Joint Fairness Model with Applications to Risk Predictions for Under-represented Populations

Hyungrok Do[1]   Shinjini Nandi[2]   Preston Putzel[3]
Padhraic Smyth[3]   Judy Zhong[1*]

[1]Department of Population Health, New York University Grossman School of Medicine, New York, NY 10016, USA
[2]Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717, USA
[3]Department of Computer Science, University of California, Irvine, CA 92697, USA

May 24, 2021

arXiv:2105.04648v3 [stat.AP] 21 May 2021

Abstract

Under-representation of certain populations, based on gender, race/ethnicity, and age, in data collection for predictive modeling may yield less-accurate predictions for the under-represented groups. Recently, this issue of fairness in predictions has attracted significant attention, as data-driven models are increasingly utilized to perform crucial decision-making tasks. Methods to achieve fairness in the machine learning literature typically build a single prediction model subject to some fairness criteria in a manner that encourages fair prediction performances for all groups. These approaches have two major limitations: i) fairness is often achieved by compromising accuracy for some groups; ii) the underlying relationship between dependent and independent variables may not be the same across groups. We propose a Joint Fairness Model (JFM) approach for binary outcomes that estimates group-specific classifiers using a joint modeling objective function that incorporates fairness criteria for prediction. We introduce an Accelerated Smoothing Proximal Gradient Algorithm to solve the convex objective function, and demonstrate the properties of the proposed JFM estimates. Next, we presented the key asymptotic properties for the JFM parameter estimates. We examined the efficacy of the JFM approach in achieving prediction performances and parities, in comparison with the Single Fairness Model, group-separate model, and group-ignorant model through extensive simulations. Finally, we demonstrated the utility of the JFM method in the motivating example to obtain fair risk predictions for under-represented older patients diagnosed with coronavirus disease 2019 (COVID-19).

Keywords: algorithmic fairness, algorithmic bias, joint estimation, under-represented population

2

1 Introduction

1.1 Applied Context

The issue of making fair predictions has attracted significant attention recently in machine learning as a critical issue in the application of data-driven models. Though machine learning models are increasingly utilized to perform crucial decision-making tasks, recent evidence reveals that many carefully designed algorithms learn biases from the underlying data and exploit these inequalities when making predictions. For example, large systematic biases in prediction performance have been detected for machine learning models in areas such as recidivism prediction relative to race [Angwin et al., 2016], ranking of job candidates relative to gender [Lahoti et al., 2018] and face recognition relative to both race and gender [Ryu et al., 2018, Buolamwini and Gebru, 2018]. There is an emerging recognition that such biases are also likely to be a significant issue in data-derived predictive models in healthcare [Char et al., 2018]. Data obtained through clinical trials are often biased and not representative of racial/ethnic minority groups and/or people over 75 with multiple chronic conditions [Gianfrancesco et al., 2018], a phenomenon which has appeared in studies of cancer incidence and mortality [Murthy et al., 2004], cardiovascular diseases [Sardar et al., 2014] and diabetes [Chow et al., 2012], etc. Biased representation of different populations in biomedical studies limits the benefits that can be potentially achieved for these communities.

One motivating example is to predict mortality for patients infected with coronavirus disease 2019 (COVID-19). As of January 23 2021, COVID-19 has infected more than 96 million people globally, accounting for more than 2 million known deaths. Older patients are particularly vulnerable to severe outcomes and death due to COVID-19. The Centers for Disease Control and Prevention (CDC) reported that the fatality rate was 18.8% for

patients older than 80 years whereas the overall fatality rate is estimated at up to 5% for all patients [Kompaniyets et al., 2021]. This difference in survival highlights an urgent need for risk stratification of older patients with COVID-19 based on routine clinical assessments. However, most COVID-19 studies have not been stratified by age groups [Tehrani et al., 2021]. Thus, when a risk prediction equation generated from the general population was applied to older patients with COVID-19, the model predicted high-risk scores overall due to their older age, higher prevalence of comorbidities and more laboratory abnormalities. This resulted in insufficient and unfair risk stratification for these patients as not all older patients are at the same risk of death from COVID-19 [Tehrani et al., 2021].

1.2   Existing Approaches

Methods to address fairness in the machine learning literature typically begin with a formal probabilistic definition of fairness. In the context of risk prediction, predictive fairness at the group level means that a risk prediction model has performance characteristics (based on accuracy, ranking, calibration) that are relatively independent of group memberships. For example, if the false positive rate for a classification model is defined as $P(\hat{y} = 1|y = 0)$, where $\hat{y}$ is the model's prediction, then enforcing equality with respect to a particular binary group indicator variable G can be stated as requiring the two predictive distributions $P(\hat{y} = 1|G = 1, y = 0)$ and $P(\hat{y} = 1|G = 0, y = 0)$ to be as close as possible. Other definitions include demographic parity [Calders et al., 2009], equalized odds or equal opportunities [Hardt et al., 2016], disparate treatment, impact and mistreatment [Zafar et al., 2019, 2017a] etc. It is recognized that there is no unique optimal way to define fairness, leading to trade-offs between different approaches [Zafar et al., 2017b].

Given a fairness criterion, the second component of a fairness strategy requires an algorithmic approach, typically consisting of either 1) pre-processing the data by mapping the

training data to a transformed space where the dependencies between sensitive attributes and class labels disappear [Kamiran and Calders, 2012, Dwork et al., 2018]; or 2) post-processing of a trained prediction model to modify the probability of the decision being positive from an existing classifier to limit unfair discrimination [Kamishima et al., 2012, Hardt et al., 2016]; or 3) "in-process," where fairness is accounted for during training of a model, e.g., by adding a fairness constraint to the objective function during training. Zemel et al. [2013] proposed to learning a fair representation of the data and classifier parameters by optimizing a non-convex function. Zafar et al. [2017b] further defined a convex function as a measure of (un)fairness, and suggested optimizing accuracy subject to the convex fairness constraints as well as their converse.

A key feature of nearly all existing approaches is that a single set of classifier parameters is estimated, using fairness criteria that encourage fair prediction performance across all groups. This approach has two main limitations: i) fairness is often achieved by compromising accuracy of some groups; ii) the underlying relationship between dependent and independent variables may not be the same across group, and the differences in predictive features may be of interest. In the example of predicting mortality risk for patients with COVID-19, while one would expect some features to have the same association with mortality for both older and younger patients, the associations between mortality and other features may be different between age groups. For instance, overweight and obesity (Body Mass Index [BMI] $> 25 kg/m^2$) increase the risk for COVID-19 associated mortality, particularly among adults aged < 65 years [Kompaniyets et al., 2021] However, geriatric BMI guidelines are different from younger adults. For older adults, higher BMIs are often associated with greater energy stores and a better nutritional state overall, which is beneficial for patients' survival outcomes when serious infections are developed. Estimating separate prediction models for each group does not leverage potential similarities between the

5

groups. Moreover, estimating a single prediction model, even with the fairness criteria, will likely result in sub-optimal estimation or prediction performances for one group in order to achieve fair performances with one set of parameters shared across groups. Danaher et al. [2014] proposed the joint graphical lasso method, a technique for jointly estimating multiple models corresponding to distinct but related conditions. Their approach is based upon a penalized log-likelihood approach, which penalizes the differences between parameter estimates across groups. Penalized log-likelihood approaches have often been used by other authors like Yuan and Lin [2007], Friedman et al. [2007b] etc. for similar estimation purposes while minimizing the disparities in estimates across groups. In all such cases, however, prediction performances are not considered.

In this paper, we propose a Joint Fairness Model, a technique for jointly estimating multiple prediction models corresponding to distinct but related groups, to achieve fair prediction performances across groups. The model parameters are estimated by encouraging prediction fairness, while simultaneously ensuring high predictive accuracy irrespective of the heterogeneity across the groups. The rest of this paper is organized as follows. In Section 2, we present the proposed joint fairness model. Section 3 describes the algorithm to find its optimal solution, and discusses hyperparameter selection. In Section 4, we discuss asymptotic consistency of the estimators. We illustrate the performance of our proposal in simulation studies in Section 5; and an application to the motivating example of predicting COVID-19 mortality outcomes for patients of different age groups in Section 6. Section 7 extends the proposed joint fairness model to generalized linear models for other types of outcomes. Finally, we summarize and discuss our findings in Section 8.

2 Problem Formulation

For binary outcomes, consider we are given K groups of datasets $S^k = \{(\mathbf{X}_i^k, y_i^k) \in \mathbb{R}^p \times \{0,1\} : i = 1, \cdots, n^k\}$ with $K \geq 2$ representing group membership. Assuming that the $n = \sum_{k=1}^{K} n^k$ observations are independently distributed: $y_i^k \sim \text{Bernoulli}(\text{p}_i^k)$, $\hat{y}_i^k : \mathbb{R}^p \to \{0,1\}$ is the predicted value based on predictor features $\mathbf{X}_i^k$. We focus on the development of the fair prediction approach for the widely-used logistic regression model. The log-likelihood of the logistic model for the data from all groups takes the form

$$\sum_{k=1}^{K} \ell(\boldsymbol{\beta}^k; \mathbf{X}^k, \mathbf{y}^k) = \sum_{k=1}^{K} \sum_{i=1}^{n^k} \left(y_i^k \mathbf{X}_i^k \boldsymbol{\beta}^k - \log\left(1 + \exp\left(\mathbf{X}_i^k \boldsymbol{\beta}^k\right)\right)\right). \tag{1}$$

Define $\boldsymbol{\beta} = (\boldsymbol{\beta}^1 \ldots \boldsymbol{\beta}^K) \in \mathbb{R}^{pK}$. Maximizing the likelihood function (1) with respect to $\boldsymbol{\beta}^k$ in each group separately yields the maximum likelihood estimates $\hat{\boldsymbol{\beta}}^k$ of group k, thus making separate predictions $\hat{y}^k$ per group. If we ignore group memberships, $\hat{\boldsymbol{\beta}}$ can be estimated by maximizing the likelihood function in equation (1) setting all $\boldsymbol{\beta}^k$ equal to a single global parameter vector $\hat{\boldsymbol{\beta}}$ and making predictions $\hat{y}$ per individual (irrespective of group) using that parameter vector.

If the K datasets correspond to observations collected from K distinct but related groups, then one might wish to borrow strength across the K groups to estimate $\boldsymbol{\beta}$ and predict $\hat{y}$, rather than estimating parameters $\boldsymbol{\beta}^k$ for each group separately, or estimating one set of $\boldsymbol{\beta}^k$ for all k which can lead to heterogeneous prediction performance across the groups. Therefore, instead of estimating $\boldsymbol{\beta}$ by maximizing the likelihood in equation (1), we consider a penalized log-likelihood approach and seek to jointly estimate $\boldsymbol{\beta}$ by solving an objective function of $\sum_{k=1}^{K} \ell(\boldsymbol{\beta}^k; \mathbf{X}^k, \mathbf{y}^k)$ in equation (1) subject to constraints on (i) fairness, $\mathcal{P}_{\text{F}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\text{F}})$ (ii) parameter similarity, $\mathcal{P}_{\text{Sim}}(\boldsymbol{\beta}; \lambda_{\text{Sim}})$, and (iii) parameter

sparsity, $\mathcal{P}_{\mathrm{Sp}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sp}})$.

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\; F(\boldsymbol{\beta}) = -\sum_k \frac{1}{n_k} \ell(\boldsymbol{\beta}^k; \mathbf{X}^k, \mathbf{y}^k) + \mathcal{P}_{\mathrm{F}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) + \mathcal{P}_{\mathrm{Sim}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sim}}) + \mathcal{P}_{\mathrm{Sp}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sp}}).$$

(2)

We propose choosing a fairness penalty function $\mathcal{P}_{\mathrm{F}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}})$ that encourages each group to have similar predictive performance. In this work, we use equalized odds [Hardt et al., 2016] which encourages each group to have similar false positive rates (FPRs) and false negative rates (FNRs). Thus, we want to minimize the absolute difference between FPR^j and FPR^k $|P(\hat{y} = 1|G = j, y = 0) - P(\hat{y} = 1|G = k, y = 0)|$, and that between FNR^j and FNR^k: $|P(\hat{y} = 0|G = j, y = 1) - P(\hat{y} = 0|G = k, y = 1)|$.

Under the logistic regression model, $|P(\hat{y} = 1|G = j, y = 0) - P(\hat{y} = 1|G = k, y = 0)| = \left| \mathbb{E}\left[\frac{\exp(\mathbf{X}\boldsymbol{\beta}^j)}{1+\exp(\mathbf{X}\boldsymbol{\beta}^j)} \Big| G = j, y = 0 \right] - \mathbb{E}\left[\frac{\exp(\mathbf{X}\boldsymbol{\beta}^k)}{1+\exp(\mathbf{X}\boldsymbol{\beta}^k)} \Big| G = k, y = 0 \right] \right|$ which is nonconvex due to the nonconvexity of the sigmoid function. We instead minimize the absolute difference of the expected linear components of the two groups $\left| \mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}^j \big| G = j, y = 0 \right] - \mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}^k \big| G = k, y = 0 \right] \right|$. The inequality below, which follows from a first order Taylor series approximation of the sigmoid function, guarantees that minimizing the difference of the linear components results in minimizing the difference of the FPRs:

$$\left| \mathbb{E}\left[\frac{\exp(\mathbf{X}\boldsymbol{\beta}^j)}{1 + \exp(\mathbf{X}\boldsymbol{\beta}^j)} \Big| G = j, y = 0 \right] - \mathbb{E}\left[\frac{\exp(\mathbf{X}\boldsymbol{\beta}^k)}{1 + \exp(\mathbf{X}\boldsymbol{\beta}^k)} \Big| G = k, y = 0 \right] \right|$$
$$\leq \left| \mathbb{E}\left[\frac{1}{2} + \frac{\mathbf{X}\boldsymbol{\beta}^j}{4} \Big| G = j, y = 0 \right] - \mathbb{E}\left[\frac{1}{2} + \frac{\mathbf{X}\boldsymbol{\beta}^k}{4} \Big| G = k, y = 0 \right] \right|.$$

Similar approximation can be used for the absolute difference between FNR^j and FNR^k. Note that the empirical estimate of the expectation is

$$\mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}^k \big| G = k, y = y \right] = \frac{1}{|S_y^k|} \sum_{i \in S_y^k} \mathbf{X}_i \boldsymbol{\beta}^k,$$

8

where $S_y^k = \{(\mathbf{X}_i, y_i) : G_i = k, y_i = y\}$ is a subgroup defined by group k and the true response value y with $y \in \{0, 1\}$. Thus, our fairness penalty to bridge the between-group gaps in the linear components of FPR^k and FNR^k is defined as:

$$\mathcal{P}_{\mathrm{F}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) = \mathcal{P}_{\mathrm{FPR}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) + \mathcal{P}_{\mathrm{FNR}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) \tag{3}$$

$$= \lambda_{\mathrm{F}} \sum_{j<k} \left| \frac{1}{|S_0^j|} \sum_{i \in S_0^j} \mathbf{X}_i \boldsymbol{\beta}^j - \frac{1}{|S_0^k|} \sum_{i \in S_0^k} \mathbf{X}_i \boldsymbol{\beta}^k \right| + \lambda_{\mathrm{F}} \sum_{j<k} \left| \frac{1}{|S_1^j|} \sum_{i \in S_1^j} \mathbf{X}_i \boldsymbol{\beta}^j - \frac{1}{|S_1^k|} \sum_{i \in S_1^k} \mathbf{X}_i \boldsymbol{\beta}^k \right|$$

where the summation $\sum_{j<k}$ represents $\sum_{k=1}^{K} \sum_{j=1}^{k-1}$ for the simplicity.

