } p(x|-,g) = \sum_{k\in\mathcal{K}^-} v_{gk} p(x|-,k),$$
$$\tilde{p}(x|+,g) = \sum_{k\in\mathcal{K}^+} \tilde{w}_{gk} p(x|+,k) \text{ and } \tilde{p}(x|-,g) = \sum_{k\in\mathcal{K}^-} \tilde{v}_{gk} p(x|-,k),$$

where $\boldsymbol{w}_g = \left[w_{gk}\right]_{k\in\mathcal{K}^+}$ and $\boldsymbol{v}_g = \left[v_{gk}\right]_{k\in\mathcal{K}^-}$ are the mixing proportions for the unbiased group class-conditionals while $\tilde{\boldsymbol{w}}_g = [\tilde{w}_{gk}]_{k\in\mathcal{K}^+}$ and $\tilde{\boldsymbol{v}}_g = [\tilde{v}_{gk}]_{k\in\mathcal{K}^-}$ are the mixing proportions for the biased group class-conditionals. In summary, the differences between the class-conditionals across the groups and within a group (due to selection bias) are explained by the differing mixing proportions, reflecting over and under sampling from components. Note that **SCC**-invariance is identical to the PCC-invariance when the components are non-overlapping [50].

In this work, we will consider a special case of the **SCC**-invariance model, where all the components take a Gaussian form similar to MB-GC; i.e., $p(x|+,k) = \phi(x;\mu_k^+,\Sigma_k^+)$ and $p(x|-,k) = \phi(x;\mu_k^-,\Sigma_k^-)$, where $\phi(x;\mu,\Sigma)$ is the Gaussian density with mean μ and covariance Σ. Thus, all the class-conditionals follow a Gaussian Mixture Model (GMM). Gaussian mixtures are universal approximators of continuous probability densities with enough components and are considered to become nonparametric for $K^+, K^- \to \infty$ [12], making them a reasonable choice for a flexible yet tractable selection bias model. Under the Gaussian mixture assumption, **SCC**-invariance is a generalization of MB-GC to account for groups. In Appendix F, we demonstrate the practical advantage of **SCC**-invariance over MB-GC.

4.2 Parameter estimation

We derive two multi-sample Expectation Maximization (EM) algorithms, MS-GMM-$\boldsymbol{\Theta}_1$ and MS-GMM-$\boldsymbol{\Theta}_2$, to obtain maximum loglikelihood estimates of the Gaussian mixture model (GMM) parameters by fitting the labeled ($\mathcal{L}$) and unlabeled data ($\mathcal{U}$). The parameters will be later used to compute importance sampling weights (Table 1) to mitigate the bias in fairness metric estimates.

Let $\mathcal{L}^+ \subseteq \mathcal{L}$ and and $\mathcal{L}^- \subseteq \mathcal{L}$ contain all positive and negative examples, respectively, in $\mathcal{L}$. Let $\mathcal{U}_g \subseteq \mathcal{U}$ and $\mathcal{L}_g \subseteq \mathcal{L}$ contain all examples that belong to group g in $\mathcal{U}$ and $\mathcal{L}$, respectively. Let $\mathcal{L}_g^+ \subseteq \mathcal{L}_g$ and $\mathcal{L}_g^- \subseteq \mathcal{L}_g$ contain all positive and negative examples in $\mathcal{L}$ belonging to group g, respectively. Next, we list all the parameters of the model. For $* \in \{+,-\}$, $\boldsymbol{\mu}^* = [\mu_k^*]_{k\in\mathcal{K}^*}$ and $\Sigma^* = [\Sigma_k^*]_{k\in\mathcal{K}^*}$ contain the means and covariance matrices of the Gaussian components in the mixture model. Let $\boldsymbol{w}$ ($\boldsymbol{v}$) and $\tilde{\boldsymbol{w}}$ ($\tilde{\boldsymbol{v}}$) be the mixing proportions in the mixture formulation of the unbiased and biased positive (negative) class-conditionals, respectively, as defined earlier. Let $\boldsymbol{w}_g$ ($\boldsymbol{v}_g$) and $\tilde{\boldsymbol{w}}_g$ ($\tilde{\boldsymbol{v}}_g$) be the mixing proportions of the unbiased and biased positive (negative) class-conditionals, respectively, for group g. Let $\alpha = p(y=+)$ and $\tilde{\alpha} = \tilde{p}(y=+)$ be the proportion of positives in the unbiased and biased distributions, respectively. Let $\alpha_g = p(y=+|g)$ and $\tilde{\alpha}_g = \tilde{p}(y=+|g)$ be the proportion of positives in the unbiased and biased distributions for group g, respectively. Let $\boldsymbol{\eta} = [\eta_g]_{g\in\mathcal{G}}$ and $\tilde{\boldsymbol{\eta}} = [\tilde{\eta}_g]_{g\in\mathcal{G}}$ contain the proportion of group g individuals, $p(g)$ and $\tilde{p}(g)$, in the unbiased and biased distributions, respectively. Notice that $\tilde{\alpha}$ can be estimated directly from $\mathcal{L}$ as the fraction of positives in it ($|\mathcal{L}^+|/|\mathcal{L}|$) and $\tilde{\alpha}_g$, by computing the fraction in group wise manner ($|\mathcal{L}_g^+|/|\mathcal{L}_g|$). Similarly, $\boldsymbol{\eta}$ and $\tilde{\boldsymbol{\eta}}$ can be estimated from the fraction of group g individuals in the $\mathcal{U}$ ($|\mathcal{U}_g|/|\mathcal{U}|$) and $\mathcal{L}$ ($|\mathcal{L}_g|/|\mathcal{L}|$), respectively, provided they contain the group information. All of the remaining parameters are hidden and cannot be directly estimated with simple formulas.

Let $\boldsymbol{\Theta}_1 = \{\boldsymbol{w}_g, \tilde{\boldsymbol{w}}_g, \boldsymbol{v}_g, \tilde{\boldsymbol{v}}_g, \alpha_g\}_{g\in\mathcal{G}} \cup \{\boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \Sigma^+, \Sigma^-\}$ denote a subset of the hidden parameters. Note that $\boldsymbol{w}, \tilde{\boldsymbol{w}}, \boldsymbol{v}, \tilde{\boldsymbol{v}}$ and α can be inferred from $\boldsymbol{\eta}, \tilde{\boldsymbol{\eta}}, \tilde{\alpha}_g$ and the parameters in $\boldsymbol{\Theta}_1$ as follows.

$$\boldsymbol{w} = \frac{\sum_{g\in\mathcal{G}} \alpha_g \eta_g \boldsymbol{w}_g}{\sum_{g\in\mathcal{G}} \alpha_g \eta_g}, \quad \boldsymbol{v} = \frac{\sum_{g\in\mathcal{G}} (1-\alpha_g)\eta_g \boldsymbol{v}_g}{\sum_{g\in\mathcal{G}} (1-\alpha_g)\eta_g}, \quad \alpha = \sum_{g\in\mathcal{G}} \alpha_g \eta_g$$
$$\tilde{\boldsymbol{w}} = \frac{\sum_{g\in\mathcal{G}} \tilde{\alpha}_g \tilde{\eta}_g \tilde{\boldsymbol{w}}_g}{\sum_{g\in\mathcal{G}} \tilde{\alpha}_g \tilde{\eta}_g}, \quad \tilde{\boldsymbol{v}} = \frac{\sum_{g\in\mathcal{G}} (1-\tilde{\alpha}_g)\tilde{\eta}_g \tilde{\boldsymbol{v}}_g}{\sum_{g\in\mathcal{G}} (1-\tilde{\alpha}_g)\tilde{\eta}_g}. \tag{1}$$

MS-GMM-$\boldsymbol{\Theta}_1$: Our first algorithm, MS-GMM-$\boldsymbol{\Theta}_1$, iteratively estimates $\boldsymbol{\Theta}_1$ using the following update equations, derived under the EM framework [6, 28] (Appendix C). Here $* \in \{+,-\}$ and symbols accented with $\check{}$ and $\hat{}$ denote the old parameters, before the applying the updates, and the new parameters, respectively.