The similarity penalty $\mathcal{P}_{\mathrm{Sim}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sim}})$ is chosen to encourage similarity across the K estimated parameters. Here we use the generalized fused Lasso penalty [Hoefling, 2010, Danaher et al., 2014, Dondelinger et al., 2018] defined as

$$\mathcal{P}_{\mathrm{Sim}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sim}}) = \lambda_{\mathrm{Sim}} \sum_{j<k} \|\boldsymbol{\beta}^j - \boldsymbol{\beta}^k\|_1. \tag{4}$$

The sparsity penalty $\mathcal{P}_{\mathrm{Sp}}(\boldsymbol{\beta})$ is chosen to encourage sparse estimates and to avoid ill-defined maximum likelihood estimates when $n^k < p$.

$$\mathcal{P}_{\mathrm{Sp}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sp}}) = \sum_k \lambda_{\mathrm{Sp}_k} \|\boldsymbol{\beta}^k\|_1. \tag{5}$$

In the three penalty functions, λ_{F}, λ_{Sim}, and λ_{Sp} are nonnegative hyperparameters. Here $\mathcal{P}_{\mathrm{Sp}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sp}})$, $\mathcal{P}_{\mathrm{F}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}})$, and $\mathcal{P}_{\mathrm{Sim}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sim}})$ are convex penalty functions, so that the objective in equation (2) is convex in $\boldsymbol{\beta}$. The proposed model jointly estimates $\boldsymbol{\beta}$ to achieve fair performances across groups, herein referred to as the Joint Fairness Model (JFM). In contrast, the dominant approach for fair predictions in the current literature is to estimate a single set of $\boldsymbol{\beta}$ parameters with constraints on quality of performance metrics across groups [Bechavod and Ligett, 2017].

Penalty functions in (3), (4), and (5) are based on the *L1* norm. They can be flexibly adapted to *L2* penalization or a combination of *L1* and *L2* penalizations. The difference between *L1* and *L2* penalties have been well discussed [Tibshirani, 1996, Zou and Hastie, 2005]. For the fairness penalty, Bechavod and Ligett [2017] showed that there are no remarkable differences in the empirical performances between *L1* and *L2* fairness penalty forms. When we use the *L2* form of the similarity penalty, it penalizes large differences more aggressively so that models have less chance to obtain group-specific estimates. Note that other formats of the similarity penalty can be used in the JFM framework. For example, the group Lasso penalty [Yuan and Lin, 2006] has been shown to encourage similar sparsity patterns across groups [Obozinski et al., 2010, Danaher et al., 2014], while the fused lasso term is more aggressive in encouraging similar $\hat{\boldsymbol{\beta}}^k$ estimates.

3 Accelerated Smoothing Proximal Gradient Algorithm for JFM

In this section, we introduce an Accelerated Smoothing Proximal Gradient (ASPG) Algorithm [Chen et al., 2012] to solve the optimization problem (2) for JFM. The objective function of (2) is convex in $\boldsymbol{\beta}$ so that a global optimal solution can be attained. However, conventional proximal gradient-based or coordinate descent approaches (generally used for Lasso-like methods) cannot be directly applied to solve Problem (2) because there is no closed form solution for a proximal operator associated with $\mathcal{P}_{\text{FPR}}$ and $\mathcal{P}_{\text{FNR}}$.

3.1 Nesterov smooth approximation

To overcome the difficulty originating from the non-differentiability of the fairness and similarity penalties, we decouple the terms into a linear combination of the decision variables via the dual norm, then apply the Nesterov smoothing approximation [Nesterov, 2005]. We start with matrix representations of the fairness penalty terms $\mathcal{P}_{\mathrm{FPR}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) = \lambda_{\mathrm{F}} \|\mathbf{D}_0 \boldsymbol{\beta}\|_1$ and $\mathcal{P}_{\mathrm{FNR}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) = \lambda_{\mathrm{F}} \|\mathbf{D}_1 \boldsymbol{\beta}\|_1$, where $\mathbf{D}_y \in \mathbb{R}^{K(K-1)/2 \times pK}$ is defined as below. Similarly, the matrix representation of the similarity penalty $\mathcal{P}_{\mathrm{Sim}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sim}}) = \lambda_{\mathrm{Sim}} \|\mathbf{F} \boldsymbol{\beta}\|_1$ with $\mathbf{F}$ defined as below.

$$\mathbf{D}_y = \begin{pmatrix} \bar{\mathbf{X}}_y^1 & -\bar{\mathbf{X}}_y^2 & \mathbf{0} & \cdots & \mathbf{0} \\ & & \vdots & & \\ \mathbf{0} & \bar{\mathbf{X}}_y^2 & -\bar{\mathbf{X}}_y^3 & \cdots & \mathbf{0} \\ & & \vdots & & \end{pmatrix} \quad \mathbf{F} = \begin{pmatrix} \mathbf{I}_p & -\mathbf{I}_p & \mathbf{0} & \cdots & \mathbf{0} \\ & & \vdots & & \\ \mathbf{0} & \mathbf{I}_p & -\mathbf{I}_p & \cdots & \mathbf{0} \\ & & \vdots & & \end{pmatrix}$$

Here, $\bar{\mathbf{X}}_y^j = \frac{1}{|S_y^j|} \sum_{\mathbf{X}^j \in S_y^j} \mathbf{X}^j$ is the average logit vector for group j with outcome y, $\mathbf{I}_p$ is the p-dimensional identity matrix. The single matrix form of the fairness penalty term and the similarity penalty term is therefore defined as:

$$\mathcal{P}_{\mathrm{F}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) + \mathcal{P}_{\mathrm{Sim}}(\boldsymbol{\beta}; \lambda_{\mathrm{Sim}}) = \left\| \begin{pmatrix} \lambda_{\mathrm{F}} \mathbf{D}_0 \\ \lambda_{\mathrm{F}} \mathbf{D}_1 \\ \lambda_{\mathrm{Sim}} \mathbf{F} \end{pmatrix} \boldsymbol{\beta} \right\|_1 = \|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta}\|_1.$$

Thus, the objective function (2) can be written in matrix form:

$$\underset{\boldsymbol{\beta}}{\mathrm{minimize}} \; -\sum_k \ell(\boldsymbol{\beta}^k; \mathbf{X}^k, \mathbf{y}^k) + \|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta}\|_1 + \sum_k \lambda_{\mathrm{Sp}_k} \|\boldsymbol{\beta}^k\|_1, \qquad (6)$$

where the associated proximal operator of $\|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta}\|_1$ does not have a closed form solution. We apply the Nesterov smooth approximation to approximate $\|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta}\|_1$ by a smooth function $f_\mu(\boldsymbol{\beta})$. Since the dual norm of the *L1* norm is the L_∞ norm, we have

$$\|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta}\|_1 = \sup\{\boldsymbol{\alpha}^T \mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta} : \|\boldsymbol{\alpha}\|_\infty \le 1\},$$

and thus, for $\mu > 0$, Nesterov smooth approximation of $\|\mathbf{D}_{\lambda_\mathrm{F},\lambda_\mathrm{Sim}}\boldsymbol{\beta}\|_1$ is

$$f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim}) = \sup\left\{\boldsymbol{\alpha}^T\mathbf{D}_{\lambda_\mathrm{F},\lambda_\mathrm{Sim}}\boldsymbol{\beta} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 : \|\boldsymbol{\alpha}\|_\infty \le 1\right\}. \tag{7}$$

The following proposition provides the maximum gap between $\|\mathbf{D}_{\lambda_\mathrm{F},\lambda_\mathrm{Sim}}\boldsymbol{\beta}\|_1$ and its Nesterov approximation $f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim})$.

Proposition 3.1 *For any $\mu > 0$, the Nesterov smooth approximation satisfies the following inequalities:*

$$0 \le \|\mathbf{D}_{\lambda_\mathrm{F},\lambda_\mathrm{Sim}}\boldsymbol{\beta}\|_1 - f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim}) \le \frac{\mu p K}{2}.$$

Proof: See Supplementary Material S.2.

The proposition implies that we can control the upper bound of the approximation error by manipulating μ. We can achieve an arbitrary accuracy δ by letting $\mu = \frac{2\delta}{pK}$.

The next proposition dictates that the gradient $\nabla f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim})$ has a simple form and is thus easy to compute.

Proposition 3.2 *For any $\mu > 0$, $f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim})$ is smooth and convex with respect to $\boldsymbol{\beta}$, whose gradient takes the following form:*

$$\nabla f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim}) = \mathbf{D}_{\lambda_\mathrm{F},\lambda_\mathrm{Sim}}^T\boldsymbol{\alpha}^*, \tag{8}$$

where $\boldsymbol{\alpha}^* = \mathrm{argmax}\left\{\boldsymbol{\alpha}^T\mathbf{D}_{\lambda_\mathrm{F},\lambda_\mathrm{Sim}}\boldsymbol{\beta} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 : \|\boldsymbol{\alpha}\|_\infty \le 1\right\}$. Moreover, the gradient is Lipschitz continuous with the Lipschitz constant $L_\mu = \mu^{-1}\|\mathbf{D}_{\lambda_\mathrm{F},\lambda_\mathrm{Sim}}\|_2^2$, where $\|\cdot\|_2$ denotes the matrix spectral norm (which is equivalent to the largest singular value of the matrix).

Proof: See Supplementary Material S.3.

Computational Remark: Matrix multiplication $\mathbf{D}_{\lambda_\mathrm{F}\lambda_\mathrm{Sim}}^T\boldsymbol{\alpha}^*$ requires $\mathcal{O}(p^2K^3)$ operations, thus making it computationally intensive when p is large. However, $\mathbf{D}_{\lambda_\mathrm{F}\lambda_\mathrm{Sim}}^T\boldsymbol{\alpha}^*$ can be

computed efficiently without matrix multiplication. Because of its special structure, its computation can be substituted by a series of scalar multiplications and vector additions. We can reduce the complexity to $\mathcal{O}(pK^3)$. Details are provided in Supplementary Material S.1.

The following proposition yields to attain $\boldsymbol{\alpha}^*$ in Proposition 3.2, which is essential to compute the gradient $\nabla f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim})$.

Proposition 3.3 *For any $\mu > 0$, we have*

$$\boldsymbol{\alpha}^* = S_\infty \left(\mu^{-1} \mathbf{D}_{\lambda_\mathrm{F}, \lambda_\mathrm{Sim}} \boldsymbol{\beta} \right),$$

where $S_\infty(\cdot)$ is the projection onto the unit L_∞ ball, which is defined by

$$[S_\infty(\boldsymbol{x})]_i = \begin{cases} x_i & \text{if } x_i \in [-1, 1] \\ 1 & \text{if } x_i \in (1, \infty) \\ -1 & \text{if } x_i \in (-\infty, -1) \end{cases}.$$

Proof: See Supplementary Material S.4.

Computational Remark: The matrix multiplication $\mathbf{D}_{\lambda_\mathrm{F}, \lambda_\mathrm{Sim}} \boldsymbol{\beta}$ is computationally expensive as well. It requires $\mathcal{O}(p^2 K^3)$ operations, however, we can simplify it to $\mathcal{O}(pK^2)$ by performing a series of vector subtractions. The details are presented in Supplementary Material S.1.

3.2 Accelerated Smoothing Proximal Gradient Algorithm

With $\|\mathbf{D}_{\lambda_\mathrm{F}, \lambda_\mathrm{Sim}} \boldsymbol{\beta}\|_1$ substituted by the Nesterov smooth approximation $f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim})$, Problem (6) becomes

$$\underset{\boldsymbol{\beta}}{\text{minimize }} \tilde{F}(\boldsymbol{\beta}) = -\sum_k \ell(\boldsymbol{\beta}^k; \mathbf{X}^k, \mathbf{y}^k) + f_\mu(\boldsymbol{\beta}; \lambda_\mathrm{F}, \lambda_\mathrm{Sim}) + \sum_k \lambda_{\mathrm{Sp}_k} \|\boldsymbol{\beta}^k\|_1, \qquad (9)$$

13

whose first two terms are convex smooth functions. Although the sparsity penalty term $\sum_k \lambda_{\mathrm{Sp}_k} \|\boldsymbol{\beta}^k\|_1$ is non-differentiable, it can be managed through the proximal gradient method using the soft-thresholding operator $\mathcal{S}$ with a closed form solution [Friedman et al., 2007a].

Algorithm 1 presents the proposed ASPG algorithm, starting from parameter initialization, to gradient descent iterations with proximal and momentum steps, until convergence. The gradient descent step tries to improve the current solution $\boldsymbol{\gamma}^{(t-1)}$ by using the gradients $\nabla \ell$ of the log-likelihood and ∇f_μ of function (8). Subsequently, it performs a proximal step for the sparsity penalty. Finally, a momentum-based update is performed to accelerate the convergence. Specifically, we adopted the momentum coefficients in the fast iterative shrinkage thresholding algorithm [Beck and Teboulle, 2009].

Although Algorithm 1 minimizes the Nesterov smooth approximation $\tilde{F}(\boldsymbol{\beta})$ instead of the original objective function $F(\boldsymbol{\beta})$ in equation (2), it can be proven that the solution is sufficiently close to the optimal solution of equation (2). We first present a lemma demonstrating a convergence property of the algorithm.

Lemma 3.1 *Let* $\{\boldsymbol{\beta}^{(t)} : t = 1, 2, \cdots\}$ *be a sequence generated by Algorithm 1. Then for any* $t \geq 1$,

$$\tilde{F}(\boldsymbol{\beta}^{(t)}) - \tilde{F}(\boldsymbol{\beta}^*) \leq \frac{2L\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2^2}{t^2},$$

where $\boldsymbol{\beta}^*$ *is a global minimizer of Problem* (9).

Proof: Proof of this theorem is analogous to the proof of Theorem 4.4 in Beck and Teboulle [2009] because $-\sum_k \ell(\boldsymbol{\beta}^{(k)}; \mathbf{X}^k, \mathbf{y}^k) + f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}})$ is a convex differentiable function and it has Lipschitz continuous gradient with Lipschitz constant

$$L = \frac{1}{4} \max \left\{ \lambda_{\max}(\mathbf{X}^{kT}\mathbf{X}^k) : k = 1, \cdots, K \right\} + \mu^{-1}\|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}}\|_2^2 > 0,$$

14

Algorithm 1 Accelerated Smoothing Proximal Gradient (ASPG) Algorithm for the JFM

1: **Input:** Data $\mathbf{X}^k, \mathbf{y}^k$ for $k = 1 \ldots K$, hyperparameters $\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}, \lambda_{\mathrm{Sp}}, \epsilon, \mu$

2: **Output:** $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}^1, \cdots, \hat{\boldsymbol{\beta}}^K)$ solving the joint fairness objective function (2).

3: **Initialize:** $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, $\boldsymbol{\gamma}^{(0)} = \mathbf{0}$, $s^{(0)} = 1$

4: $L = \frac{1}{4} \max \left\{ \lambda_{\max}(\mathbf{X}^{kT}\mathbf{X}^k) : k = 1, \cdots, K \right\} + \mu^{-1} \|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}}\|_2^2$

5: **for** $t \geq 1$ **do**

6: $\qquad \boldsymbol{\alpha}^{(t)} = \boldsymbol{\gamma}^{(t-1)} - L^{-1} \left(-\nabla \ell \left(\boldsymbol{\gamma}^{(t-1)} \right) + \nabla f_\mu \left(\boldsymbol{\gamma}^{(t-1)} \right) \right)$

7: $\qquad \boldsymbol{\beta}^{(t)} = \mathcal{S} \left(\boldsymbol{\alpha}^{(t)}; L^{-1}\lambda_{\mathrm{Sp}} \right)$

8: $\qquad$ **if** $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2 \leq \epsilon$ **break**

9: $\qquad s^{(t)} = \frac{1+\sqrt{1+4s^{(t-1)2}}}{2}$

10: $\qquad \boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t)} + \left(\frac{s^{(t-1)}-1}{s^{(t)}} \right) \left(\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)} \right)$

11: $\qquad t \leftarrow t + 1$

12: **end for**

13: $\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(t)}$.

where $\lambda_{\max}(\mathbf{A})$ denotes the largest eigenvalue of $\mathbf{A}$.