$$\hat{\alpha}_g \leftarrow \frac{1}{|\mathcal{U}_g|} \sum_{x\in\mathcal{U}_g} \sum_{k\in\mathcal{K}^+} \check{\pi}_{gk}^+(x),$$
$$\hat{\tilde{w}}_{gk} \leftarrow \frac{1}{|\mathcal{L}_g^+|} \sum_{x\in\mathcal{L}_g^+} \check{\rho}_{gk}^+(x), \quad \hat{w}_{gk} \leftarrow \frac{1}{\hat{\alpha}_g |\mathcal{U}_g|} \sum_{x\in\mathcal{U}_g} \check{\pi}_{gk}^+(x),$$
$$\hat{\tilde{v}}_{gk} \leftarrow \frac{1}{|\mathcal{L}_g^-|} \sum_{x\in\mathcal{L}_g^-} \check{\rho}_{gk}^-(x), \quad \hat{v}_{gk} \leftarrow \frac{1}{(1-\hat{\alpha}_g)|\mathcal{U}_g|} \sum_{x\in\mathcal{U}_g} \check{\pi}_{gk}^-(x),$$
$$\hat{\mu}_k^* \leftarrow \frac{\sum_{g\in\mathcal{G}} \left[\sum_{x\in\mathcal{L}_g^*} \check{\rho}_{gk}^*(x) x + \sum_{x\in\mathcal{U}_g} \check{\pi}_{gk}^*(x) x\right]}{\sum_{g\in\mathcal{G}} \left[\sum_{x\in\mathcal{L}_g^*} \check{\rho}_{gk}^*(x) + \sum_{x\in\mathcal{U}_g} \check{\pi}_{gk}^*(x)\right]},$$
$$\hat{\Sigma}_k^* \leftarrow \frac{\sum_{g\in\mathcal{G}} \left[\sum_{x\in\mathcal{L}_g^*} \check{\rho}_{gk}^*(x) \check{x}_k^* \otimes \check{x}_k^* + \sum_{x\in\mathcal{U}_g} \check{\pi}_{gk}^*(x) \check{x}_k^* \otimes \check{x}_k^*\right]}{\sum_{g\in\mathcal{G}} \left[\sum_{x\in\mathcal{L}_g^*} \check{\rho}_{gk}^*(x) + \sum_{x\in\mathcal{U}_g} \check{\pi}_{gk}^*(x)\right]},$$

where $\check{x}_k^* = x - \check{\mu}_k^*$; $\otimes$ denotes the outer product between two vectors; $\check{\rho}_{gk}^*(x)$ and $\check{\pi}_{gk}^*(x)$ are computed with the old value of the

parameters in $\check{\mathbf{\Theta}}_1$ in the formula for $\rho^*_{gk}(x)$ and $\pi^*_{gk}(x)$, respectively, given below. Let $\rho^+_{gk}(x)$ ($\rho^-_{gk}(x)$) denote the probability that a random positive (negative) labeled example from group g comes from the k^{th} positive (negative) component. Let $\pi^+_{gk}(x)$ ($\pi^-_{gk}(x)$) be the probability that a random unlabeled example from group g comes from the k^{th} positive (negative) component. Mathematically,

$$\rho^+_{gk}(x) = \tilde{p}(k|x,g,+) = \frac{\tilde{w}_{gk}\phi^+_k(x)}{\text{NMix}(x;\tilde{\theta}^+_g)}$$

$$\rho^-_{gk}(x) = \tilde{p}(k|x,g,-) = \frac{\tilde{v}_{gk}\phi^-_k(x)}{\text{NMix}(x;\tilde{\theta}^-_g)}$$

$$\pi^+_{gk}(x) = p(k,+|x,g) = \frac{\alpha_g w_{gk}\phi^+_k(x)}{\alpha_g \text{NMix}(x;\theta^+_g) + (1-\alpha_g)\text{NMix}(x;\theta^-_g)}$$

$$\pi^-_{gk}(x) = p(k,-|x,g) = \frac{(1-\alpha_g)v_{gk}\phi^-_k(x)}{\alpha_g \text{NMix}(x;\theta^+_g) + (1-\alpha_g)\text{NMix}(x;\theta^-_g)},$$

where θ^*_g and $\tilde{\theta}^*_g$ denote the parameters of the unbiased and biased class-conditionals for group g, respectively, as given in the caption of Table 1, while $\text{NMix}(x;\theta^*_g)$ and $\text{NMix}(x;\tilde{\theta}^*_g)$ are the unbiased and biased class-conditional density functions for group g, respectively, also defined in the caption of Table 1. The initial parameters, serving as input to MS-GMM-$\mathbf{\Theta}_1$, are given in Appendix E.1. The estimated $\mathbf{\Theta}_1$ allows computing weights $\psi^*_g(x)$ in Table 1. To compute $\chi^*_g(x)$ estimates of parameters $\tilde{\boldsymbol{w}}$ and $\tilde{\boldsymbol{v}}$ are required, which can be obtained by applying Eq. 1.

MS-GMM-$\mathbf{\Theta}_2$: Our second algorithm, MS-GMM-$\mathbf{\Theta}_2$, estimates $\mathbf{\Theta}_2 = \{\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{v}}\} \cup \{\boldsymbol{w}_g, \boldsymbol{v}_g, \alpha_g\}_{g\in\mathcal{G}} \cup \{\boldsymbol{\mu}^+, \boldsymbol{\mu}^-, \Sigma^+, \Sigma^-\}$, instead of $\mathbf{\Theta}_1$, by ignoring the group information in $\mathcal{L}$. MS-GMM-$\mathbf{\Theta}_2$ can not be used to compute weights $\psi^*_g(x)$ (Table 1), however, it can still be used to compute $\chi^*_g(x)$ for a valid fairness metric correction approach. Since it does not require the group information in $\mathcal{L}$, it can be used when group annotations are missing from $\mathcal{L}$; for example, in our clinical genetics data (Section 7). The parameter update equations for MS-GMM-$\mathbf{\Theta}_2$ are given by

$$\hat{\tilde{w}}_k \leftarrow \frac{1}{|\mathcal{L}^+|}\sum_{x\in\mathcal{L}^+}\check{\rho}^+_k(x), \quad \hat{\tilde{v}}_k \leftarrow \frac{1}{|\mathcal{L}^-|}\sum_{x\in\mathcal{L}^-}\check{\rho}^-_k(x),$$

$$\hat{\mu}^*_k \leftarrow \frac{\sum_{x\in\mathcal{L}^*}\check{\rho}^*_k(x)x + \sum_{g\in\mathcal{G}}\sum_{x\in\mathcal{U}_g}\check{\pi}^*_{gk}(x)x}{\sum_{x\in\mathcal{L}^*}\check{\rho}^*_k(x) + \sum_{g\in\mathcal{G}}\sum_{x\in\mathcal{U}_g}\check{\pi}^*_{gk}(x)},$$

$$\hat{\Sigma}^*_k \leftarrow \frac{\sum_{x\in\mathcal{L}^*}\check{\rho}^*_k(x)\check{x}^*_k\otimes\check{x}^*_k + \sum_{g\in\mathcal{G}}\sum_{x\in\mathcal{U}_g}\check{\pi}^*_{gk}(x)\check{x}^*_k\otimes\check{x}^*_k}{\sum_{x\in\mathcal{L}^*}\check{\rho}^*_k(x) + \sum_{g\in\mathcal{G}}\sum_{x\in\mathcal{U}_g}\check{\pi}^*_{gk}(x)},$$

where $\check{x}^*_k$, $\otimes$ and $\check{\pi}^*_{gk}(x)$ are as defined above for MS-GMM-$\mathbf{\Theta}_1$. The update rules for $\hat{\alpha}_g$, $\hat{w}_{gk}$, and $\hat{v}_{gk}$ are the same as MS-GMM-$\mathbf{\Theta}_1$. $\check{\rho}^+_k(x)$ is computed with the old value of the parameters in $\check{\mathbf{\Theta}}_2$ in the formula for $\rho^*_k(x)$ given below. Let $\rho^+_k(x)$ ($\rho^-_k(x)$) denote the probability that a random positive (negative) labeled example comes from the k^{th} positive (negative) component. Mathematically,

$$\rho^+_k(x) = \tilde{p}(k|x,+) = \frac{\tilde{w}_k\phi^+_k(x)}{\text{NMix}(x;\tilde{\theta}^+)}$$

$$\rho^-_k(x) = \tilde{p}(k|x,-) = \frac{\tilde{v}_k\phi^-_k(x)}{\text{NMix}(x;\tilde{\theta}^-)}$$

Weights	General formula	Under GMM assumption
$\psi^*_g(x)$	$=\frac{p(x\|*,g)}{\tilde{p}(x\|*,g)}$	$=\frac{\text{NMix}(x;\theta^*_g)}{\text{NMix}(x;\tilde{\theta}^*_g)}$
$\chi^*_g(x)$	$=\frac{p(x\|*,g)}{\tilde{p}(x\|*)}$	$=\frac{\text{NMix}(x;\theta^*_g)}{\text{NMix}(x;\tilde{\theta}^*)}$

Table 1: Weights used in the unbiased estimation of fairness metrics. $* \in \{+,-\}$. $\alpha^+ = \alpha$ and $\alpha^- = (1-\alpha)$. $\theta^+ = \{\boldsymbol{\mu}^+, \Sigma^+, \boldsymbol{w}\}$ and $\theta^- = \{\boldsymbol{\mu}^-, \Sigma^-, \boldsymbol{v}\}$ denote the parameters of the unbiased class-conditionals under the Gaussian mixture assumption. Here $\boldsymbol{\mu}^* = [\mu^*_k]_{k\in\mathcal{K}^*}$ and $\Sigma^* = [\Sigma^*_k]_{k\in\mathcal{K}^*}$ contain the means and covariance matrices of the Gaussian components and $\tilde{\theta}^+ = \{\boldsymbol{\mu}^+, \Sigma^+, \tilde{\boldsymbol{w}}\}$ and $\tilde{\theta}^- = \{\boldsymbol{\mu}^-, \Sigma^-, \tilde{\boldsymbol{v}}\}$ denote the parameters of the biased class-conditionals. $\theta^+_g = \{\boldsymbol{\mu}^+, \Sigma^+, \boldsymbol{w}_g\}$ and $\theta^-_g = \{\boldsymbol{\mu}^-, \Sigma^-, \boldsymbol{v}_g\}$ denote the parameters of the unbiased class-conditionals for group g. $\tilde{\theta}^+_g = \{\boldsymbol{\mu}^+, \Sigma^+, \tilde{\boldsymbol{w}}_g\}$ and $\tilde{\theta}^-_g = \{\boldsymbol{\mu}^-, \Sigma^-, \tilde{\boldsymbol{v}}_g\}$ denote the parameters of the biased class-conditionals for group g. $\text{NMix}(x;\theta^+) = \sum_{k\in\mathcal{K}^+} w_k\phi(x;\mu^+_k,\Sigma^+_k)$ and $\text{NMix}(x;\theta^-) = \sum_{k\in\mathcal{K}^-} v_k\phi(x;\mu^-_k,\Sigma^-_k)$ gives the positive and negative unbiased class-conditional density functions, respectively, where $\phi(x;\mu,\Sigma)$ is the Gaussian density function with mean μ and covariance Σ. $\text{NMix}(x;\tilde{\theta}^*)$, $\text{NMix}(x;\theta^*_g)$ and $\text{NMix}(x;\tilde{\theta}^*_g)$ are similarly defined as Gaussian mixture densities using the corresponding parameters.

where $\tilde{\theta}^+$ and $\tilde{\theta}^-$ denote the parameters of the biased positive and negative class-conditionals (marginalized over g), respectively, as given in the caption of Table 1, while $\text{NMix}(x;\tilde{\theta}^+)$ and $\text{NMix}(x;\tilde{\theta}^-)$ are the corresponding class-conditional density functions, also defined in the caption of Table 1. The initial parameters, serving as input to MS-GMM-$\mathbf{\Theta}_2$, are given in Appendix E.1.

5 Bias-Mitigated Fairness Metric Estimates

When the labeled data is biased, the standard formulas used to compute a fairness metric could have significant systematic errors and cannot be trusted to evaluate or vet models. We derive two importance sampling based approaches to improve the fairness metric estimates, in spite of the selection bias. Most fairness metrics can be expressed in terms of the group-wise true positive rate (TPR), false positive rate (FPR) and the unbiased proportion of positives within each group, α_g. We first derive bias mitigation approaches for TPR and FPR estimation. Proof provided in Appendix A.

5.1 Bias-mitigated TPR, FPR and PPV

Groups in labeled data (GIL) : To get unbiased estimates of a classifier's ($\hat{y}$) TPR and FPR for group g, we first give an approach that uses labeled points only from group g in a importance sampling based weighted average formulation.

$$\widehat{\text{TPR}}_g(\hat{y};\mathcal{L},\mathcal{U}) \overset{est}{=} \frac{1}{|\mathcal{L}^+_g|}\sum_{x\in\mathcal{L}^+_g}\hat{\psi}^+_g(x)\hat{y}(x),$$

$$\widehat{\text{FPR}}_g(\hat{y};\mathcal{L},\mathcal{U}) \overset{est}{=} \frac{1}{|\mathcal{L}^-_g|}\sum_{x\in\mathcal{L}^-_g}\hat{\psi}^-_g(x)\hat{y}(x),$$

where $\hat{\psi}^+_g(x)$ and $\hat{\psi}^-_g(x)$ are estimates of $\psi^+_g(x)$ and $\psi^-_g(x)$, respectively, obtained by substituting the estimated GMM parameters as

defined in Table 1. Note that GIL can be implemented with estimated $\boldsymbol{\Theta}_1$ parameters but not $\boldsymbol{\Theta}_2$, thereby requiring MS-GMM-$\boldsymbol{\Theta}_1$ as an intermediate step and group information for the labeled data.

Groups not in labeled data (GNIL): Although GIL gives an unbiased estimate relying solely on data points from group g, it could have large variance. To reduce the variance and still obtain an unbiased estimate, we give another correction approach that uses the entire labeled dataset (including data points from other groups) to compute group-wise TPR and FPR as follows

$$\widehat{\text{TPR}}_g(\hat{y}; \mathcal{L}, \mathcal{U}) \overset{est}{=} \frac{1}{|\mathcal{L}^+|} \sum_{x \in \mathcal{L}^+} \hat{\chi}_g^+(x)\hat{y}(x),$$

$$\widehat{\text{FPR}}_g(\hat{y}; \mathcal{L}, \mathcal{U}) \overset{est}{=} \frac{1}{|\mathcal{L}^-|} \sum_{x \in \mathcal{L}^-} \hat{\chi}_g^-(x)\hat{y}(x),$$

where $\hat{\chi}_g^+(x)$ and $\hat{\chi}_g^-(x)$ are estimates of $\chi_g^+(x)$ and $\chi_g^-(x)$, respectively, obtained by substituting the estimated GMM parameters as defined in Table 1.

Note that GNIL can be implemented with either $\boldsymbol{\Theta}_1$ (using Eq. 1) or $\boldsymbol{\Theta}_2$ estimates, thereby requiring either MS-GMM-$\boldsymbol{\Theta}_1$ or MS-GMM-$\boldsymbol{\Theta}_2$ as an intermediate step. Using MS-GMM-$\boldsymbol{\Theta}_2$ it can be implemented without group information in the labeled data.

Note that bias-mitigated estimates of many other evaluation metrics can be obtained from TPR FPR and α_g. For example, bias-mitigated precision or Positive Predictive Value (PPV) for group g can be obtained by substituting bias-mitigated TPR and FPR along with α_g estimates from MS-GMM algorithms:

$$\text{PPV}_g = \frac{\text{TPR}_g \alpha_g}{\text{TPR}_g \alpha_g + \text{FPR}_g(1 - \alpha_g)} .$$

In theory, the importance sampling approach for bias mitigation can be applied to any evaluation metric, by defining appropriate weight functions. For example, bias-mitigated Expected Calibration Error (ECE) can be estimated as shown in Appendix B.

5.2 Bias-mitigated fairness metrics

Mitigation of bias in fairness metric estimates follows directly from reducing bias in TPR and FPR, along with estimate of α_g from the MS-GMM algorithms. In this work we focus on the notion of group fairness, i.e., fairness evaluated for a group as a whole, instead of at an individual level, and consider the following three metrics:

Equal opportunity (EO): requires that for each group the probability of making a positive prediction (e.g., hired) among individuals belonging to the positive class (e.g., deserve to be hired) is identical. In other words, the TPR for all the groups be identical. Considering a two group setting, we measure EO, as the difference between the TPR of the two groups; $\text{EO} = \text{TPR}_{g_1} - \text{TPR}_{g_2}$. The metric does not consider individuals in the negative class.

Predictive Equality (PE): requires that for each group the probability of making a positive prediction (e.g., hired) among individuals belonging to the negative class (e.g., not deserve to be hired) is identical. In other words, the FPR for all the groups be identical. Considering a two group setting, we measure PE, as the difference between the FPR of the two groups; $\text{PE} = \text{FPR}_{g_1} - \text{FPR}_{g_2}$. The metric does not consider individuals in the positive class.

Predictive Parity (PP): requires that for each group the probability of making a correct positive prediction (e.g., hired and deserve to be hired) among individuals predicted to have a positive outcome (e.g., hired) is identical. In other words, the PPV for all the groups be identical. Considering a two group setting, we measure PP, as the difference between the PPV of the two groups; $\text{PP} = \text{PPV}_{g_1} - \text{PPV}_{g_2}$. The metric considers individuals in both the positive and the negative class as well the proportion of each class with the groups.

The GIL and GNIL bias mitigation approaches for EO, PE and PP follow directly from the corrections to TPR, FPR and PPV presented in Section 5.1.

6 Experimental Evaluation

6.1 Synthetic data generation

To evaluate our proposed approaches to mitigating error in fairness metric estimation when labeled data exhibits sample selection bias, we construct challenging synthetic datasets. We use synthetic data where the selection bias can be controlled and quantified precisely via the data generation model parameters. Namely, the similarity between two mixture distributions can be quantified by the AUC-ROC of an ideal classifier trained to classify points from each mixture. For synthetic data with known mixture model parameters, the ideal classifier comes from the (known) mixture posteriors.

We generate synthetic data by first picking K Gaussian components to represent the positives and another K Gaussian components to represent the negatives ($K^+ = K^- = K$ in our experiments). The base parameters, the mean and standard deviation of the components, are determined such that the two class-conditional distributions fall within a prescribed AUC$(+,-)$ range. Note that all group class-conditionals, biased or unbiased, share the same K positive or negative components in their Gaussian mixture formulation, and only vary in their mixing proportions, $\boldsymbol{w}_g$ and $\boldsymbol{v}_g$ for the unbiased mixtures, $p(x|+,g)$ and $p(x|-,g)$, respectively, and $\tilde{\boldsymbol{w}}_g$ and $\tilde{\boldsymbol{v}}_g$ for the biased mixtures, $\tilde{p}(x|+,g)$ and $\tilde{p}(x|-,g)$, respectively.