Based on the lemma, we establish a theorem that shows the solution provided by Algorithm 1 can be arbitrarily close to the global optimum of Problem (2).

Theorem 3.1 *Let* $\{\boldsymbol{\beta}^{(t)} : t = 1, 2, \cdots\}$ *be a sequence generated by Algorithm 1. Then for any* $t \geq 1$,

$$F(\boldsymbol{\beta}^{(t)}) - F(\boldsymbol{\beta}^{**}) \leq \frac{\mu p K}{2} + \frac{2L\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2^2}{t^2},$$

where $\boldsymbol{\beta}^*$ *and* $\boldsymbol{\beta}^{**}$ *are global minimizers of Problem (9) and Problem (2), respectively, and* L *is the Lipschitz constant of* $\tilde{F}$ *presented in Lemma 3.1.*

Proof: We can easily verify the inequality by applying Proposition 3.1 and Lemma 3.1, and

using $\tilde{F}(\boldsymbol{\beta}^*) \leq \tilde{F}(\boldsymbol{\beta}^{**})$ as below:

$$F(\boldsymbol{\beta}^{(t)}) - F(\boldsymbol{\beta}^{**}) = \left(F(\boldsymbol{\beta}^{(t)}) - \tilde{F}(\boldsymbol{\beta}^{(t)})\right) + \left(\tilde{F}(\boldsymbol{\beta}^{(t)}) - \tilde{F}(\boldsymbol{\beta}^*)\right) + \left(\tilde{F}(\boldsymbol{\beta}^*) - F(\boldsymbol{\beta}^{**})\right)$$

$$\leq \frac{\mu p K}{2} + \frac{2L\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2^2}{t^2} + 0.$$

Given the desired accuracy $\delta > 0$ for the approximation, we set $\mu = \frac{2\delta}{pK}$. Then, we have $F(\boldsymbol{\beta}^{(t)}) - F(\boldsymbol{\beta}^{**}) \leq \delta + \frac{2L\|\boldsymbol{\beta}^0 - \boldsymbol{\beta}^*\|_2^2}{t^2}$. This inequality implies that the accuracy of Algorithm 1 both depends on the number of iterations t and the accuracy $\delta > 0$ for the approximation. Based on the theorem, we present the rate of convergence of the algorithm in the following proposition.

Proposition 3.4 *Given a desired accuracy $\varepsilon > 0$, rate of convergence of Algorithm 1 is $\mathcal{O}\left(\sqrt{\frac{pK}{\delta(\varepsilon - \delta)}}\right)$. Note that $\delta > 0$ must be smaller than ε.*

Proof: See Supplementary Material S.5.

Proposition 3.5 *Time complexity of a single iteration of Algorithm 1 is $\mathcal{O}((n + K^2)pK)$.*

Proof: Computing the gradient $\nabla \sum_k \ell(\boldsymbol{\beta}^k)$ of the sum of the log-likelihood functions requires $\mathcal{O}(npK)$. Computing $\nabla f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}})$ requires $\mathcal{O}(pK^3)$. Thus, the gradient step requires $\mathcal{O}((n + K^2)pK)$ operations. The proximal step and momentum step both require $\mathcal{O}(pK)$, which are dominated by the complexity of the gradient step. Therefore, a single iteration of Algorithm 1 requires $\mathcal{O}((n + K^2)pK)$ operations.

4 Asymptotic properties of the JFM estimates

We now present the key asymptotic results for the JFM parameter estimates $\hat{\boldsymbol{\beta}}$ for each group by solving objective function (2) of a logistic regression for a binary outcome when

16

$K = 2$. We assume p remains constant and n increases to infinity. Consider the following assumptions

Assumption 1. $\mathcal{I}(\boldsymbol{\beta}^k)/n^k \to \mathbf{C}^k$, where $\mathbf{C}^k$ is a positive definite $p \times p$ matrix, for $k = 1$ and 2, where $\mathcal{I}(\boldsymbol{\beta}^k)$ is the information matrix of size $p \times p$. For simplicity, we assume there are no intercept terms in $\boldsymbol{\beta}^k$.

Assumption 2. As $n = \min\limits_{k=1,2} n^k \to \infty$, $\max_{\hat{\boldsymbol{\beta}}^k} \left\| \left(\mathcal{I}(\hat{\boldsymbol{\beta}}^k)^{-\frac{1}{2}} \right) \mathcal{I}(\boldsymbol{\beta}^k) \left(\mathcal{I}(\hat{\boldsymbol{\beta}}^k)^{-\frac{1}{2}} \right)^T - \mathbf{I}_p \right\|_2 \to 0$

where $\mathcal{I}(\hat{\boldsymbol{\beta}}^k)$ is the empirical information matrix, and $\mathbf{I}_p$ is a $p \times p$ identity matrix.

The following theorem proves $\sqrt{n}$-consistency for the estimators, complying with the fairness and similarity constraints between the two groups as well as the sparsity constraint.

Theorem 4.1 *Let $\hat{\boldsymbol{\beta}}^k$ for $k = 1$ and 2, minimize the loss function (2). If $\lambda_F^{(n)}/\sqrt{n} \to \lambda_F^{(0)} \geq 0$, $\lambda_{\text{Sim}}^{(n)}/\sqrt{n} \to \lambda_{\text{Sim}}^{(0)} \geq 0$, and $\lambda_{\text{Sp}}^{(n)}/\sqrt{n} \to \lambda_{\text{Sp}}^{(0)} \geq 0$, then under the assumptions 1 and 2*

$$\sqrt{n} \left(\hat{\boldsymbol{\beta}}^k - \boldsymbol{\beta}^k \right) \to \hat{\mathbf{u}}^k \tag{10}$$

where $\{\hat{\mathbf{u}}^1, \hat{\mathbf{u}}^2\} = \operatorname{argmin}(\mathcal{V})$, for $\mathbf{u}^k = (u_1^k, \ldots, u_p^k) \in \mathbb{R}^p$,

$$\mathcal{V}(\mathbf{u}^1, \mathbf{u}^2) = \mathbf{u}^{1T}\mathbf{W}^1 + \mathbf{u}^{2T}\mathbf{W}^2 + \frac{1}{2}\mathbf{u}^{1T}\mathbf{C}^1\mathbf{u}^1 + \frac{1}{2}\mathbf{u}^{2T}\mathbf{C}^2\mathbf{u}^2 +$$

$$\lambda_F^{(0)} \left[(\bar{\mathbf{X}}_0^1\mathbf{u}^1 - \bar{\mathbf{X}}_0^2\mathbf{u}^2)\operatorname{sign}(\bar{\mathbf{X}}_0^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_0^2\boldsymbol{\beta}^2)\mathbb{I}(\bar{\mathbf{X}}_0^1\boldsymbol{\beta}^1 \neq \bar{\mathbf{X}}_0^2\boldsymbol{\beta}^2) + |\bar{\mathbf{X}}_0^1\mathbf{u}^1 - \bar{\mathbf{X}}_0^2\mathbf{u}^2|\mathbb{I}(\bar{\mathbf{X}}_0^1\boldsymbol{\beta}^1 = \bar{\mathbf{X}}_0^2\boldsymbol{\beta}^2) +$$

$$(\bar{\mathbf{X}}_1^1\mathbf{u}^1 - \bar{\mathbf{X}}_1^2\mathbf{u}^2)\operatorname{sign}(\bar{\mathbf{X}}_1^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_1^2\boldsymbol{\beta}^2)\mathbb{I}(\bar{\mathbf{X}}_1^1\boldsymbol{\beta}^1 \neq \bar{\mathbf{X}}_1^2\boldsymbol{\beta}^2) + |\bar{\mathbf{X}}_1^1\mathbf{u}^1 - \bar{\mathbf{X}}_1^2\mathbf{u}^2|\mathbb{I}(\bar{\mathbf{X}}_1^1\boldsymbol{\beta}^1 = \bar{\mathbf{X}}_1^2\boldsymbol{\beta}^2) \right] +$$

$$\lambda_{\text{Sp}}^{(0)} \sum_{k=1}^{2} \sum_{j=1}^{p} \{u_j^k \operatorname{sign}(\beta_j^k)\mathbb{I}(\beta_j^k \neq 0) + |u_j^k|\mathbb{I}(\beta_j^k = 0)\} +$$

$$\lambda_{\text{Sim}}^{(0)} \sum_{j=1}^{p} \{(u_j^1 - u_j^2)\text{sign}(\beta_j^1 - \beta_j^2)\mathbb{I}(\beta_j^1 \neq \beta_j^2) + |u_j^1 - u_j^2|\mathbb{I}(\beta_j^1 = \beta_j^2)\}$$

Here $\mathbf{W}^k \sim \mathcal{N}_p(\mathbf{0}, \mathbf{C}^k)$, *where* $\mathbf{C}^k = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i^k \mathbf{X}_i^{kT}$, *and* $\frac{1}{|S_y^k|} \sum_{i \in S_y^k} \mathbf{X}_i = \bar{\mathbf{X}}_y^k$, *for* $y = 0, 1$ *and* $k = 1, 2$.

Proof: See Supplementary Material S.6.

5 Simulation Study

We performed a series of simulations to evaluate the proposed JFM, and compared it with the approaches of a group-separate individual logistic regression model, a group-ignorant vanilla logistic regression model, and a Single Fairness Model (SFM) proposed by Bechavod and Ligett [2017]. In the context of logistic regression, such an SFM minimizes the following objective function. We also established ASPG for SFM (see S.8.)

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; -\sum_k \ell(\boldsymbol{\beta}; \mathbf{X}_k, \mathbf{y}_k) + \lambda_{\text{F}} \left\{ \sum_{j<k} \left| \frac{\sum_{\mathbf{X}^j \in S_0^j} \mathbf{X}^j \boldsymbol{\beta}}{|S_0^j|} - \frac{\sum_{\mathbf{X}^k \in S_0^k} \mathbf{X}^k \boldsymbol{\beta}}{|S_0^k|} \right| \right.$$
$$\left. + \sum_{j<k} \left| \frac{\sum_{\mathbf{X}^j \in S_1^j} \mathbf{X}^j \boldsymbol{\beta}}{|S_1^j|} - \frac{\sum_{\mathbf{X}^k \in S_1^k} \mathbf{X}^k \boldsymbol{\beta}}{|S_1^k|} \right| \right\} + \lambda_{\text{Sp}} \|\boldsymbol{\beta}\|_1.$$

When applying the group-separate model, regression coefficients were estimated for each group separately with *L1* penalty. The group-ignorant model estimated one logistic regression with group membership as an additional covariate with an *L1* penalty.

5.1 Simulation Setup

We consider a two-group problem ($K = 2$) for simplicity with group 1 as the over-represented group and group 2 as the under-represented group with respect to the sample

sizes. The training samples were simulated as follows. The predictor matrix $\mathbf{X}^k$ was independently generated from a standard normal distribution. The binary outcome y_i^k was then simulated from Bernoulli$(\pi_i(\mathbf{x}_i^k))$, where $\pi_i(\mathbf{x}_i^k) = \frac{\exp(\mathbf{x}_i^k \boldsymbol{\beta}^k)}{1+\exp(\mathbf{x}_i^k \boldsymbol{\beta}^k)}$. Out of the total number of features, 40% in each group had non-zero coefficients (β's). The non-zero coefficients were each set to the value 3. The simulations were conducted under four scenarios to investigate performances at various levels of shared parameters, sample sizes and dimensionalities.

- In Scenario 1, the shared features between the two groups ranged from 0% to 100% of features with non-zero coefficients. The intercepts were selected so that the baseline event prevalence were at 10% for each group. The sample sizes were set at 500 and 200 for group 1 and 2 respectively. The number of features were set to $p = 100$.

- In Scenario 2, the baseline prevalence of the under-represented group (group 2) ranged from 10% to 50%. The baseline event prevalence of the over-represented group (group 1) was fixed at 50%. Half of the features with non-zero coefficients were shared between the groups, while the other half of the features were group-specific. The sample sizes were set at 500 and 200 for group 1 and 2 respectively. The number of features was set to $p = 100$.

- In Scenario 3, the sample size of the under-represented group (group 2) ranged from 50 to 300 while the sample size of group 1 was fixed at 500. The number of features were set to $p = 100$. Half of the features with non-zero coefficients were shared between the groups.

- In Scenario 4, the number of features p ranged from 50 to 2,000 in order to investigate model performance in high-dimensional settings. Sample sizes were 500 and 200 for group 1 and 2 respectively. For each value of p, 40 features had non-zero coefficients, with half of the non-zero features shared between the two groups.

19

We evaluated the methods on independent testing datasets with large sample sizes ($n = 1000$ for both groups) under the same simulation setups. The Area under the Receiver Operating Characteristic curve (AUC) was used to assess the predictive ability of each model. Prediction unfairness was assessed by the group difference in AUCs. Medians and interquartile ranges (IQRs) of the assessment metrics were generated from 20 replicates for each experiment. Predictive performances and their unfairness in terms of FPR and FNR were calculated with cutoff of the predicted probability at 0.5 and presented in Supplementary Material S.10. We further presented additional simulation scenarios in Supplementary Material S.11.

5.2 Choice of the Evaluation Metrics in Selecting Hyperparameters in Cross-validations

The group-ignorant model, group-separate model, SFM, and JFM contain 1, K, 2, and $K + 2$ hyperparameters respectively. For every method, 5-fold cross-validation on the training dataset was used to determine the hyperparameters. For the vanilla models (group-separate and group-ignorant), the lasso penalty term was selected by optimizing cross-validation AUCs. For the fairness-aware models, we compared a series of evaluation metrics for selecting the hyperparameters in cross-validations, including group average of AUCs/accuracies (arithmetic mean, geometric mean, and harmonic mean), overall AUCs/accuracies on all samples ignoring group memberships, and the group average of AUCs/accuracies subtracting the disparity of AUCs/accuracies (absolute differences and squared differences) in Supplementary Materials S.9. The harmonic mean of group-wise AUCs in cross-validations selected the hyperparameters generating the most robust AUCs and parities in the test datasets, therefore was used in the following simulations results.

5.3   Simulation Results

For Scenario 1, Figure 1(a) displays the estimated AUC for the under-represented group against the proportion of shared features in the two groups. The AUCs of the under-represented group from the JFM, SFM, and group-ignorant models improved as the proportion of shared features increased. The SFM and group-ignorant models were highly sensitive to the percentage of shared nonzero features as they both estimate a single set of parameters for both groups. In contrast, JFM showed consistently higher AUC than the other three methods. When the proportion of shared features is high, JFM estimated higher AUCs and smaller variances than those from the group-separate model. JFM's performance was similar to those of the SFM and the group-ignorant model. When the proportion of shared features is low, JFM estimated higher AUCs than the SFM and the group-ignorant model, and showed similar AUC to the group-separate model. Figure 1(b) displays the estimated AUC for the majority group against the proportion of shared features in the two groups. JFM was robust in achieving comparable AUC to that from the group-separate model. The SFM and group-ignorant models were highly sensitive to the percentage of shared features for the majority group with lower AUCs when the proportion of shared parameters is low. Figure 1(c) displays the estimated overall AUCs, and Figure 1(d) displays the group disparity of AUCs from the four approaches. Together, these figures demonstrate that the JFM achieves fair prediction performances robustly across the range of varying proportions of shared features between groups, by training the classifiers jointly with a flexible parameterization. Figure S.5(a) through Figure S.5(d) compares the average of TPR and TNR and disparity in TPR and TNR differences of the four methods. The patterns are similar to those found using AUCs.