For a given pair of number of components, $K \in \{2, 4, 8\}$, and dimension, $D \in \{2, 8\}$, we generate 315 datasets ($\mathcal{L}$ and $\mathcal{U}$) equally divided between three AUC$(+,-)$ buckets, $[0.65, 0.75]$, $[0.75, 0.85]$ and $[0.85, 0.95]$, and having varying degree of selection bias. Interpreting AUC as a measure of separation between two distributions, we measure the selection bias also as AUC, however, between the biased and unbiased class-conditionals. For a given selection bias bucket in Figure 1, biased and unbiased mixing proportions are selected such that each group class-conditional's selection bias is in that bucket. The 315 datasets are divided among the top three selection bias buckets, having approximately 90 datasets each, and the the bottom bucket with half the width, having 45 datasets. For each dataset, we generate a total of $25,000$ samples, where $\lambda \in \{0.5, 0.1\}$ specifies the proportion of samples that are labeled. The ratio in which the samples are distributed between the groups is determined by the parameters η and $\tilde{\eta}$ for the unlabeled and labeled data respectively. For example, $\eta = 0.2$ indicates that 20% of the unlabeled samples belong to g_1 and 80% belong to g_2. For our experiments, we consider two different configurations: (i) equal: $\eta = \tilde{\eta} = 0.5$, where the samples are split equally between the groups for labeled and unlabeled data, and (ii) minority: $\eta = 0.1$ and $\tilde{\eta} = 0.05$, where g_1 is severely underrepresented in the unlabeled data and this effect is further magnified in the labeled data.

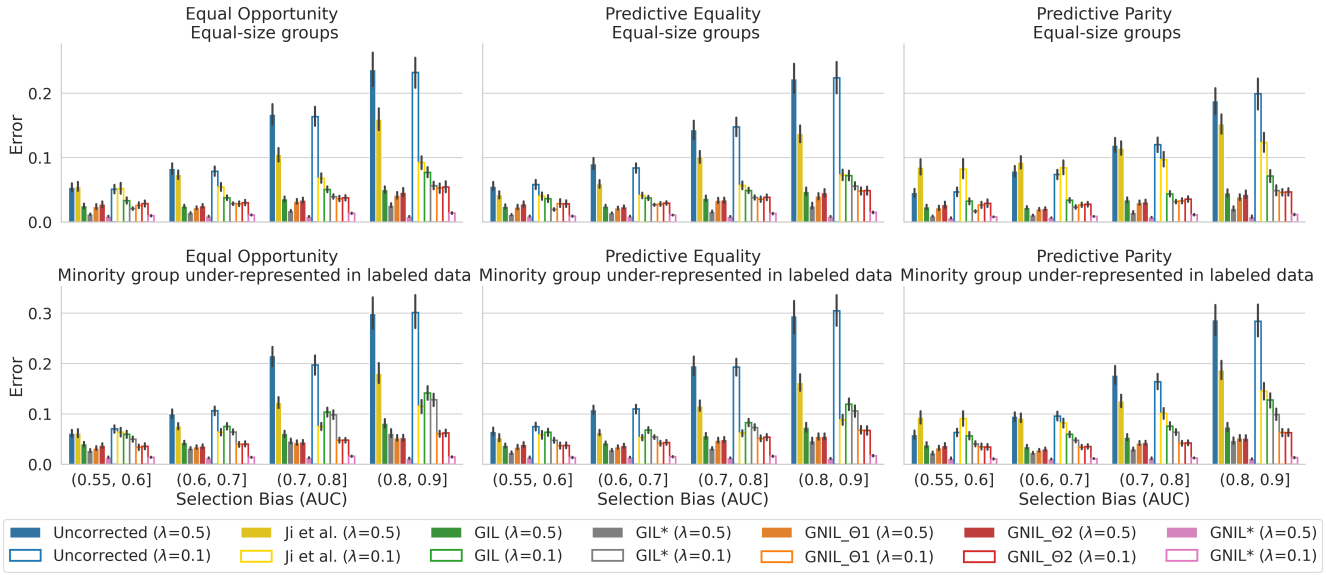

Figure 1: **Mean absolute error of fairness metric estimates for synthetic data and 95% confidence interval error bars for increasing levels of selection bias. Results for equal-size groups (top row) and when the minority group is under-represented in labeled data, $\eta = [0.1, 0.9]$, $\eta_l = [0.05, 0.95]$ (bottom row). Solid (unfilled) bars represent datasets where the proportion of data with labels is $\lambda = 0.5$ ($\lambda = 0.1$). Corrections that use ground-truth parameters are designated by an asterisk (GIL* and GNIL*). Each bar summarizes the errors across data generated with 1) 2 and 8 dimensions 2) 2,4 and 8 positive and negative components and 3) the 3 AUC$(+,-)$ buckets. Each bar in lowest selection bias interval, (0.55, 0.6], contain results from 6×45 datasets and those in the remaining three selection bias intervals contain 6×90 datasets.**

6.2 Results

Figure 1 compares the mean absolute error in the estimation of equal opportunity (EO), predictive equality (PE), and predictive parity (PP) across different values of selection bias for seven methods: 1) Uncorrected: standard approach, 2) Ji et al. [19]: baseline, 3) GIL with MS-GMM-$\boldsymbol{\Theta}_1$ based parameter estimates, 4) GIL*: GIL with true parameters, 5) GNIL-Θ_1: GNIL with MS-GMM-$\boldsymbol{\Theta}_1$ based parameter estimates, 6) GNIL-Θ_2: GNIL with MS-GMM-$\boldsymbol{\Theta}_2$ based parameter estimates, and 7) GNIL*: GNIL with true parameters. The error is computed w.r.t. the ground truth value measured on the unlabeled data, using the true labels. We ran MS-GMM-$\boldsymbol{\Theta}_1$ and MS-GMM-$\boldsymbol{\Theta}_2$ fixing the number of components to their true values used in the data generation. Experimental results in [5] suggest that the number of components can be accurately determined using the data log-likelihood via cross-validation. We provide the log-likelihood criteria for our appraoches in Appendix D. The baseline [19] was derived to obtain low-variance estimates by incorporating unlabeled data with a few labeled examples, without any consideration of selection bias.

Comparison with Uncorrected: As expected, the magnitude of error for the uncorrected estimator increases significantly with an increase in selection bias, reaching as high as 0.25 - 0.4 for the highest selection bias bucket. However, the increase in error for our GIL- and GNIL-based correction approaches is marginal. We observe that the error in fairness disparities is significantly reduced while employing the proposed bias mitigation approaches across all metrics and configurations. This shows that the approach is effective in producing fairness estimates that are close to unbiased estimates even in the presence of selection bias.

Comparison with baseline: In all settings, our GNIL-based approaches show an improvement over Ji et al. [19]. Our GIL-based approach also shows an improvement in most settings, except when proportion of labeled data is low ($\lambda = 0.1$) and additionally, the minority group is significantly under-represented (bottom row, unfilled bars). Notice that this only happens for EO and PE (not for PP) and the difference in the mean absolute error is small. This setting is particularly challenging, since there are only a small number of labeled points from the minority group available. Since GIL computes each group's term (e.g., TPR_g) in a fairness metric estimate as a weighted average on the labeled points from that group only, it is more sensitive to the size of the minority group in the labeled set. In comparison, GNIL uses the entire labeled set (from both groups) to compute each group's term, making it more robust to the size of the minority group in the labeled set. As expected, the error in the baseline approach increases significantly with selection bias, specially when the biased labeled data has more influence ($\lambda = 0.5$) over the estimate; the error increase is modest for $\lambda = 0.1$ in comparison. Overall, our GIL and GNIL-based approaches, show significant improvement over the baseline in high selection bias paradigm, specially when $\lambda = 0.5$.

Absence of groups in the labeled data: The strong performance of GNIL-Θ_2 (comparable with GNIL-Θ_1 and better than GIL) demonstrates that accurate bias mitigated fairness estimates can be obtained without access to the group memberships in the labeled data. Thus estimating the group specific parameters only from the unlabeled data (MS-GMM-$\boldsymbol{\Theta}_2$) and computing group specific fairness metric terms (e.g., TPR_g) as weighted averages over the entire labeled set provides an effective solution.

Ablation study: To isolate the impact of our importance sampling based reweighting from the EM based parameter estimation on bias mitigation, we present results for GIL- and GNIL-based oracle methods in Figure 1. In this setting, weights for bias mitigation are computed using the true parameters rather than the estimated ones. We see that the estimated errors are higher, but close to the best-case oracle corrections. The magnitude of the oracle-parameter corrections are stable across various degrees of selection bias, while the estimated-parameter corrections marginally increase with increasing selection bias in the labeled data.

7 Case Study: Variant Impact Prediction

Prediction of the impact of genomic variants is an important computational task in genomic medicine [37, 39, 41]. In complex disease, the objective is to approximate the risk of a phenotype in the form of a polygenic risk score (PRS). PRS is usually implemented as a linear combination of estimated effect sizes and the presence indicators of multiple common variants in an individual's genome as the effect of any such variant on the phenotype is small. As such, PRS can provide clinical utility for conditions such as cancer, cardiovascular disease and diabetes [9]. In contrast, the objective in rare disease is to approximate the risk that a single genomic variant causes disease. Since variants can have large effects, the ability of an algorithm to predict pathogenic ones can be used as a proxy to predicting a clinical phenotype.