Figure 2 displays the performance of the four methods when varying the baseline event prevalence of the under-represented group while holding the prevalence of the majority

21

group fixed. In Figure 2(a), the JFM showed consistently higher AUCs for the under-represented group than those from all the other models. The AUCs estimated from the group-separate method showed higher variance when the prevalence is rare. Figure 2(b) indicates that the AUC of the over-represented group was not impacted for the JFM and group-separate methods, remaining consistently higher than those from the SFM and the group-ignorant models. As seen in Figure 2(c) and 2(d), the JFM achieves overall satisfactory AUCs and parity between groups with varying sample sizes of the under-represented group. Figure S.6(a) through Figure S.6(d) compares the average of TPR and TNR and disparity in TPR and TNR differences of the four methods.

Figure 3 displays the performance of the four methods against the sample size of the under-represented group with other settings fixed. In Figure 3(a), the AUCs of the under-represented group from all models were improved as its sample size increased. The JFM showed consistently higher AUCs and smaller variances than those from all the other models. JFM outperforms the other models the most when the minority group's sample size is small, showing the benefits of borrowing information between groups in situations with unbalanced sample sizes. Figure 3(b) illustrates that the AUC of majority group was not impacted for the JFM and group-separate methods. However, the AUC of the majority group decreased as sample size of the under-represented group increased for the SFM and the group-ignorant models. This decrease highlights an undesirable performance from these two methods, namely, compromising accuracy by estimating a single set of classifier parameters. Figure 3(c) and 3(d) illustrates that the JFM achieves overall satisfactory AUCs and parity between groups across varying sample sizes of the under-represented group. Figure S.7 compares the average of TPR and TNR and disparity of TPR and TNR of the four methods.

Figure 4 displays the performance of the four methods while varying the number of

features from 200 to 2000, and holding the number of associated features constant at 40. It demonstrates that the JFM method in going from low dimensional to high dimensional settings can maintain overall satisfactory prediction performances and parity between groups. Supplementary Figure S.12 displays the performance of the four methods while varying the number of features from 200 to 2000, and setting the number of associated features to a fixed proportion of the total number of features. The resultant patterns are similar to Figure 4.

We investigated the empirical computational complexity of JFM with the increasing number of features and sample sizes in the Supplementary Materials. Figure S.1 shows that the JFM computation time is approximately $\mathcal{O}(p^{1.5})$ and $\mathcal{O}(n)$. Details are presented in Section S.7.

6 COVID-19 Risk Prediction Case Study

We applied the JFM, in comparison with other methods, to predict mortality related to COVID-19 from patients' routine ambulatory encounters and laboratory records prior to COVID-19 infection, with the goal of better stratifying patient risk for clinical management. We used a retrospective EHR dataset of 11,594 patients of age 50+ with laboratory-confirmed COVID-19 at New York University Langone Health (NYULH) from March 2020 to February 2021. Among the 11,594 patients, 1,242 (10.7%) died of COVID-19. The patients were divided into four groups by their age at the time of COVID-19 diagnosis: 50-64, 65-74, 75-84, and 85+ with $5,905$ (50.9%), $2,946$ (25.4%), $1,814$ (15.6%), and 929 (8.0%) patients, respectively. The observed mortality rates were 4.44%, 11.17%, 18.96% and 33.05%, respectively. Candidate features ($p = 82$) included demographic variables, such as age, sex, race/ethnicity, smoking status, body mass index (BMI); common chronic
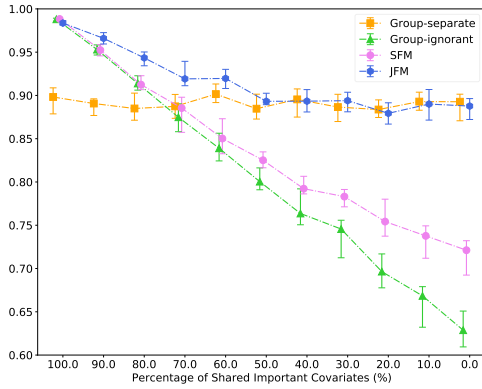
(a) AUC of the Under-represented Group
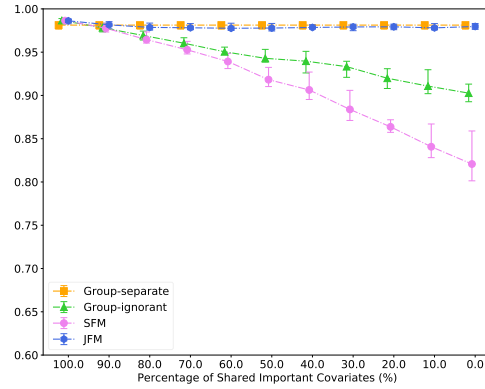
(b) AUC of the Over-represented Group
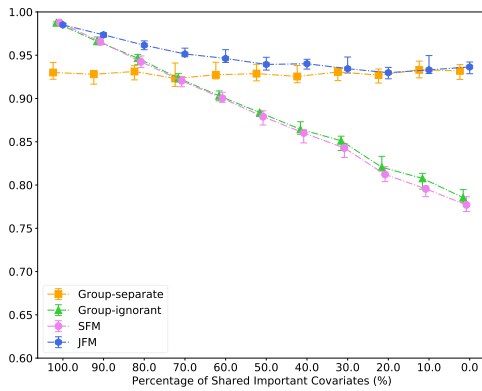
(c) Overall AUC
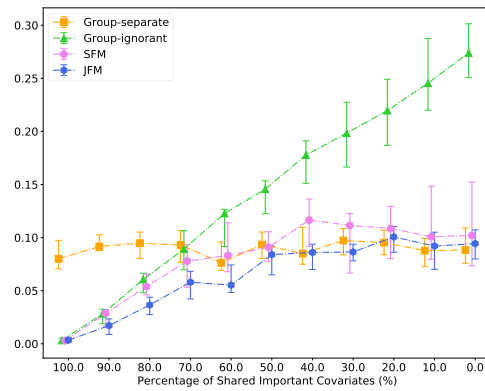
(d) Disparity of AUC

Figure 1: Experimental Results for Scenario 1

disease history such as diabetes, dementia, chronic kidney diseases (CKD); Myocardial Infarction (MI) & Atrial Fibrillation (AF); and routinely collected laboratory markers, such as lipid panels, blood panels, albumin, creatinine, aspartate aminotransferase (AST) etc. obtained from patients routine ambulatory histories before their COVID-19 infections. To build the prediction models, we randomly split the dataset into training ($n = 8,115$, 70%)
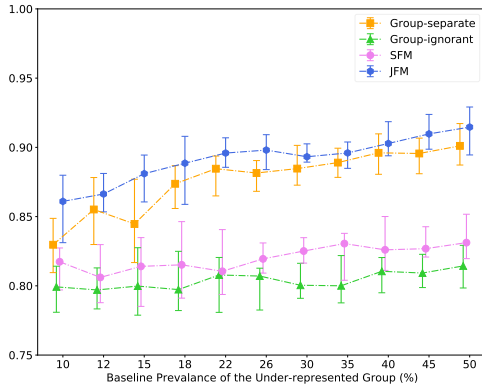
(a) AUC of the Under-represented Group
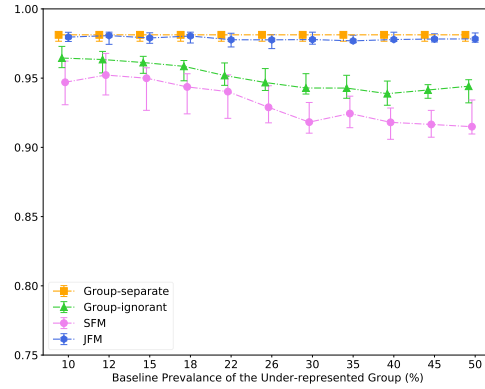
(b) AUC of the Over-represented Group
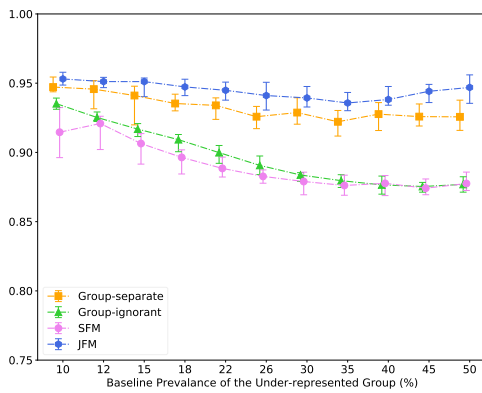
(c) Overall AUC
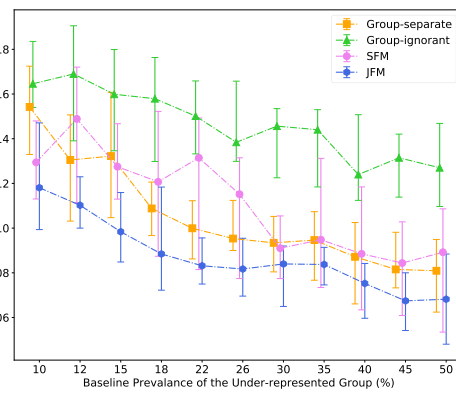
(d) Disparity of AUC

Figure 2: Experimental Results for Scenario 2

and testing ($n = 3,479$, 30%) sets. We first standardized all features to zero-mean and unit variance. Five-fold cross-validation was conducted on the training set to determine the optimal hyperparameters for each model. Hyperparameters for the group-separate and group-ignorant models were selected to maximize the groupwise AUCs and the overall AUC, respectively, while those for the SFM and JFM were determined to maximize the harmonic
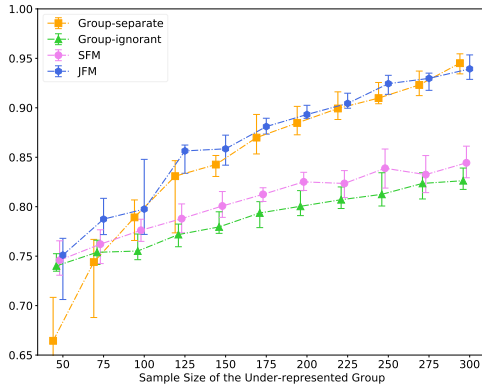
25

(a) AUC of the Under-represented Group
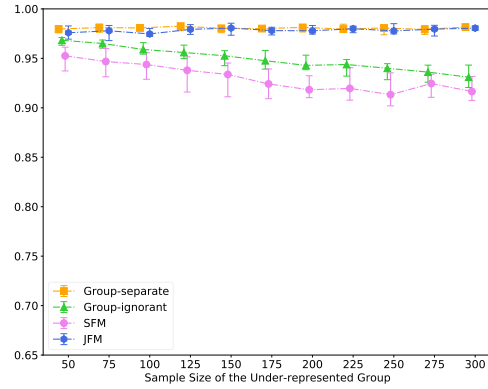
(b) AUC of the Over-represented Group
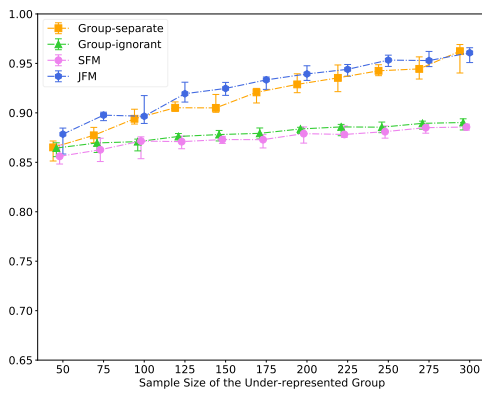
(c) Overall AUC
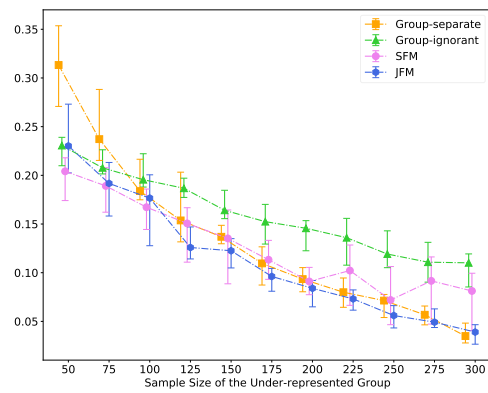
(d) Disparity of AUC

Figure 3: Experimental Results for Scenario 3

mean of groupwise AUCs. Subsequently, we trained the final models with the optimal hyperparameters using the entire training set and applied the final models to the testing dataset to demonstrate their predictive performance. We repeated the training/testing split 10 times and averaged the performances across the 10 times. Table 1 presents the AUCs and the averages of TPR and TNR of the four methods for each age group. The
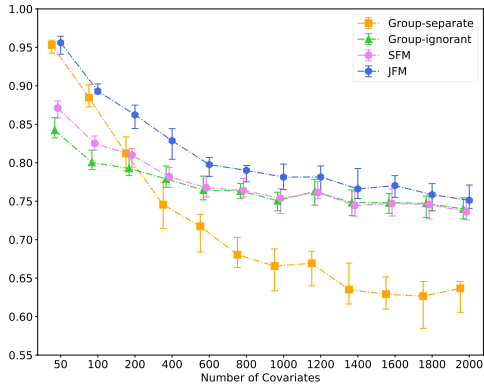
26

(a) AUC of the Under-represented Group
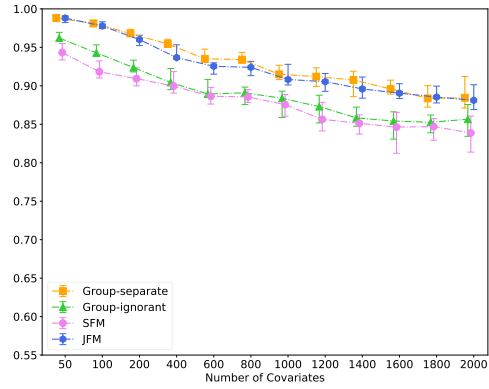
(b) AUC of the Over-represented Group
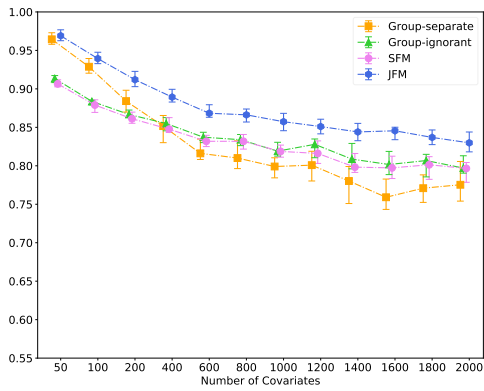
(c) Overall AUC
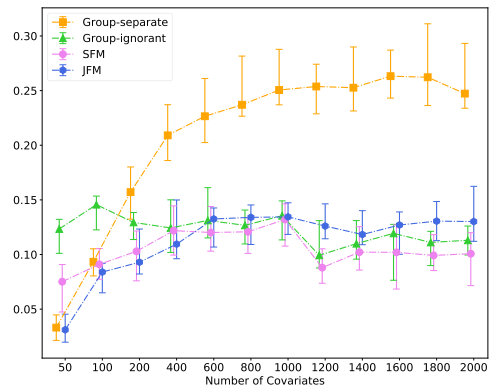
(d) Disparity of AUC

Figure 4: Experimental Results for Scenario 4

JFM performed better across all age groups than the separate model did, demonstrating that joint modeling yields higher efficiency. Compared with the group ignorant model, the JFM performed better in the three older age groups, with comparable AUC for the 50-64 age group, which resulted in smaller disparities in prediction performance overall. This phenomenon supports the observed pattern in simulation studies that the JFM reduced

disparities in prediction performances without impacting those from the majority groups. In contrast, the SFM tended to reduce prediction disparities by lowering the performances for the majority groups. Figure 5 presents the boxplots of odds ratios (ORs) of selected demographic and clinical features estimated by the JFM. These results support the hypothesis that some features have common associations between groups, and some have group-specific ORs. For example, the decreasing OR estimates of BMI along age-groups confirmed the prior hypothesis that the association between BMI and COVID-19 mortality is heterogeneous between age-groups. In JFM estimates, BMI is positively associated with higher risks of COVID-19 mortality for patients younger than 75, but with smaller and even reversed ORs in the oldest age groups. For older adults, higher BMIs are often associated with greater energy stores and a better nutritional state overall, which is beneficial for patients' survival outcomes when infected by COVID-19. The proportion of underweight patients (BMI<18) increased from 0.6% in the age group 50-64 to 5.5% in the age group 85+. The underweight status, often a proxy of frailty, has been repeatedly reported as a strong risk factor of COVID-19-induced multiorgan failure and mortality in older patients [Tehrani et al., 2021]. On the other hand, the JFM can improve efficiencies for covariates with rare prevalence in a subgroup. For instance, dementia has been reported as a risk factor with COVID-19 mortality. In the group-separate model, dementia was insignificant in patients aged 50-64, mainly due to its low prevalence in this group (0.6%). In contrast, dementia was significantly associated with mortality in all age groups with similar ORs in the JFM estimates.

Models	AUCs				Average of TPR and TNR			
	50-64	65-74	75-84	Over 85	50-64	65-74	75-84	Over 85
Group-separate	0.838	0.773	0.709	0.649	0.780	0.722	0.669	0.632
Group-ignorant	**0.855**	0.786	0.735	0.659	**0.803**	**0.731**	0.687	0.639
SFM	0.847	0.774	0.728	0.660	0.791	0.724	0.688	0.640
JFM	0.852	**0.791**	**0.736**	**0.672**	0.794	**0.731**	**0.690**	**0.659**

Table 1: Predictive Performance on COVID-19 Case Study
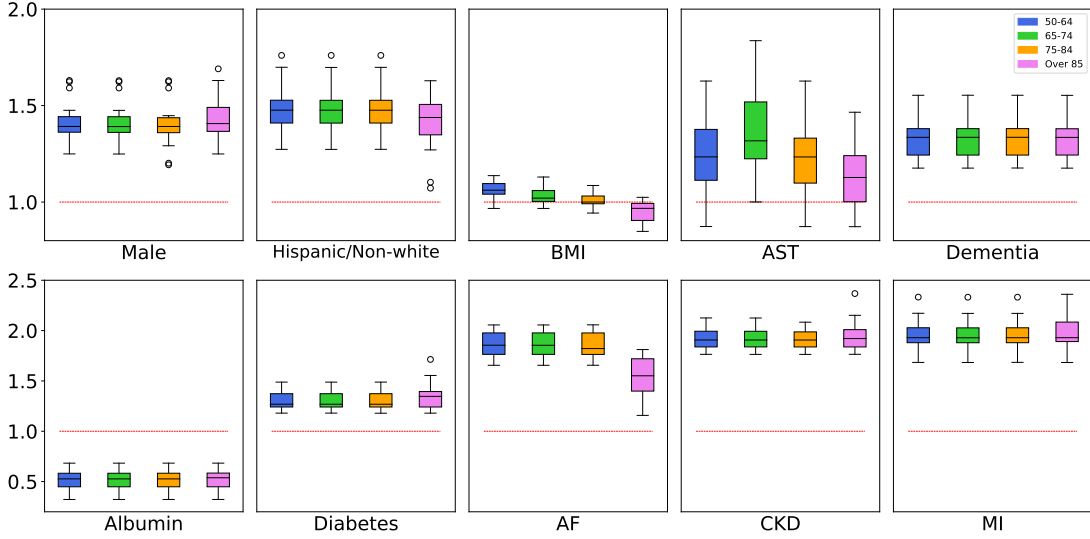
Figure 5: Estimated Odds Ratios for COVID-19 Dataset

7   JFM for Generalized Linear Models

The proposed JFM framework in (2) can be extended to Generalized Linear Models (GLMs) when the response variable $\mathbf{y}^k$ is obtained from an exponential family.

We can choose a generalized fairness penalty function to encourage each group to have

similar linear components.

$$\mathcal{P}_{\mathrm{F}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\mathrm{F}}) = \lambda_{\mathrm{F}} \sum_{j<k} \mathbb{E}_{\mathbf{y}} \left[\left| \mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}^j | G = j, \mathbf{y} = y\right] - \mathbb{E}\left[\mathbf{X}\boldsymbol{\beta}^k | G = k, \mathbf{y} = y\right] \right| \right],$$

The proposed accelerated smoothing proximal gradient method can also be extended to solve the generalized JFMs.

8   Conclusions and Discussion

In this study we introduced a new method, the joint fairness model, for jointly estimating sparse parameters on the basis of observations drawn from distinct but related groups with the goal of achieving fair performances across groups. We employ an efficient accelerated smoothing proximal gradient algorithm to solve the joint fair objective function, which has convex penalty functions. Our algorithm is tractable on high-dimensional datasets (thousands of features on thousands of samples.) Further, we presented the asymptotic distributions of parameter estimates $\hat{\boldsymbol{\beta}}^k$ and provided a framework to perform hypothesis testing of the overall $\boldsymbol{\beta}$ or the individual elements of β_j. Our JFM predictions outperform competing approaches, including group separate models, group ignorant models and single fairness models, on a range of simulated scenarios.

We note that the JFM's reliance on separate hyperparameters ($K+2$ hyperparameters) that control sparsity, fairness and similarity can be viewed as a strength rather than a drawback because one can vary separately the amount of similarity, sparsity and fairness to enforce in the group specific estimates. In situations with many groups, further assumptions can be made to reduce the number of sparsity hyperparameters (i.e. $\lambda_{\mathrm{Sp}_k} = c_k \lambda_{\mathrm{Sp}}$). Possible choices of c_k include $\frac{1}{\sqrt{n^k}}$ so that sparsity is inversely proportional to the number of samples, and 1 for the simplicity.

As an exception of nearly all existing fairness-aware prediction approaches estimating a single set of classifier parameters, recent studies have proposed to use multi-task learning (MTL) to improve algorithm fairness [Oneto et al., 2019]. However, most MTL researches have focused on joint architecture, optimization, and task relationship learning, which is a different emphasis from the proposed JFM approach to improve risk prediction performance for under-represented populations.

Moving forward, the proposed JFM framework can be extended for time-to-event outcomes by putting similar constraints. It can also be extended to non-linear models by adding a suitable fairness penalty term to the objective function. Given the increasing ability to subclassify diseases according to their molecular features and the recognition that substantial heterogeneity exists in many molecular subtypes, most diseases will be eventually classified into a collection of multiple subtypes with unbalanced sample sizes. Therefore, the proposed JFM has wide application potential to improve prediction efficiencies and reduce subgroup prediction disparities beyond applications addressing gender, race/ethnicity and age disparities.

A Python package implementing the JFM will be made available at
`https://github.com/hyungrok-do/joint-fairness-model`.

References

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. *ProPublica, May*, 23:2016, 2016.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, 2009. doi: 10.1109/ICDMW.2009.83.

Danton S Char, Nigam H Shah, and David Magnus. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine*, 378 (11):981, 2018.

Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, Eric P Xing, et al. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

Edward A. Chow, Henry Foster, Victor Gonzalez, and LaShawn McIver. The disparate impact of diabetes on racial/ethnic minority populations. *Clinical Diabetes*, 30(3):130–133, 2012. ISSN 0891-8929. doi: 10.2337/diaclin.30.3.130.

Patrick Danaher, Pei Wang, and Daniela M Witten. The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(2):373–397, 2014.

Frank Dondelinger, Sach Mukherjee, and The Alzheimer's Disease Neuroimaging Initiative. The joint lasso: high-dimensional regression for group structured data. *Biostatistics*, 21 (2):219–235, 09 2018. ISSN 1465-4644. doi: 10.1093/biostatistics/kxy035.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In Sorelle A. Friedler and Christo Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 119–133, New York, NY, USA, 23–24 Feb 2018. PMLR.

Jerome Friedman, Trevor Hastie, Holger Höfling, Robert Tibshirani, et al. Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332, 2007a.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2007b. ISSN 1465-4644. doi: 10.1093/biostatistics/kxm045.

Milena A. Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178(11):1544–1547, Nov 2018. ISSN 2168-6114. doi: 10.1001/jamainternmed.2018.3763. 30128552[pmid].

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Holger Hoefling. A path algorithm for the fused lasso signal approximator. *Journal of*

Computational and Graphical Statistics, 19(4):984–1006, 2010. doi: 10.1198/jcgs.2010.09208.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012. ISSN 0219-3116. doi: 10.1007/s10115-011-0463-8.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. pages 35–50, 09 2012. doi: 10.1007/978-3-642-33486-3_3.

Lyudmyla Kompaniyets, Alyson B. Goodman, Brook Belay, Freedman David S., Marissa S. Sucosky, Samantha J. Lange, Adi V. Gundlapalli, Tegan K. Boehmer, and Heidi M. Blanck. Body mass index and risk for covid-19–related hospitalization, intensive care unit admission, invasive mechanical ventilation, and death — united states, march – december, 2020. *MMWR Morb Mortal Wkly Rep 2021*, 70:355–361, March 2021.

Preethi Lahoti, Gerhard Weikum, and Krishna P. Gummadi. ifair: Learning individually fair data representations for algorithmic decision making. *CoRR*, abs/1806.01059, 2018.

Vivek H. Murthy, Harlan M. Krumholz, and Cary P. Gross. Participation in Cancer Clinical TrialsRace-, Sex-, and Age-Based Disparities. *JAMA*, 291(22):2720–2726, 06 2004. ISSN 0098-7484. doi: 10.1001/jama.291.22.2720.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Guillaume Obozinski, Ben Taskar, and Michael I Jordan. Joint covariate selection and joint

subspace selection for multiple classification problems. *Statistics and Computing*, 20(2): 231–252, 2010.

Luca Oneto, Michele Doninini, Amon Elders, and Massimiliano Pontil. Taking advantage of multitask learning for fair classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 227–237, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242. doi: 10.1145/3306618. 3314255.

Hee Jung Ryu, Hartwig Adam, and Margaret Mitchell. Inclusivefacenet: Improving face attribute detection with race and gender diversity. In *Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*. Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), 2018.

Muhammad Rizwan Sardar, Marwan Badri, Catherine T. Prince, Jonathan Seltzer, and Peter R. Kowey. Underrepresentation of Women, Elderly Patients, and Racial Minorities in the Randomized Trials Used for Cardiovascular Guidelines. *JAMA Internal Medicine*, 174(11):1868–1870, 11 2014. ISSN 2168-6106. doi: 10.1001/jamainternmed.2014.4758.

Sara Tehrani, Anna Killander, Per Åstrand, Jan Jakobsson, and Patrik Gille-Johnson. Risk factors for death in adult covid-19 patients: Frailty predicts fatal outcome in older patients. *International Journal of Infectious Diseases*, 102:415–421, 2021. ISSN 1201-9712. doi: https://doi.org/10.1016/j.ijid.2020.10.071.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996. doi: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007. ISSN 00063444.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE, 2017a. International World Wide Web Conferences Steering Committee. ISBN 9781450349130. doi: 10.1145/3038912.3052660.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In Aarti Singh and Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970, Fort Lauderdale, FL, USA, 20–22 Apr 2017b. PMLR.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: A flexible approach for fair classification. *Journal of Machine Learning Research*, 20(75):1–42, 2019.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 67(2):301–320, 2005. ISSN 13697412, 14679868.

SUPPLEMENTARY MATERIAL

S.1 Computational Remark

Although $\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}^{T}\boldsymbol{\alpha}^*$ and $\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta}$ in Proposition 3.2 and 3.3 seem computationally expensive due to the high-dimensionality of $\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}} \in \mathbb{R}^{(p+1)K(K-1)\times pK}$, we can reduce the complexity because of their structure.

For $\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}^{T}\boldsymbol{\alpha}^*$, we have

$$\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}^{T}\boldsymbol{\alpha}^* = \lambda_{\mathrm{F}}\mathbf{D}_0\alpha_1^* + \lambda_{\mathrm{F}}\mathbf{D}_1\alpha_2^* + \lambda_{\mathrm{Sim}}\mathbf{A}^*, \tag{S.1}$$

where

$$\mathbf{A}^* = \begin{pmatrix} \boldsymbol{\alpha}_{3+}^* & \boldsymbol{\alpha}_{3+}^* & \boldsymbol{\alpha}_{3+}^* & \cdots & \mathbf{0} \\ -\boldsymbol{\alpha}_{3+}^* & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & -\boldsymbol{\alpha}_{3+}^* & \mathbf{0} & \cdots & \mathbf{0} \\ & & \vdots & & \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\alpha}_{3+}^* \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \cdots & -\boldsymbol{\alpha}_{3+}^* \end{pmatrix},$$

and $\boldsymbol{\alpha}_{3+}^* = (\alpha_3^*, \cdots, \alpha_{pK}^*)$ is sub-vector of $\boldsymbol{\alpha}^*$ that obtained by removing first two elements from it. (S.1) requires scalar-matrix multiplication and matrix addition and thus its computational complexity is $\mathcal{O}(pK^3)$, which is lower than $\mathcal{O}(p^2K^3)$ of the matrix multiplication.

On the other hand, we have

$$\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta} = \begin{pmatrix} \lambda_{\mathrm{F}}\mathbf{D}_0\boldsymbol{\beta} \\ \lambda_{\mathrm{F}}\mathbf{D}_1\boldsymbol{\beta} \\ \lambda_{\mathrm{Sim}}\mathbf{F}\boldsymbol{\beta} \end{pmatrix}. \tag{S.2}$$

Here, $\mathbf{F}$ is a sparse matrix consists of identity matrices and thus $\mathbf{F}\boldsymbol{\beta}$ can be computed

38

without matrix multiplication by

$$\mathbf{F}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}^1 - \boldsymbol{\beta}^2 \\ \vdots \\ \boldsymbol{\beta}^{K-1} - \boldsymbol{\beta}^K \end{pmatrix}. \tag{S.3}$$

Its complexity is $\mathcal{O}(pK^2)$ which is lower than $\mathcal{O}(p^2K^3)$, the complexity of the standard matrix multiplication for $\mathbf{F}\boldsymbol{\beta}$. Since (S.3) only requires a series of vector subtraction operations, it is more efficient than multiplying large matrices. Note that the complexity of (S.2) is also $\mathcal{O}(pK^2)$ because $\mathbf{D}_0\boldsymbol{\beta}$ and $\mathbf{D}_1\boldsymbol{\beta}$ both require $\mathcal{O}(pK)$ computations.

S.2   Proof of Proposition 3.1

Note that the proof of Proposition 3.1 to Proposition 3.3 are based on the work of Chen et al. [2012]. The left-hand side of the inequalities is trivial by definition. For the right-hand side, we have

$$\|\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta}\|_1 - f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}) \leq \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2, \quad \forall \boldsymbol{\alpha} \in \mathbb{R}^{pK} \text{ s.t. } \|\boldsymbol{\alpha}\|_\infty \leq 1,$$

and it is easy to verify that $\|\boldsymbol{\alpha}\|_2^2 \leq pK$ given that $\boldsymbol{\alpha} \in \mathbb{R}^{pK}$ and $\|\boldsymbol{\alpha}\|_\infty \leq 1$, which completes the proof.