It has been well documented that PRS estimators display limited transferability between populations of different genetic ancestries [26, 27]. The underlying factors implicate population-specific allele frequencies for common variants as well as the sizes of haplotype blocks that impact estimation of variant effect sizes. These findings have spurred large-scale sequencing projects for populations with less publicly available data as well as research on computational transferability of PRS akin to domain adaptation [25]. The situation is less clear in rare disease [7, 33], where predictors are traditionally constructed from the reference genome sequence and evolutionary conservation between species [22]. In this section, our objective is to investigate rare disease and gain insight into the performance of variant impact predictors on individuals from different genetic ancestries. We focus on missense variants, those that result in a single amino acid substitution in an individual's protein sequence relative to the reference genome.

7.1 Data and Estimation Method

Missense variants with pathogenicity class labels were collected from ClinVar [23] while the unlabeled data was extracted from gnomAD [21]. Variants in ClinVar come from the patient population and are skewed towards well-studied disease genes and more frequently sequenced populations. ClinVar is often believed, though not empirically demonstrated, to not be representative of the distribution of variants in a healthy population, except for an overrepresentation of pathogenic variants. As such, there is a possibility of fairness issues for the machine learning models, particularly those trained in a supervised manner. A selection bias-aware approach to estimating a fairness metric is therefore warranted. In contrast, variants in gnomAD come from the general population and can be considered to represent an unbiased distribution of variants. A

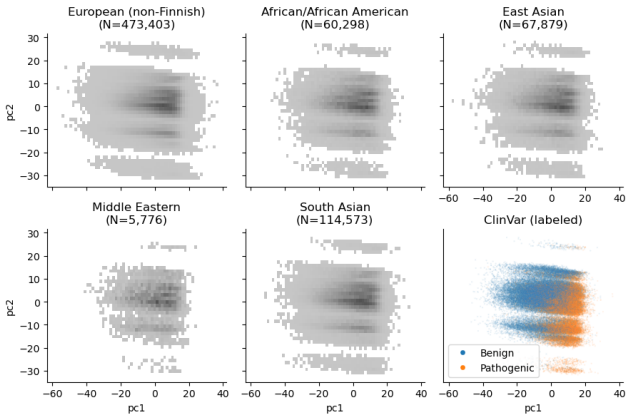

Figure 2: **Histogram of first two principal components of MutPred2 features for unlabeled genomic variants by population where the variant is observed (first 5 plots) and scatter plot of labeled variants in the ClinVar database (lower right plot).**

distinct feature of this problem is that, for privacy reasons, ClinVar data often omits the genetic ancestry of an affected individual, whereas gnomAD data contains both allele count and allele frequency in individuals of broad ancestral groups. The absence of group information in ClinVar makes it impossible to estimate a fairness metric using the standard approach, let alone mitigate bias in the estimation. However, our GNIL approach with MS-GMM-$\boldsymbol{\Theta}_2$ as an intermediate step, not only mitigates bias, but also makes fairness estimation feasible in the first place.

To investigate fairness in variant impact prediction, we extracted all 1-star or higher variants in ClinVar with a designated clinical significance of "benign", "benign/likely benign," "likely benign" (negative label) and "pathogenic", "pathogenic/likely pathogenic," "likely pathogenic" (positive label) to serve as the labeled set. We also extracted 3,061,576 variants from gnomAD together with their allele frequency by population, a measure of how often a variant has been observed in a population, and discarded variants with no allele frequency data available. Of the 1,458,756 (47.6%) remaining variants, we identified those that have been observed in only one population, leaving 851,704 (58.4%) population-unique variants, which serves as the unlabeled set. We then derived bias-mitigated fairness metric estimates comparing the most represented group, European (non-Finnish), to the variants observed in the other (under-represented) genetic ancestry groups with a sufficient number of observed variants: African/African American, East Asian, Middle Eastern and South Asian (Figure 2).

We evaluated fairness of the MutPred2 tool [35], one of the top performing predictors in the field [38, 45]. Its features and pathogenicity predictions were collected from the portal of the IGVF Consortium [18], which hosts MutPred2 data for all possible missense variants. MutPred2 is a supervised model, but it does not use allele frequencies as features, which is important as such models have been found to have unequal prediction accuracy across different allele frequency bins [38]. We used the clinical thresholds of 0.737, 0.829, and 0.932, the minimum scores required to qualify for supporting, moderate, and strong evidence towards pathogenicity, respectively [34]. Each threshold has direct significance for a

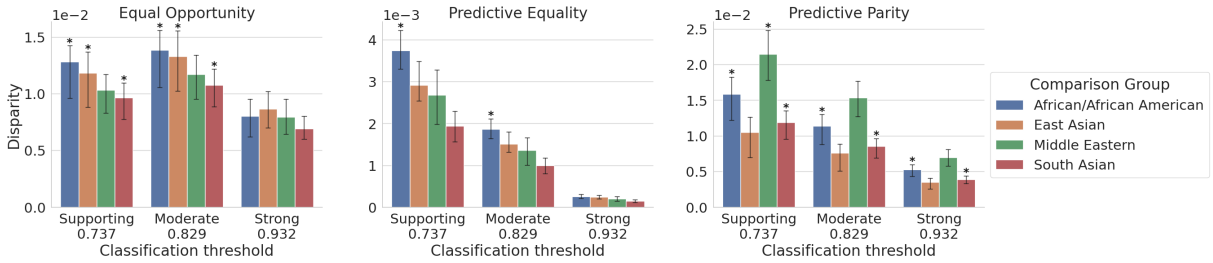

Figure 3: **Fairness metrics estimates were obtained as signed difference between the majority (European non-Finnish) and each under-represented group. Error bars show 68% confidence intervals based on 2000 bootstrap samples. Statistical significance of positive differences (indicated by *) was assessed at Bonferroni-adjusted significance level of 0.05/36, accounting for 36 multiple comparisons.**

variant's eligibility to be used in genetic diagnosis [39], but was here applied in a simpler manner to obtain pathogenic (positive) and benign (negative) prediction labels.

Using 8 components for each label (pathogenic and benign), MS-GMM-Θ_2 was applied to the ClinVar and gnomAD variants, represented with MutPred2 features after running Principal Component Analysis (PCA) and retaining the top 16 principal components. The estimated parameters were incorporated in the GNIL correction approach to obtain fairness estimates of EO, PE and PP.

7.2 Results

As a signed difference between the majority group and a minority group, each fairness metric attained a small positive value (<0.021) when computed using all data points in the two groups. Using 2000 bootstrap samples, 15 out of the 36 differences were positive with statistical significance at Bonferroni-corrected significance level of 0.05/36, as denoted by * in Figure 3. The maximum EO discrepancy of 0.01085 was observed between the European non-Finnish and the African/African American group (TPR = 0.5659 vs. TPR = 0.5551) for the moderate strength of evidence threshold. The maximum PE discrepancy of 0.005322 was observed between the European non-Finnish and African/African American group (FPR = 0.0683 vs. FPR = 0.0629) for the supporting strength of evidence threshold. The maximum PP discrepancy of 0.020301, was observed between the European non-Finnish and Middle Eastern group (PPV = 0.8402 vs. PPV = 0.8199) for the supporting strength of evidence threshold. Overall, the evidence of unfairness towards minority population groups in MutPred2 pathogenicity scores is minimal based on the small effect sizes, in spite of the statistical significance in some comparisons. Note further that our analysis was carried out only on population-unique variants, which constituted 58.4% of all variants with ancestry information.

Considering an aggregated analysis of all pairwise comparisons, we observe that TPR, FPR and PPV of the European non-Finish group are always higher than those of the minority groups. With respect to EO and PP this suggests a pattern of better prediction outcomes for European non-Finnish group. However, when PE is considered, this suggests better prediction outcomes for the minority groups, since higher FPR implies higher error. This seems to suggest rather than using a single classification threshold across groups, a group-specific threshold might reduce these discrepancies. A slightly lower threshold for minority groups could bring their TPR and FPR values closer to the European non-Finish group.

8 Discussion

In this work, we investigated the effect of selection bias on fairness metric estimation. Our synthetic data results (Figure 1) demonstrate the pernicious effect of selection bias in fairness estimation. The standard approach without corrections could have errors as high as 0.25-0.4 in some cases of selection bias. This could give vastly misleading conclusions on the fairness assessment of a model, potentially leading to an unfair model getting deployed. Selection bias should therefore be considered an important factor when estimating fairness and evaluating models for sensitive applications.

Fortunately, access to an unbiased, unlabeled dataset could address this issue effectively as demonstrated by our two bias mitigation approaches. Among these, GNIL (ignoring the group information in the labeled data) has a slight advantage over GIL both in terms of the estimation error and wider applicability to sensitive applications where the demographic information is excluded from labeled data. This key feature of the GNIL approach made it feasible to study the fairness of a widely used pathogenicity predictor, MutPred2, across ancestral groups. Applying our bias correction approaches, we find minimal evidence of unfairness in MutPred2's predictions, due to small effect sizes.