S.3   Proof of Proposition 3.2

The smoothness of $f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}})$ can be proved by applying the following Theorem 26.3 in Rockafellar [1970]. We start by the conjugate ϕ^* of $\phi(\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2$ defined on $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_\infty \leq 1\}$, which is given by

$$\phi^*(\boldsymbol{\beta}) = \sup_{\{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_\infty \leq 1\}} \left(\boldsymbol{\alpha}^T\boldsymbol{\beta} - \phi(\boldsymbol{\alpha})\right).$$

39

By plugging $\frac{\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}}\boldsymbol{\beta}}{\mu}$ into the conjugate function, we have

$$\mu\phi^*\left(\frac{\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}}\boldsymbol{\beta}}{\mu}\right) = \sup_{\{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_\infty \leq 1\}}\left(\boldsymbol{\alpha}^T\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2\right) = f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}).$$

Therefore, f_μ has the essentially smooth conjugate function (we can easily verify that $\phi(\boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{\alpha}\|_2^2$ defined on $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_\infty \leq 1\}$ is essentially convex) and thus it is a smooth function.

To obtain the gradient ∇f_μ, we apply Danskin's theorem. Let

$$\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \boldsymbol{\alpha}^T\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}}\boldsymbol{\beta} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2.$$

Then,

$$f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}) = \max_{\{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_\infty \leq 1\}}\psi(\boldsymbol{\alpha}, \boldsymbol{\beta}).$$

Since $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_\infty \leq 1\}$ is a compact set, f_μ is continuous in both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, and it is convex in $\boldsymbol{\beta}$ for every $\boldsymbol{\alpha}$ such that $\|\boldsymbol{\alpha}\|_\infty \leq 1$. Under these three conditions, Danskin's theorem grants that f_μ is convex in $\boldsymbol{\beta}$. Moreover,

$$\nabla f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}) = \frac{\partial}{\partial\boldsymbol{\beta}}\psi(\boldsymbol{\alpha}^*, \boldsymbol{\beta}) = \mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}}^T\boldsymbol{\alpha}^*,$$

since the set

$$\left\{\boldsymbol{\alpha}^* : \psi(\boldsymbol{\alpha}^*, \boldsymbol{\beta}) = \max_{\{\boldsymbol{\alpha}:\|\boldsymbol{\alpha}\|_\infty \leq 1\}}\psi(\boldsymbol{\alpha}, \boldsymbol{\beta})\right\}$$

has a single element because ψ is strongly convex in $\boldsymbol{\alpha}$.

S.4 Proof of Proposition 3.3

$\boldsymbol{\alpha}^*$ can be attained by solving the following optimization problem

$$\max_{\boldsymbol{\alpha}} \boldsymbol{\alpha}^T\mathbf{D}_{\lambda_{\mathrm{F}}}, \lambda_{\mathrm{Sim}}\boldsymbol{\beta} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 \text{ s.t. } \|\boldsymbol{\alpha}\|_\infty \leq 1,$$

which can be rewritten as the following minimization problem

$$\min_{\boldsymbol{\alpha}} \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 - \boldsymbol{\alpha}^T \mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta} \quad \text{s.t. } \|\boldsymbol{\alpha}\|_\infty \le 1.$$

It is equivalent to

$$\min_{\boldsymbol{\alpha}} \left\|\boldsymbol{\alpha} - \frac{\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta}}{\mu}\right\|_2^2 \quad \text{s.t. } \|\boldsymbol{\alpha}\|_\infty \le 1,$$

whose optimal solution satisfies

$$\alpha_i^* = \begin{cases} d_i & \text{if } d_i \in [-1,1] \\ 1 & \text{if } d_i \in (1,\infty) \\ -1 & \text{if } d_i \in (-\infty,-1) \end{cases},$$

where $d_i = \left[\frac{\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta}}{\mu}\right]_i$ is the i-th element of $\frac{\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta}}{\mu}$. Note that solving the minimization problem is equivalent to finding a Euclidean projection of $\frac{\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta}}{\mu}$ onto the unit L_∞ ball.

S.5 Proof of Proposition 3.4

From Theorem 3.1, with $\mu = \frac{2\delta}{pK}$ for the approximation accuracy $0 < \delta < \varepsilon$, we have

$$F(\boldsymbol{\beta}^{(t)}) - F(\boldsymbol{\beta}^{**}) \le \delta + \frac{2\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2^2}{t^2}\left(\frac{1}{4}\max\left\{\lambda_{\max}(\mathbf{X}^{kT}\mathbf{X}^k) : k = 1,\cdots,K\right\} + \frac{pK}{2\delta}\|\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\|_2^2\right).$$

Thus, the number of iterations t to achieve $F(\boldsymbol{\beta}^{(t)}) - F(\boldsymbol{\beta}^{**}) \le \varepsilon$, is bounded by

$$\sqrt{\frac{2\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2^2}{\varepsilon - \delta}\left(\frac{1}{4}\max\left\{\lambda_{\max}(\mathbf{X}^{kT}\mathbf{X}^k) : k = 1,\cdots,K\right\} + \frac{pK}{2\delta}\|\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\|_2^2\right)},$$

which can be simplified to $\mathcal{O}\left(\sqrt{\frac{pK}{\delta(\varepsilon-\delta)}}\right)$.

41

S.6 Proof of Theorem 4.1

To prove the theorem, it is sufficient to show that $\mathcal{V}_n(\mathbf{u}^1, \mathbf{u}^2) \to \mathcal{V}(\mathbf{u}^1, \mathbf{u}^2)$ as $n \to \infty$, where $\mathcal{V}_n(\mathbf{u}^1, \mathbf{u}^2)$ is defined in (S.4).

From Theorem 4.1, we can re-write $\mathcal{V}(\mathbf{u}^1, \mathbf{u}^2)$ as

$$\mathcal{V}(\mathbf{u}^1, \mathbf{u}^2) = g(\mathbf{u}^1, \mathbf{u}^2) + h(\mathbf{u}^1, \mathbf{u}^2)$$

where

$$g(\mathbf{u}^1, \mathbf{u}^2) = \mathbf{u}^{1T}\mathbf{W}^1 + \mathbf{u}^{2T}\mathbf{W}^2 + \frac{1}{2}\mathbf{u}^{1T}\mathbf{C}^1\mathbf{u}^1 + \frac{1}{2}\mathbf{u}^{2T}\mathbf{C}^2\mathbf{u}^2$$

and

$$h(\mathbf{u}^1, \mathbf{u}^2) =$$

$$\lambda_{\mathrm{F}}^{(0)}\bigg[(\bar{\mathbf{X}}_0^1\mathbf{u}^1 - \bar{\mathbf{X}}_0^2\mathbf{u}^2)\mathrm{sign}(\bar{\mathbf{X}}_0^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_0^2\boldsymbol{\beta}^2)\mathbb{I}(\bar{\mathbf{X}}_0^1\boldsymbol{\beta}^1 \neq \bar{\mathbf{X}}_0^2\boldsymbol{\beta}^2) + |\bar{\mathbf{X}}_0^1\mathbf{u}^1 - \bar{\mathbf{X}}_0^2\mathbf{u}^2|\mathbb{I}(\bar{\mathbf{X}}_0^1\boldsymbol{\beta}^1 = \bar{\mathbf{X}}_0^2\boldsymbol{\beta}^2) +$$

$$(\bar{\mathbf{X}}_1^1\mathbf{u}^1 - \bar{\mathbf{X}}_1^2\mathbf{u}^2)\mathrm{sign}(\bar{\mathbf{X}}_1^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_1^2\boldsymbol{\beta}^2)\mathbb{I}(\bar{\mathbf{X}}_1^1\boldsymbol{\beta}^1 \neq \bar{\mathbf{X}}_1^2\boldsymbol{\beta}^2) + |\bar{\mathbf{X}}_1^1\mathbf{u}^1 - \bar{\mathbf{X}}_y^2\mathbf{u}^2|\mathbb{I}(\bar{\mathbf{X}}_1^1\boldsymbol{\beta}^1 = \bar{\mathbf{X}}_1^2\boldsymbol{\beta}^2)\bigg] +$$

$$\lambda_{\mathrm{Sp}}^{(0)}\sum_{k=1}^{2}\sum_{j=1}^{p}\{u_j^k\mathrm{sign}(\beta_j^k)\mathbb{I}(\beta_j^k \neq 0) + |u_j^k|\mathbb{I}(\beta_j^k = 0)\} +$$

$$\lambda_{\mathrm{Sim}}^{(0)}\sum_{j=1}^{p}\{(u_j^1 - u_j^2)\mathrm{sign}(\beta_j^1 - \beta_j^2)\mathbb{I}(\beta_j^1 \neq \beta_j^2) + |u_j^1 - u_j^2|\mathbb{I}(\beta_j^1 = \beta_j^2)\}.$$

Let

$$\mathcal{V}_n(\mathbf{u}^1, \mathbf{u}^2) = g_n(\mathbf{u}^1, \mathbf{u}^2) + h_n(\mathbf{u}^1, \mathbf{u}^2) \tag{S.4}$$

where

$$g_n(\mathbf{u}^1, \mathbf{u}^2) = -\left\{\ell\left(\boldsymbol{\beta}^1 + \frac{\mathbf{u}^1}{\sqrt{n}}\right) - \ell(\boldsymbol{\beta}^1)\right\} - \left\{\ell\left(\boldsymbol{\beta}^2 + \frac{\mathbf{u}^2}{\sqrt{n}}\right) - \ell(\boldsymbol{\beta}^2)\right\},$$

42

and

$$h_n(\mathbf{u}^1, \mathbf{u}^2) = \lambda_{\mathrm{F}}\left\{ \left| \bar{\mathbf{X}}_0^1\left(\boldsymbol{\beta}^1 + \frac{\mathbf{u}^1}{\sqrt{n}}\right) - \bar{\mathbf{X}}_0^2\left(\boldsymbol{\beta}^2 + \frac{\mathbf{u}^2}{\sqrt{n}}\right)\right| - |\bar{\mathbf{X}}_0^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_0^2\boldsymbol{\beta}^2| \right. \tag{S.5}$$

$$\left. + \left| \bar{\mathbf{X}}_1^1\left(\boldsymbol{\beta}^1 + \frac{\mathbf{u}^1}{\sqrt{n}}\right) - \bar{\mathbf{X}}_1^2\left(\boldsymbol{\beta}^2 + \frac{\mathbf{u}^2}{\sqrt{n}}\right)\right| - |\bar{\mathbf{X}}_1^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_1^2\boldsymbol{\beta}^2| \right\}$$

$$+ \lambda_{\mathrm{Sp}} \sum_{k=1}^{2}\sum_{j=1}^{p}\left\{ \left|\beta_j^k + \frac{u_j^k}{\sqrt{n}}\right| - |\beta_j^k| \right\}$$

$$+ \lambda_{\mathrm{Sim}} \sum_{j=1}^{p}\left\{ \left|\left(\beta_j^1 + \frac{u_j^1}{\sqrt{n}}\right) - \left(\beta_j^2 + \frac{u_j^2}{\sqrt{n}}\right)\right| - |\beta_j^1 - \beta_j^2| \right\}.$$

We first show $g_n(\mathbf{u}^1, \mathbf{u}^2) \to g(\mathbf{u}^1, \mathbf{u}^2)$, that is,

$$\ell_k\left(\boldsymbol{\beta}^k + \frac{\mathbf{u}^k}{\sqrt{n}}\right) - \ell_k(\boldsymbol{\beta}^k) \to (\mathbf{u}^k)^T\mathbf{W}^k + \frac{1}{2}(\mathbf{u^k})^T\mathbf{C}^k\mathbf{u^k}. \tag{S.6}$$

Following the arguments of Viallon et al. [2013], we apply Taylor series expansion on the left side of (S.6) which yields

$$\ell_k\left(\boldsymbol{\beta}^k + \frac{\mathbf{u}^k}{\sqrt{n}}\right) - \ell_k(\boldsymbol{\beta}^k) = \frac{\nabla\ell_k(\boldsymbol{\beta}^k)^T\mathbf{u}^k}{\sqrt{n}} + \frac{1}{2}\mathbf{u^k}^T\frac{\mathcal{I}(\boldsymbol{\beta}^k)}{n}\mathbf{u}^k + o_{\mathrm{P}}\left(\frac{1}{n}\right).$$

Here, o_{P} is the small o with respect to the probability measure P. Assumption 1 ensures $\frac{1}{2}\mathbf{u}^{kT}\frac{\mathcal{I}(\boldsymbol{\beta}^k)}{n}\mathbf{u}^k \to \frac{1}{2}\mathbf{u}^{kT}\mathbf{C}^k\mathbf{u}^k$ and assumption 2 ensures $\nabla\ell_k(\boldsymbol{\beta}^k)^T\mathbf{u}^k/\sqrt{n} \to \mathbf{W}^{kk}$.

On the other hand, to show $h_n(\mathbf{u}^1, \mathbf{u}^2) \to h(\mathbf{u}^1, \mathbf{u}^2)$, we follow the arguments in Theorem 2 of Knight and Fu [2000]. For the first term of (S.5), we have

$$\lambda_{\mathrm{F}}^{(n)}\left\{ \left| \bar{\mathbf{X}}_y^1\left(\boldsymbol{\beta}^1 + \frac{\mathbf{u}^1}{\sqrt{n}}\right) - \bar{\mathbf{X}}_y^2\left(\boldsymbol{\beta}^2 + \frac{\mathbf{u}^2}{\sqrt{n}}\right)\right| - |\bar{\mathbf{X}}_y^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_y^2\boldsymbol{\beta}^2| \right\}$$

$$= \lambda_{\mathrm{F}}^{(n)}\left\{ \left| \bar{\mathbf{X}}_y^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_y^2\boldsymbol{\beta}^2 + \frac{\bar{\mathbf{X}}_y^1\mathbf{u}^1 - \bar{\mathbf{X}}_y^2\mathbf{u}^2}{\sqrt{n}}\right| - |\bar{\mathbf{X}}_y^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_y^2\boldsymbol{\beta}^2| \right\}$$

$$\to \lambda_{\mathrm{F}}^{(0)}(\bar{\mathbf{X}}_y^1\mathbf{u}^1 - \bar{\mathbf{X}}_y^2\mathbf{u}^2)\mathrm{sign}(\bar{\mathbf{X}}_y^1\boldsymbol{\beta}^1 - \bar{\mathbf{X}}_y^2\boldsymbol{\beta}^2)\mathbb{I}(\bar{\mathbf{X}}_y^1\boldsymbol{\beta}^1 \neq \bar{\mathbf{X}}_y^2\boldsymbol{\beta}^2) + |\bar{\mathbf{X}}_y^1\mathbf{u}^1 - \bar{\mathbf{X}}_y^2\mathbf{u}^2|\mathbb{I}(\bar{\mathbf{X}}_y^1\boldsymbol{\beta}^1 = \bar{\mathbf{X}}_y^2\boldsymbol{\beta}^2),$$

43

as $n \to \infty$, for $y = 0, 1$. Similarly,

$$\lambda_{\text{Sim}}^{(n)} \sum_{j=1}^{p} \left\{ \left| \left(\beta_j^1 + \frac{u_j^1}{\sqrt{n}} \right) - \left(\beta_j^2 + \frac{u_j^2}{\sqrt{n}} \right) \right| - |\beta_j^1 - \beta_j^2| \right\}$$

$$= \lambda_{\text{Sim}}^{(n)} \sum_{j=1}^{p} \left\{ \left| \beta_j^1 - \beta_j^2 + \frac{u_j^1 - u_j^2}{\sqrt{n}} \right| - |\beta_j^1 - \beta_j^2| \right\}$$

$$\to \lambda_{\text{Sim}}^{(0)} \sum_{j=1}^{p} \left\{ (u_j^1 - u_j^2)\text{sign}(\beta_j^1 - \beta_j^2)\mathbb{I}(\beta_j^1 \neq \beta_j^2) + |u_j^1 - u_j^2|\mathbb{I}(\beta_j^1 = \beta_j^2) \right\},$$

as $n \to \infty$, for $k = 1, 2$. We also have

$$\lambda_{\text{Sp}}^{(n)} \sum_{j=1}^{p} \left\{ \left| \beta_j^k + \frac{u_j^k}{\sqrt{n}} \right| - |\beta_j^k| \right\} \to \lambda_{\text{Sp}}^{(0)} \sum_{j=1}^{p} \{u_j^k\text{sign}(\beta_j^k)\mathbb{I}(\beta_j^k \neq 0) + |u_j^k|\mathbb{I}(\beta_j^k = 0)\},$$

as $n \to \infty$.

We showed that $g_n(\mathbf{u}^1, \mathbf{u}^2) \to g(\mathbf{u}^1, \mathbf{u}^2)$ and $h_n(\mathbf{u}^1, \mathbf{u}^2) \to h(\mathbf{u}^1, \mathbf{u}^2)$ as $n \to \infty$. Thus, $\mathcal{V}_n(\mathbf{u}^1, \mathbf{u}^2) \to \mathcal{V}(\mathbf{u}^1, \mathbf{u}^2)$ as $n \to \infty$ as desired.

Note: The Theorem 4.1 is proved for the JFM with *L1* penalization. For a model defined with *L2* penalization, we can simply modify $h(\mathbf{u}^1, \mathbf{u}^2)$ as below.

$$h(\mathbf{u}^1, \mathbf{u}^2) = \lambda_{\text{F}}^{(0)} \left[(\bar{\mathbf{X}}_0^1 \mathbf{u}^1 - \bar{\mathbf{X}}_0^2 \mathbf{u}^2)\text{sign}(\bar{\mathbf{X}}_0^1 \boldsymbol{\beta}^1 - \bar{\mathbf{X}}_0^2 \boldsymbol{\beta}^2)|\bar{\mathbf{X}}_0^1 \mathbf{u}^1 - \bar{\mathbf{X}}_0^2 \mathbf{u}^2| + \right.$$

$$\left. (\bar{\mathbf{X}}_1^1 \mathbf{u}^1 - \bar{\mathbf{X}}_1^2 \mathbf{u}^2)\text{sign}(\bar{\mathbf{X}}_1^1 \boldsymbol{\beta}^1 - \bar{\mathbf{X}}_1^2 \boldsymbol{\beta}^2)|\bar{\mathbf{X}}_1^1 \mathbf{u}^1 - \bar{\mathbf{X}}_y^2 \mathbf{u}^2| \right] +$$

$$\lambda_{\text{Sp}}^{(0)} \sum_{k=1}^{2} \sum_{j=1}^{p} \{u_j^k \text{sign}(\beta_j^k)|u_j^k| + \lambda_{\text{Sim}}^{(0)} \sum_{j=1}^{p} \{(u_j^1 - u_j^2)\text{sign}(\beta_j^1 - \beta_j^2)|u_j^1 - u_j^2|\}.$$

Following Knight and Fu [2000] and the arguments above, we can show $\sqrt{n}$-consistency of the estimates obtained from a model with *L2* penalization. The consistency of estimates obtained form a model utilizing mixture of *L1* and *L2* penalization can be proved similarly.

44

S.7 Computational Analysis

Figure S.1 displays JFM's empirical computational complexity against the number of features and sample sizes.

For the first experiment, we increased the number of covariates from 100 to 5,000 while fixing the sample size at 200 and 500 respectively. Figure S.1(a) shows that the JFM computation time is approximately $\mathcal{O}(p^{1.5})$, which is because Algorithm 1's per iteration complexity is linear in p and its rate of convergence is proportional to $\sqrt{p}$. With $5,000$ features, JFM finishes in 9 seconds on one Intel Xeon Platinum 8268 Processor (2.90 GHz, 24 cores) and 32GB RAM.

We then varied the sample size to 7,000 (5:2 ratio between groups) while the number of features was fixed at 1,000. In Figure S.1(b), the computation time is approximately $\mathcal{O}(n)$ for $n > 1,000$, as shown in Proposition 3.5. For $n < 1,000$, the computation time is inversely proportional to n because Problem (3) is ill-posed ($p > n$) and requires more iterations for convergence.

S.8 Accelerated Smoothing Proximal Gradient Algorithm for SFM

Bechavod and Ligett [2017] suggested to use CVXPY [Diamond and Boyd, 2016] to solve the SFM optimization problem. Since the problem is convex, CVXPY it can easily be handled. However, CVXPY is equipped with a general quadratic optimization solver and it is not efficient enough to be scalable for high-dimensional problems. Here, we introduce a variant of Algorithm 1 to solve the SFM more efficiently. Consider the following SFM optimization problem:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; -\ell(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) + \lambda_{\text{F}_0} \sum_{j<k} |(\bar{\mathbf{X}}_0^j - \bar{\mathbf{X}}_0^k)\boldsymbol{\beta}| + \lambda_{\text{F}_1} \sum_{j<k} |(\bar{\mathbf{X}}_1^j - \bar{\mathbf{X}}_1^k)\boldsymbol{\beta}| + \lambda_{\text{Sp}}\|\boldsymbol{\beta}\|_1. \quad \text{(S.7)}$$
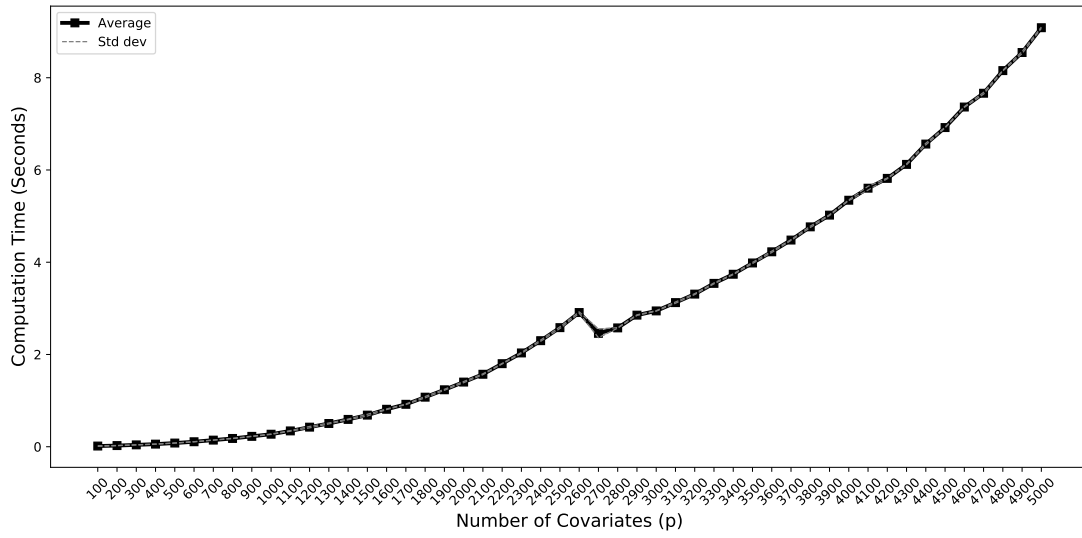
(a) Increasing Number of Covariates
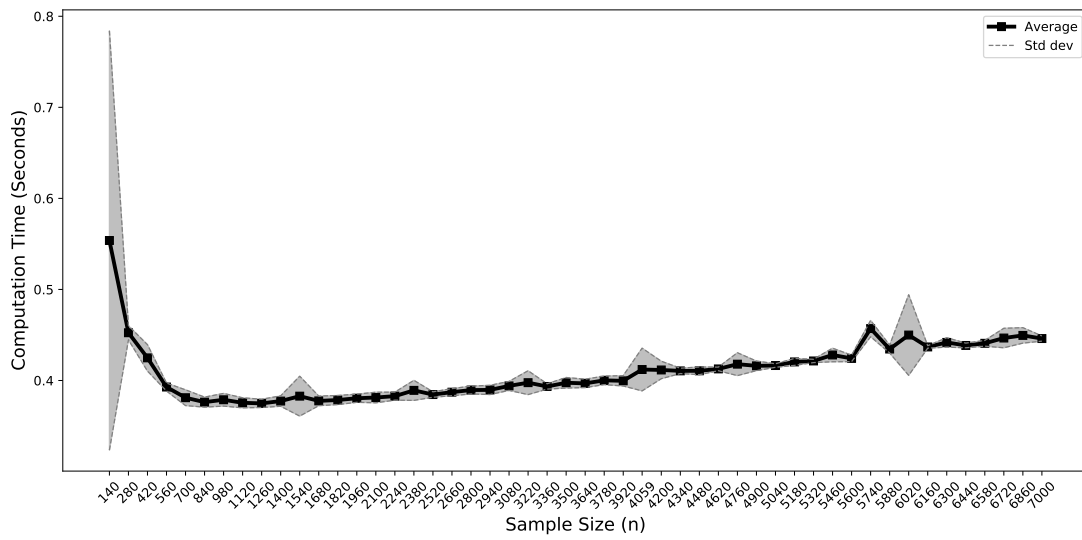
(b) Increasing Sample Size

Figure S.1: Experimental Results for Computational Analysis

46

Analogous to the matrix representation in Chapter 3, we can rewrite the objective function in matrix form:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} - \ell(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) + \|\mathbf{D}_{\lambda_{\mathrm{F}}}\boldsymbol{\beta}\|_1 + \lambda_{\mathrm{Sp}}\|\boldsymbol{\beta}\|_1,$$

where :

$$\mathbf{D}_{\lambda_{\mathrm{F}}} = \begin{pmatrix} \lambda_{\mathrm{F}_0}(\bar{\mathbf{X}}_0^1 - \bar{\mathbf{X}}_0^2) \\ \vdots \\ \lambda_{\mathrm{F}_0}(\bar{\mathbf{X}}_0^{K-1} - \bar{\mathbf{X}}_0^K) \\ \lambda_{\mathrm{F}_1}(\bar{\mathbf{X}}_1^1 - \bar{\mathbf{X}}_1^2) \\ \vdots \\ \lambda_{\mathrm{F}_1}(\bar{\mathbf{X}}_1^{K-1} - \bar{\mathbf{X}}_1^K) \end{pmatrix}.$$

We can easily verify the Nesterov smooth approximation can be applied to approximate $\|\mathbf{D}_{\lambda_{\mathrm{F}}}\boldsymbol{\beta}\|_1$ and Proposition 3.1, 3.2, and 3.3 hold. Therefore, Algorithm 2 solves the SFM optimization problem.

Algorithm 2 Accelerated Smoothing Proximal Gradient Method for SFM

1: **Input:** Data $\mathbf{X}, \mathbf{y}$, hyperparameters $\lambda_{\mathrm{F_0}}, \lambda_{\mathrm{F_1}}, \lambda_{\mathrm{Sp}}, \epsilon, \mu$

2: **Output:** $\hat{\boldsymbol{\beta}}$ solving the Single Fairness optimization problem (S.7).

3: **Initialize:** $\boldsymbol{\beta}^{(0)} = \mathbf{0}, \boldsymbol{\gamma}^{(0)} = \mathbf{0}, s^{(0)} = 1$

4: Compute $L = \frac{1}{4}\lambda_{\max}(\mathbf{X}^T\mathbf{X}) + \mu^{-1}\|\mathbf{D}_{\lambda_{\mathrm{F}}}\|_2^2$

5: **for** $m \geq 1$ **do**

6:   $\quad \boldsymbol{\alpha}^{(m)} = \boldsymbol{\gamma}^{(m-1)} - L^{-1}\left(-\nabla\ell\left(\boldsymbol{\gamma}^{(m-1)}\right) + \nabla f_\mu\left(\boldsymbol{\gamma}^{(m-1)}\right)\right)$

7:   $\quad \boldsymbol{\beta}^{(m)} = \mathcal{S}\left(\boldsymbol{\alpha}^{(m)}; L^{-1}\lambda_{\mathrm{Sp}}\right)$

8:   $\quad$ **if** $\|\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}\|_2 \leq \epsilon$ **break**

9:   $\quad s^{(m)} = \frac{1+\sqrt{1+4s^{(m-1)2}}}{2}$

10: $\quad \boldsymbol{\gamma}^{(m)} = \boldsymbol{\beta}^{(m)} + \left(\frac{s^{(m-1)}-1}{s^{(m)}}\right)\left(\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}\right)$

11: $\quad m \leftarrow m + 1$

12: **end for**

13: $\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(m)}$.

S.9 Choice of the Evaluation Metrics in Selecting Hyperparameters in Cross-validations

The group-ignorant model, group-separate model, SFM, and JFM contain 1, K, 2, and $K + 2$ hyperparameters respectively. For every method, 5-fold cross-validation on the training dataset was used to determine the hyperparameters. For the vanilla models (group-separate and group-ignorant), the lasso penalty term was selected by optimizing cross-validation AUCs. For the fairness-aware models, we compared a series of evaluation metrics for selecting the hyperparameters in cross-validations, including group average of AUCs/accuracies (arithmetic mean, geometric mean, and harmonic mean), overall AUCs/accuracies on all samples ignoring group memberships, and the group average of

48

AUCs/accuracies subtracting the disparity of AUCs/accuracies (absolute differences and squared differences). Supplementary Figures S.2, S.3, and S.4 show the prediction performances in the test datasets with the optimal hyperparameters selected by various metrics. They demonstrate that the performances in the test datasets with the hyperparameters optimizing group-average AUCs in cross-validations were more optimal than those with the hyperparameters optimizing overall AUCs in cross-validations. Although the hyperparameters chosen to optimize group average of AUCs subtracting disparities provided better fairness performance in test datasets, it was often achieved by lowering the performance of the over-represented group. We also note that the hyperparameters optimizing AUC-based evaluation metrics generated more robust performances in test datasets than those optimizing threshold-based metrics such as accuracies and TPRs/TNRs. Therefore, the simulation results use the hyperparameters optimized by the harmonic mean of group-wise AUCs in cross-validations.
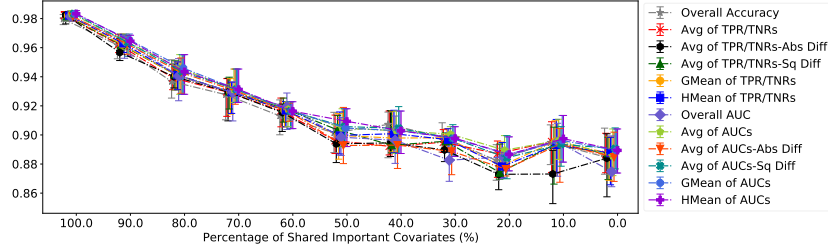
(a) AUC of the Under-represented Group
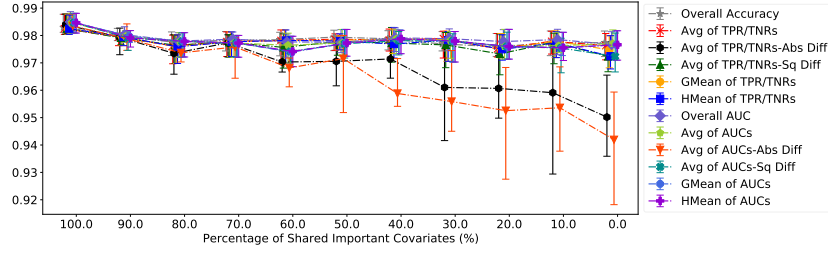
(b) AUC of the Over-represented Group
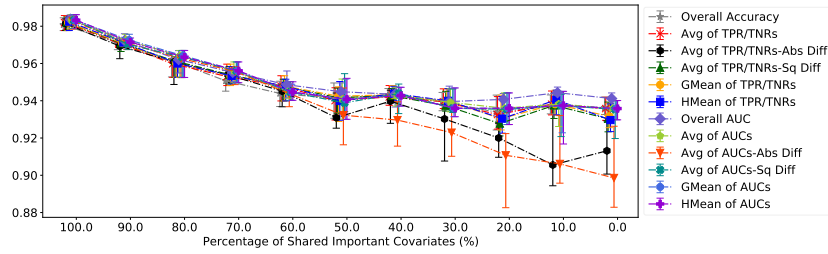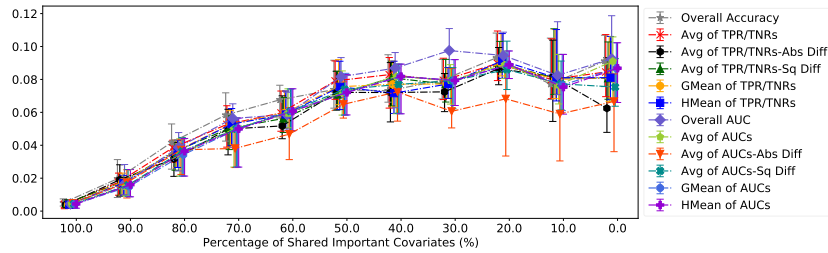
(c) Overall AUC