While our work introduces a framework to improve the estimation of fairness metrics under selection bias, we emphasize that fairness is a complex issue that cannot be fully resolved through algorithmic means alone. Metrics such as demographic parity or equalized odds capture only limited aspects of fairness and may not reflect its broader ethical and social dimensions. Furthermore, different fairness definitions can conflict with one another, and selecting an appropriate metric often involves normative judgments that lie beyond the scope of algorithmic decision-making. Nevertheless, improving the accuracy of fairness metric estimates contributes to a more informed and transparent evaluation of algorithmic impact in the real world.

9 Funding Information

This work was supported in part by the National Institutes of Health award U01HG012022 (PR).

References

[1] Ricardo Baeza-Yates. 2016. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science*. 1–1.

[2] Lingjiao Chen, Matei Zaharia, and James Y Zou. 2022. Estimating and explaining model performance when both covariates and labels shift. In *Advances in Neural Information Processing Systems*, Vol. 35. 11467–11479.

[3] Alexandra Chouldechova, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan. 2018. A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions. In *Conference on Fairness, Accountability and Transparency*. PMLR, 134–148.

[4] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2019. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems*, Vol. 32.

[5] M Clara De Paolis Kaluza, Shantanu Jain, and Predrag Radivojac. 2023. An approach to identifying and quantifying bias in biomedical data. In *Pacific Symposium on Biocomputing*, Vol. 28. 311–322.

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

[7] Ron Do, Daniel Balick, Heng Li, Ivan Adzhubei, Shamil Sunyaev, and David Reich. 2015. No evidence that selection has been less effective at removing deleterious mutations in Europeans than in Africans. *Nature Genetics* 47, 2 (2015), 126–131.

[8] Wei Du and Xintao Wu. 2021. Fair and robust classification under sample selection bias. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2999–3003.

[9] Laramie Duncan, Hanyang Shen, Bizu Gelaye, J Meijsen, K Ressler, M Feldman, R Peterson, and Ben Domingue. 2019. Analysis of polygenic risk score usage and performance in diverse human populations. *Nature Communications* 10, 1 (2019), 3328.

[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[11] Hadi Elzayn, Emily Black, Patrick Vossler, Nathanael Jo, Jacob Goldin, and Daniel E Ho. 2024. Estimating and implementing conventional fairness metrics with probabilistic protected features. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*. IEEE, 161–193.

[12] Willliam Feller. 2008. *An introduction to probability theory and its applications, vol 2*. John Wiley & Sons.

[13] Raymond Feng, Flavio Calmon, and Hao Wang. 2024. Adapting fairness interventions to missing values. In *Advances in Neural Information Processing Systems*, Vol. 36.

[14] Aaron Fraenkel. 2020. *Fairness and algorithmic decision making*. https://afraenkel.github.io/fairness-book/intro.html

[15] Saurabh Garg, Yifan Wu, Sivaraman Balakrishnan, and Zachary Lipton. 2020. A unified view of label shift estimation. In *Advances in Neural Information Processing Systems*, Vol. 33. 3290–3300.

[16] James Heckman. 1990. Varieties of selection bias. *The American Economic Review* 80, 2 (1990), 313–318.

[17] Ayanna Howard and Jason Borenstein. 2018. The ugly truth about ourselves and our robot creations: The problem of bias and social inequity. *Science and Engineering Ethics* 24, 5 (2018), 1521–1536.

[18] IGVF Consortium. 2024. Deciphering the impact of genomic variation on function. *Nature* 633, 8028 (2024), 47–57.

[19] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can I trust my fairness metric? Assessing fairness with unlabeled data and Bayesian inference. In *Advances in Neural Information Processing Systems*, Vol. 33. 18600–18612.

[20] Fanny Jourdan, Laurent Risser, Jean-Michel Loubes, and Nicholas Asher. 2023. Are fairness metric scores enough to assess discrimination biases in machine learning? *arXiv preprint arXiv:2306.05307* (2023).

[21] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, Daniel P Birnbaum, et al. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581, 7809 (2020), 434–443.

[22] Panagiotis Katsonis, Kevin Wilhelm, Amanda Williams, and Olivier Lichtarge. 2022. Genome interpretation using in silico predictors of variant impact. *Human Genetics* 141, 10 (2022), 1549–1577.

[23] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, Jeffrey Hoover, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Research* 44, D1 (2016), D862–D868.

[24] Zachary Lipton, Yu-Xiang Wang, and Alexander Smola. 2018. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning*. PMLR, 3122–3130.

[25] Diego Machado Reyes, Aritra Bose, Ehud Karavani, and Laxmi Parida. 2023. Fair-PRS: adjusting for admixed populations in polygenic risk scores using invariant risk minimization. In *Pacific Symposium on Biocomputing*, Vol. 28. 198–208.

[26] Arjun K Manrai, Birgit H Funke, Heidi L Rehm, Morten S Olesen, Bradley A Maron, Peter Szolovits, David M Margulies, Joseph Loscalzo, and Isaac S Kohane. 2016. Genetic misdiagnoses and the potential for health disparities. *New England Journal of Medicine* 375, 7 (2016), 655–665.

[27] Alicia R Martin, Masahiro Kanai, Yoichiro Kamatani, Yukinori Okada, Benjamin M Neale, and Mark J Daly. 2019. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nature Genetics* 51, 4 (2019), 584–591.

[28] Geoffrey J McLachlan and Thriyambakam Krishnan. 2008. *The EM algorithm and extensions*. John Wiley & Sons.

[29] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[30] Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *The 2018 ACM Conference on Fairness, Accountability, and Transparency*, Vol. 1170. 3.

[31] Osonde Osoba and W Welser. 2017. The risks of bias and errors in artificial intelligence. *Santa Monica: Rand Corporation* (2017). https://www.rand.org/pubs/research_reports/RR1744.html

[32] Tiago P Pagano, Rafael B Loureiro, Fernanda VN Lisboa, Rodrigo M Peixoto, Guilherme AS Guimarães, Gustavo OR Cruz, Maira M Araujo, Lucas L Santos, Marco AS Cruz, Ewerton LS Oliveira, et al. 2023. Bias and unfairness in machine learning models: A systematic review on datasets, tools, fairness metrics, and identification and mitigation methods. *Big Data and Cognitive Computing* 7, 1 (2023), 15.

[33] Ankit K Pathak, Nikita Bora, Mihaly Badonyi, Benjamin J Livesey, Joanne Ngeow, and Joseph A Marsh. 2024. Pervasive ancestry bias in variant effect predictors. *bioRxiv 2024.05.20.594987* (2024).

[34] Vikas Pejaver, Alicia B Byrne, Bing-Jian Feng, Kymberleigh A Pagel, Sean D Mooney, Rachel Karchin, Anne O'Donnell-Luria, Steven M Harrison, Sean V Tavtigian, Marc S Greenblatt, et al. 2022. Calibration of computational tools for missense variant pathogenicity classification and ClinGen recommendations for PP3/BP4 criteria. *The American Journal of Human Genetics* 109, 12 (2022), 2163–2177.

[35] Vikas Pejaver, Jorge Urresti, Jose Lugo-Martinez, Kymberleigh A Pagel, Guan Ning Lin, Hyun-Jun Nam, Matthew Mort, David N Cooper, Jonathan Sebat, Lilia M Iakoucheva, et al. 2020. Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. *Nature Communications* 11, 1 (2020), 5918.

[36] Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–44.

[37] Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. 2022. AI in health and medicine. *Nature Medicine* 28, 1 (2022), 31–38.

[38] Ruchir Rastogi, Ryan Chung, Sindy Li, Chang Li, Kyoungyeul Lee, Junwoo Woo, Dong-Wook Kim, Changwon Keum, Giulia Babbi, Pier Luigi Martelli, et al. 2025. Critical assessment of missense variant effect predictors on disease-relevant variant data. *Human Genetics* 144, 2 (2025), 281–293.

[39] Sue Richards, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W Grody, Madhuri Hegde, Elaine Lyon, Elaine Spector, et al. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in Medicine* 17, 5 (2015), 405–423.

[40] Marco Saerens, Patrice Latinne, and Christine Decaestecker. 2002. Adjusting the outputs of a classifier to new a priori probabilities: a simple procedure. *Neural Computation* 14, 1 (2002), 21–41.

[41] Jay Shendure, Gregory M Findlay, and Matthew W Snyder. 2019. Genomic medicine–progress, pitfalls, and promise. *Cell* 177, 1 (2019), 45–57.

[42] Amos Storkey. 2008. *When training and test sets are different: Characterizing learning transfer*. MIT Press, 3–28.

[43] Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. 2007. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research* 8, 5 (2007).

[44] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30.

[45] The Critical Assessment of Genome Interpretation Consortium. 2024. CAGI, the Critical Assessment of Genome Interpretation, establishes progress and prospects for computational genetic variant interpretation methods. *Genome Biology* 25, 1 (2024), 53.

[46] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. 1–7.

[47] Slobodan Vucetic and Zoran Obradovic. 2001. Classification on data with biased class distribution. In *Machine Learning: ECML 2001: 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings 12*. Springer, 527–538.

[48] Yanchen Wang and Lisa Singh. 2021. Analyzing the impact of missing values and selection bias on fairness. *International Journal of Data Science and Analytics* 12, 2 (2021), 101–119.

[49] Christopher Winship and Robert D Mare. 1992. Models for sample selection bias. *Annual Review of Sociology* 18, 1 (1992), 327–350.

[50] Daniel Zeiberg, Shantanu Jain, and Predrag Radivojac. 2023. Leveraging structure for improved classification of grouped biased data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11113–11120.

Appendix

A TPR and FPR Bias Correction Proof

$$\begin{aligned}
\textbf{GIL}: \quad \mathrm{TPR}_g &= \mathbb{E}[\hat{y}(x)|+,g] \\
&= \int_x \hat{y}(x) p(x|+,g)dx \\
&= \int_x \hat{y}(x) \frac{p(x|+,g)}{\tilde{p}(x|+,g)} \tilde{p}(x|+,g)dx \\
&= \tilde{\mathbb{E}}\left[\hat{y}(x) \frac{p(x|+,g)}{\tilde{p}(x|+,g)} \middle| +, g \right] \\
&\overset{est}{=} \frac{1}{|\mathcal{L}_g^+|} \sum_{x \in \mathcal{L}_g^+} \hat{y}(x) \psi_g^+(x),
\end{aligned}$$

$$\begin{aligned}
\textbf{GNIL}: \quad \mathrm{TPR}_g &= \mathbb{E}[\hat{y}(x)|+,g] \\
&= \int_x \hat{y}(x) p(x|+,g)dx \\
&= \int_x \hat{y}(x) \frac{p(x|+,g)}{\tilde{p}(x|+)} \tilde{p}(x|+)dx \\
&= \tilde{\mathbb{E}}\left[\hat{y}(x) \frac{p(x|+,g)}{\tilde{p}(x|+)} \middle| + \right] \\
&\overset{est}{=} \frac{1}{|\mathcal{L}^+|} \sum_{x \in \mathcal{L}^+} \hat{y}(x) \chi_g^+(x),
\end{aligned}$$

where $\mathbb{E}$ and $\tilde{\mathbb{E}}$ denote expectations w.r.t. the unbiased and biased distributions, respectively. Correction for FPR_g follows similarly.

B Bias-Corrected Expected Calibration Error

We provide the bias formula to correctly measure probabilistic calibration of a continuous classifier score function $s : \mathbb{R}^D \to [0,1]$ for group g. Let $\{b_i\}_{i=1}^B$ be B bins (non overlapping intervals) partitioning $[0,1]$. Theoretically, Expected Calibration Error (ECE) for group g is given by

$$\begin{aligned}
\mathrm{ECE}_g(s) &= \sum_{i=1}^B p(s(x) \in b_i|g) \Big| p\big(+ | s(x) \in b_i, g\big) - \mathbf{E}\big(s(x) | s(x) \in b_i, g\big) \Big| \\
&= \sum_{i=1}^B p(s(x) \in b_i|g) \left| \frac{\alpha_g p\big(s(x) \in b_i|+,g\big)}{p\big(s(x) \in b_i|g\big)} - \int_x s(x) p\big(x|s(x) \in b_i, g\big)dx \right| \\
&= \sum_{i=1}^B \left| \alpha_g p\big(s(x) \in b_i|+,g\big) - \int_x s(x) p\big(s(x) \in b_i|x,g\big) p(x|g)dx \right| \\
&= \sum_{i=1}^B \left| \alpha_g \int_{s(x) \in b_i} p(x|+,g)dx - \int_{s(x) \in b_i} s(x) p(x|g)dx \right| \\
&\overset{\text{GIL}}{=} \sum_{i=1}^B \left| \alpha_g \int_{s(x) \in b_i} \psi_g^+(x) \tilde{p}(x|+,g)dx - \int_{s(x) \in b_i} s(x) \xi_g(x) \tilde{p}(x|g)dx \right| \\
&\overset{\text{GNIL}}{=} \sum_{i=1}^B \left| \alpha_g \int_{s(x) \in b_i} \chi_g^+(x) \tilde{p}(x|+)dx - \int_{s(x) \in b_i} s(x) v_g(x) \tilde{p}(x)dx \right|,
\end{aligned}$$

where $\xi_g(x) = \frac{p(x|g)}{\tilde{p}(x|g)} = \frac{\alpha_g \mathrm{NMix}(x;\theta_g^+) + (1-\alpha_g)\mathrm{NMix}(x;\theta_g^-)}{\tilde{\alpha}_g \mathrm{NMix}(x;\tilde{\theta}_g^+) + (1-\tilde{\alpha}_g)\mathrm{NMix}(x;\tilde{\theta}_g^-)}$ and $v_g(x) = \frac{p(x|g)}{\tilde{p}(x)} = \frac{\alpha_g \mathrm{NMix}(x;\theta_g^+) + (1-\alpha_g)\mathrm{NMix}(x;\theta_g^-)}{\tilde{\alpha}\mathrm{NMix}(x;\tilde{\theta}^+) + (1-\tilde{\alpha})\mathrm{NMix}(x;\tilde{\theta}^-)}$; $\psi_g^+(x)$ and $\chi_g^+(x)$ are as defined in Table 1. The above expressions lead to the following empirical estimates of ECE_g

$$\textbf{GIL}: \mathrm{ECE}_g \overset{est}{=} \sum_{i=1}^B \left| \frac{\alpha_g}{|\mathcal{L}_g^+|} \sum_{x \in \mathcal{L}_{gi}^+} \psi_g^+(x) - \frac{1}{|\mathcal{L}_g|} \sum_{x \in \mathcal{L}_{gi}} s(x) \xi_g(x) \right|$$

$$\textbf{GNIL}: \mathrm{ECE}_g \overset{est}{=} \sum_{i=1}^B \left| \frac{\alpha_g}{|\mathcal{L}^+|} \sum_{x \in \mathcal{L}_{\_i}^+} \chi_g^+(x) - \frac{1}{|\mathcal{L}|} \sum_{x \in \mathcal{L}_{\_i}} s(x) v_g(x) \right|,$$

where $\mathcal{L}_{\_i}$ and $\mathcal{L}_{\_i}^+$ are the labeled points and the positive labeled points with scores in bin b_i, respectively; $\mathcal{L}_{gi}$ and $\mathcal{L}_{gi}^+$ are the labeled points and the positive labeled points from group g with scores in bin b_i, respectively. $\mathcal{L}$, $\mathcal{L}^+$, $\mathcal{L}_g$ and $\mathcal{L}_g^+$ are as defined in Section 4.2.

C EM parameter updates

We use w_{gk}^+ and w_{gk}^- instead of w_{gk} and v_{gk}, respectively. We also use α_g^+ and α_g^- for α_g and $1-\alpha_g$, respectively. Let X (input variable), Y (class label: + or −), G (group index) and K (component index) be the random variables. Their joint probability is given by

$$\begin{aligned}
p(X,Y,G,K) &= \prod_{g \in \mathcal{G}} \prod_{y \in \mathcal{Y}} (\alpha_g^y)^{\mathbb{I}[Y=y]\mathbb{I}[G=g]} \\
&\times \prod_{g \in \mathcal{G}} \prod_{y \in \mathcal{Y}} \prod_{k \in \mathcal{K}^y} (w_{gk}^y)^{\mathbb{I}[G=g]\mathbb{I}[Y=y]\mathbb{I}[K=k]} \\
&\times \prod_{y \in \mathcal{Y}} \prod_{k \in \mathcal{K}^y} (\phi_k^y(X))^{\mathbb{I}[Y=y]\mathbb{I}[K=k]}
\end{aligned}$$

Next, taking log

$$\begin{aligned}
\log p(X,Y,G,K) &= \sum_{g \in \mathcal{G}} \sum_{y \in \mathcal{Y}} \mathbb{I}[Y=y]\mathbb{I}[G=g] \log \alpha_g^y \\
&+ \sum_{g \in \mathcal{G}} \sum_{y \in \mathcal{Y}} \sum_{k \in \mathcal{K}^y} \mathbb{I}[G=g]\mathbb{I}[Y=y]\mathbb{I}[K=k] \log w_{gk}^y \\
&+ \sum_{y \in \mathcal{Y}} \sum_{k \in \mathcal{K}^y} \mathbb{I}[Y=y]\mathbb{I}[K=k] \log(\phi_k^y(X))
\end{aligned}$$

For the unlabeled set the E-step (both MS-GMM-$\boldsymbol{\Theta}_1$ and MS-GMM-$\boldsymbol{\Theta}_2$) corresponds to conditioning on X and G and taking expectation using the old parameters.

$$\begin{aligned}
q(x,g) &= \mathbb{E}_{\check{\Theta}_1}[\log p(X,Y,G,K)|X=x, G=g] \\
&= \sum_{y \in \mathcal{Y}} \log \alpha_g^y \sum_{k \in \mathcal{K}^y} (\check{\pi}_{gk}^y) + \sum_{y \in \mathcal{Y}} \sum_{k \in \mathcal{K}^y} \check{\pi}_{gk}^y \log w_{gk}^y \\
&+ \sum_{y \in \mathcal{Y}} \sum_{k \in \mathcal{K}^y} \check{\pi}_{gk}^y \log(\phi_k^y(X))
\end{aligned}$$

Following the same steps, $\log \tilde{p}(X,Y,G,K)$ can be derived. For MS-GMM-$\boldsymbol{\Theta}_1$, the E-step for the labeled set corresponds to conditioning on X, Y and G and taking expectation using the old parameters.

$$\begin{aligned}
\tilde{q}(x,y,g) &= \mathbb{E}_{\check{\Theta}_1}[\log \tilde{p}(X,Y,G,K)|X=x, Y=y, G=g] \\
&= \log \tilde{\alpha}_g^y + \sum_{k \in \mathcal{K}^y} \check{\rho}_{gk}^y \log \tilde{w}_{gk}^y + \sum_{k \in \mathcal{K}^y} \check{\rho}_{gk}^y \log(\phi_k^y(X))
\end{aligned}$$

For MS-GMM-$\boldsymbol{\Theta}_2$, we consider the labeled data distribution without groups

$$\tilde{p}(X,Y,K) = \prod_{y \in \mathcal{Y}} (\tilde{\alpha}^y)^{\mathbb{I}[Y=y]} \times \prod_{y \in \mathcal{Y}} \prod_{k \in \mathcal{K}^y} (\tilde{w}_k^y \phi_k^y(X))^{\mathbb{I}[Y=y]\mathbb{I}[K=k]}$$

	Equal size groups		Minority group underrepresented	
Dim	$\lambda = 0.5$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.1$
2	0.832	0.828	0.917	0.889
8	0.956	0.947	0.995	0.971

Table 2: Fraction of runs for which MS-GMM-Θ_1 obtains better unbiased group class-conditional estimates compared to MB-GC estimates for datasets described in Section 6.1. Each cell represents an evaluation over 945 simulated datasets.

Next, taking log

$$\log \tilde{p}(X, Y, K) = \sum_{y \in \mathcal{Y}} \mathbb{I}[Y = y] \log \tilde{\alpha}^y + \sum_{y \in \mathcal{Y}} \sum_{k \in \mathcal{K}^y} \mathbb{I}[Y = y]\mathbb{I}[K = k]\Big(\log \tilde{w}_k^y + \log \phi_k^y(X)\Big).$$

For MS-GMM-$\mathbf{\Theta}_2$, the E-step for the labeled data corresponds to conditioning on X and Y and taking expectation using the old parameters.

$$\begin{aligned}\tilde{q}(x, y) &= \mathbb{E}_{\check{\Theta}_2}[\log \tilde{p}(X, Y, K)|X = x, Y = y] \\ &= \log \tilde{\alpha}^y + \sum_{k \in \mathcal{K}^y} \check{\rho}_k^y \log \tilde{w}_k^y + \sum_{k \in \mathcal{K}^y} \check{\rho}_k^y \log(\phi_k^y(X))\end{aligned}$$

The Q-functions for MS-GMM-$\mathbf{\Theta}_1$ and MS-GMM-$\mathbf{\Theta}_2$ are

$$Q(\Theta_1|\check{\Theta}_1) = \frac{1}{|\mathcal{U}| + |\mathcal{L}|}\Bigg[\sum_{(x,g) \in \mathcal{U}} q(x, g) + \sum_{(x,g,y) \in \mathcal{L}} \tilde{q}(x, y, g)\Bigg]$$

$$Q(\Theta_2|\check{\Theta}_2) = \frac{1}{|\mathcal{U}| + |\mathcal{L}|}\Bigg[\sum_{(x,g) \in \mathcal{U}} q(x, g) + \sum_{(x,y) \in \mathcal{L}} \tilde{q}(x, y)\Bigg],$$

respectively. Taking partial derivative of the Q-function w.r.t. μ_k^*, Σ_k^*, α_g^+ (substituting $\alpha_g^- = 1 - \alpha_g^+$), equating to 0 gives their parameter update. For w_{kg}^*, $\tilde{w}_{kg}^*$ (MS-GMM-$\mathbf{\Theta}_1$) and $\tilde{w}_k^*$ (MS-GMM-$\mathbf{\Theta}_2$) the update equations are derived via KKT condition for satisfying 1) $w_{kg}^* \geq 0$, $\tilde{w}_{kg}^* \geq 0$ and $\tilde{w}_k^* \geq 0$, and 2) $\sum_{k \in \mathcal{K}^*} w_{kg}^* = 1$, $\sum_{k \in \mathcal{K}^*} \tilde{w}_{kg}^* = 1$ and $\sum_{k \in \mathcal{K}^*} \tilde{w}_k^* = 1$, respectively.

D Selecting the number of components

We provide the log-likelihood criteria to select the number of components via cross-validation for MS-GMM-$\mathbf{\Theta}_1$ and MS-GMM-$\mathbf{\Theta}_2$ as follows

$$\frac{1}{|\mathcal{L}| + |\mathcal{U}|}\Bigg[\sum_{(x,g,y) \in \mathcal{L}} \log \text{NMix}\Big(x; \tilde{\theta}_g^y\Big) + \sum_{(x,g) \in \mathcal{U}} \log\Big(\alpha_g \text{NMix}\Big(x; \theta_g^+\Big) + \alpha_g^- \text{NMix}\Big(x; \theta_g^-\Big)\Big)\Bigg]$$

$$\frac{1}{|\mathcal{L}| + |\mathcal{U}|}\Bigg[\sum_{(x,y) \in \mathcal{L}} \log \text{NMix}\Big(x; \tilde{\theta}^y\Big) + \sum_{(x,g) \in \mathcal{U}} \log\Big(\alpha_g \text{NMix}\Big(x; \theta_g^+\Big) + \alpha_g^- \text{NMix}\Big(x; \theta_g^-\Big)\Big)\Bigg],$$

respectively; $\tilde{\theta}^*$, $\tilde{\theta}_g^*$, θ_g^* and $\text{NMix}(x; \theta)$ are defined in Table 1.

E Implementation Details

E.1 Parameter Initialization. MS-GMM-$\mathbf{\Theta}_1$ and MS-GMM-$\mathbf{\Theta}_2$, are sensitive to parameter initialization. We initialize parameters by first fitting the standard GMM on the labeled positives and negatives, separately. For each group, the biased component weights in $\tilde{\boldsymbol{w}}_g$ and $\tilde{\boldsymbol{v}}_g$ ($\tilde{\boldsymbol{w}}$ and $\tilde{\boldsymbol{v}}$ for MS-GMM-$\mathbf{\Theta}_2$) are set to be equal. Each unlabeled point is assigned to the nearest component mean from the fitted means $\boldsymbol{\mu}$. Class priors α_g and unbiased component weights $\boldsymbol{w}_g$ ($\boldsymbol{v}_g$) are computed as proportions of assigned unlabeled points to corresponding classes and components. Component covariance matrices are also initialized using the unlabeled point assignments.

E.2 Optimization Details. While not required by the optimization, the implementation is simplified by enforcing that no component is empty. If any component weight is updated to zero, we add a small $\epsilon = 1e - 300$ to all weights for that mixture and normalize the weights by dividing by their sum. Similarly, when covariance matrices Σ_k^* become close to singular, we add ϵ to the diagonal until the matrix is invertible.

E.3 Baseline Implementation. We get results for the methods of [19] by using the work's provided code at https://github.com/disiji/bayesian-fairness-assess. The Beta Calibration (BC) method is applied using the same Multi-layer Perceptron model as used for scores in our proposed methods. The entire labeled set is used for the BC metric estimation.

F Improved Estimation with SCC-invariance Compared to MB-GC

Our proposed sample selection bias model, **SCC**-invariance, extends MB-GC to account for groups. It makes an assumption that the group class-conditionals are related as mixtures sharing the same components. The proposed algorithms, MS-GMM-$\mathbf{\Theta}_1$ and MS-GMM-$\mathbf{\Theta}_2$, exploit this structure by pooling the data from the groups, in addition to pooling the labeled and unlabeled data, to estimate the shared components (Section 4.2). Applying the MB-GC based multi-sample gaussian mixture estimation algorithm [5], independently to the data from each group, fails to capture the shared structure between the groups and leads to suboptimal performance in estimating the unbiased group class-conditionals (Table 2). We compared MS-GMM-$\mathbf{\Theta}_1$ with the MB-GC based estimation on our simulated data using the following metric,

$$\sum_{g \in \mathcal{G}} \sum_{y \in \mathcal{Y}} \text{AUC}\Big(\text{NMix}\Big(\cdot; \hat{\theta}_g^y\Big), \text{NMix}\Big(\cdot; \theta_g^y\Big)\Big),$$

where each AUC term measures the difference between a pair of estimated and true group class-conditional, by training a classifier to separate the the data generated from the two distributions. A lower value of the AUC based metric indicate better estimation. In Table 2, we present the fraction of runs for which MS-GMM-$\mathbf{\Theta}_1$ achieves a lower AUC across various configurations of dimensions, number of components, AUC(+,-), selection bias buckets, labeled data proportion (λ) and equal and unequal group representation, as described in Section 6.1.