(d) Disparity of AUC

Figure S.2: Experimental Results for Evaluation Metrics on Scenario 1
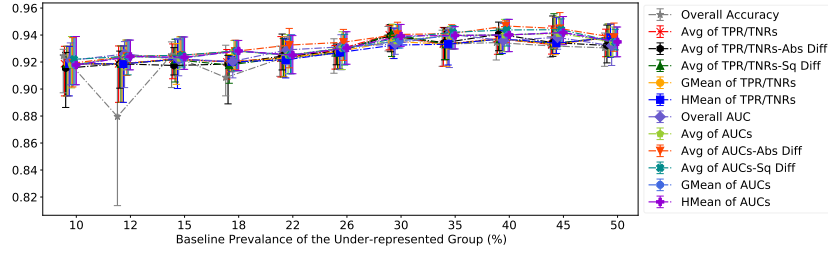
(a) AUC of the Under-represented Group
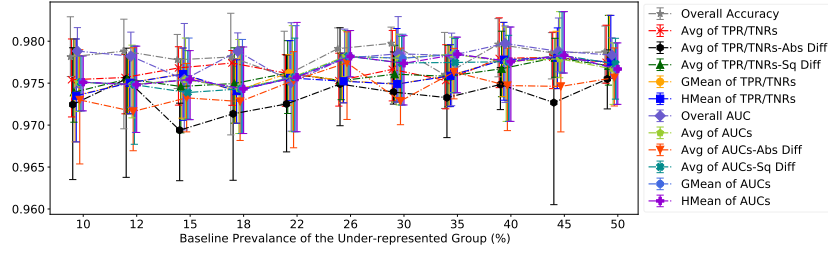
(b) AUC of the Over-represented Group
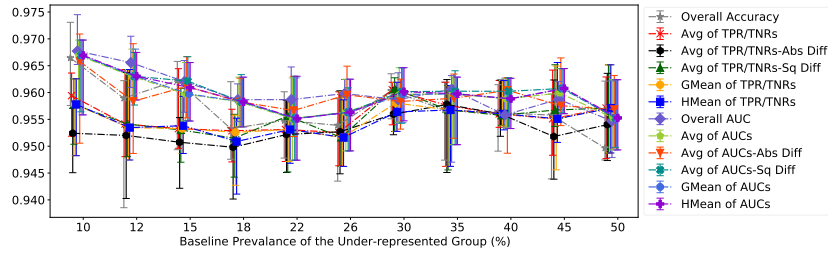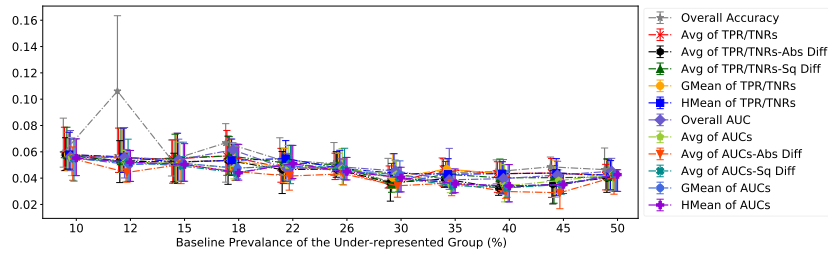
(c) Overall AUC

(d) Disparity of AUC

Figure S.3: Experimental Results for Evaluation Metrics on Scenario 2
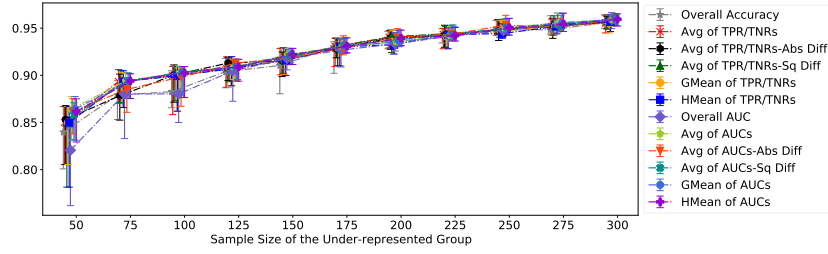
51

(a) AUC of the Under-represented Group
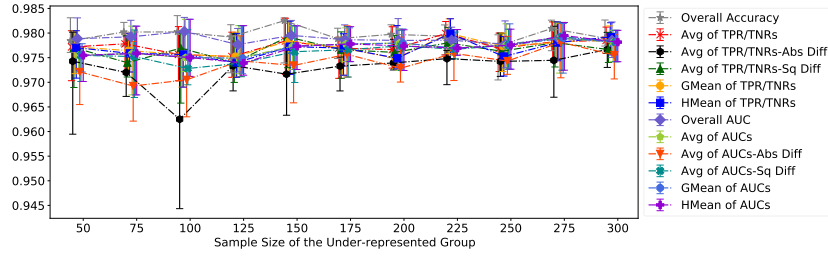
(b) AUC of the Over-represented Group
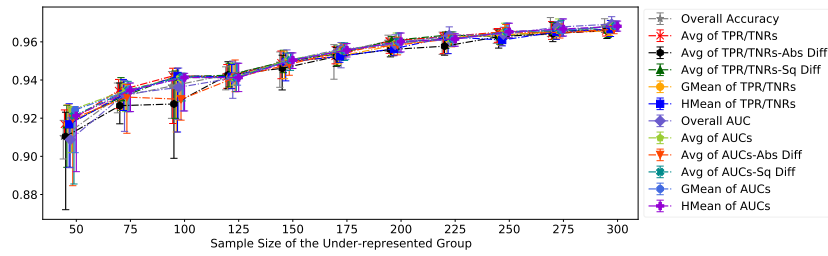
(c) Overall AUC
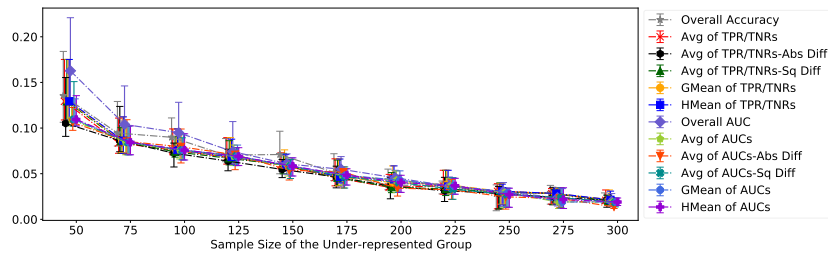
(d) Disparity of AUC

Figure S.4: Experimental Results for Evaluation Metrics on Scenario 3

S.10 Average of TPR and TNR Plots for Simulation Study
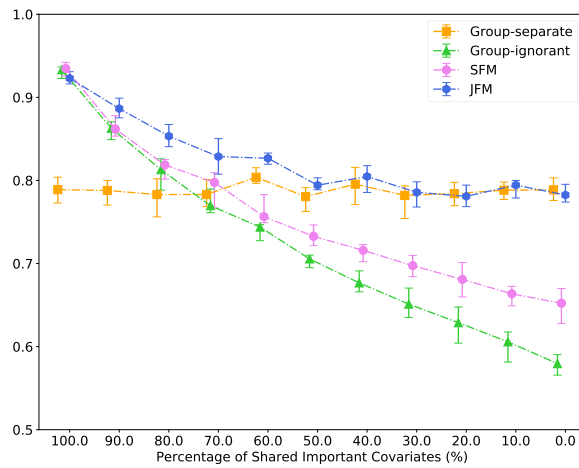
(a) Average of TPR and TNR of the Under-represented Group
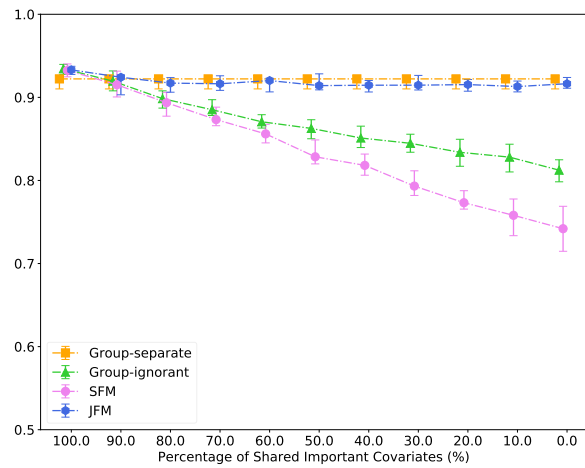
(b) Average of TPR and TNR of the Over-represented Group
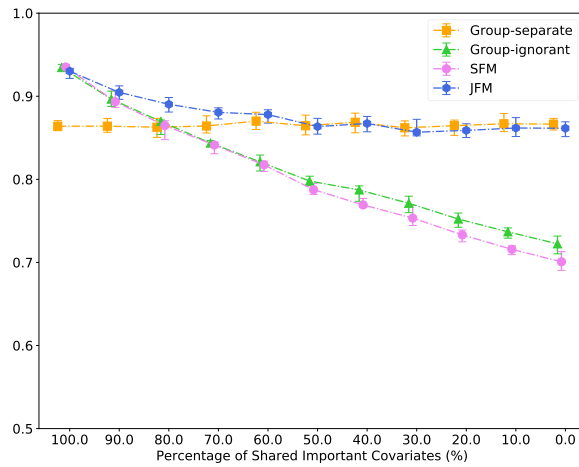
(c) Overall Average of TPR and TNR
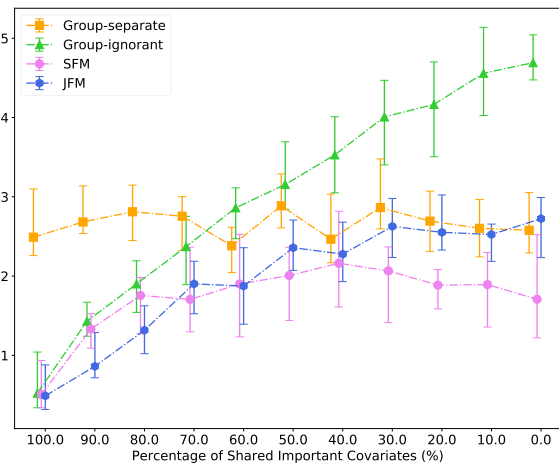
(d) Disparity of TPR and TNR

Figure S.5: Experimental Results for Scenario 1 (TPR + TNR)
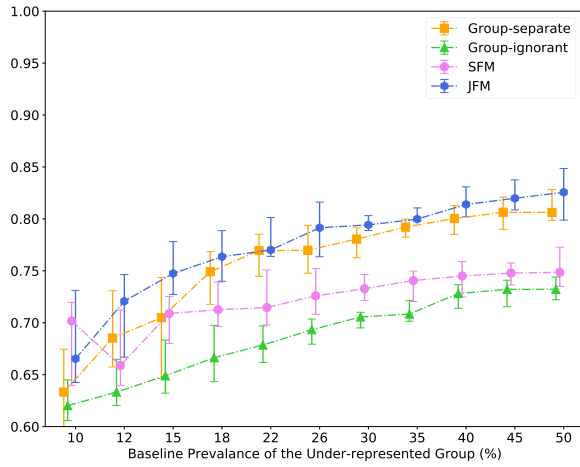
(a) Average of TPR and TNR of the Under-represented Group
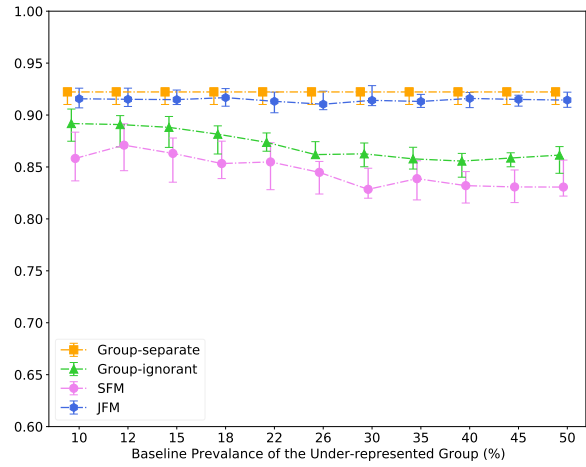
(b) Average of TPR and TNR of the Over-represented Group
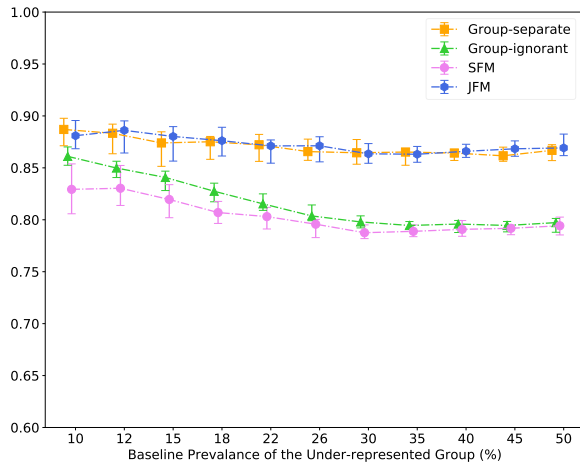
(c) Overall Average of TPR and TNR
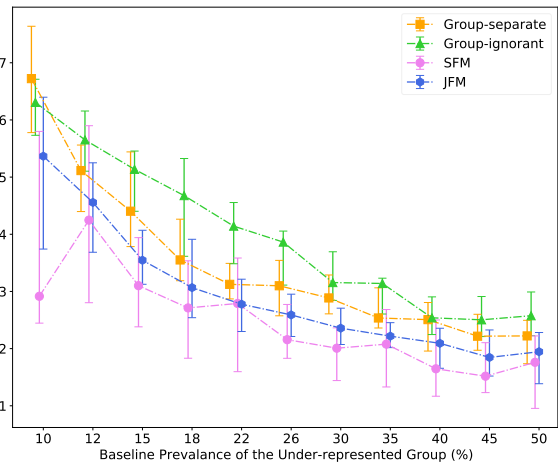
(d) Disparity of TPR and TNR

Figure S.6: Experimental Results for Scenario 2 (TPR + TNR)
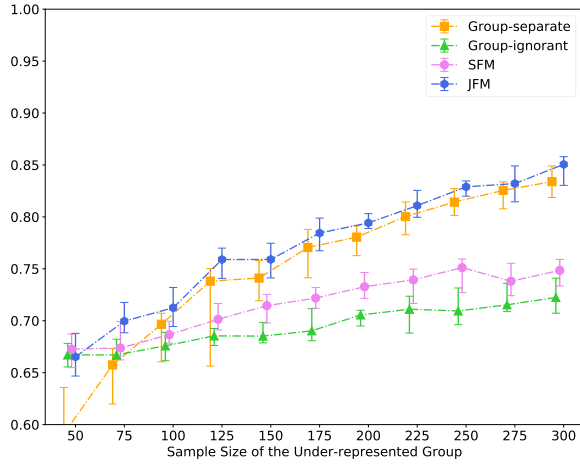
(a) Average of TPR and TNR of the Under-represented Group
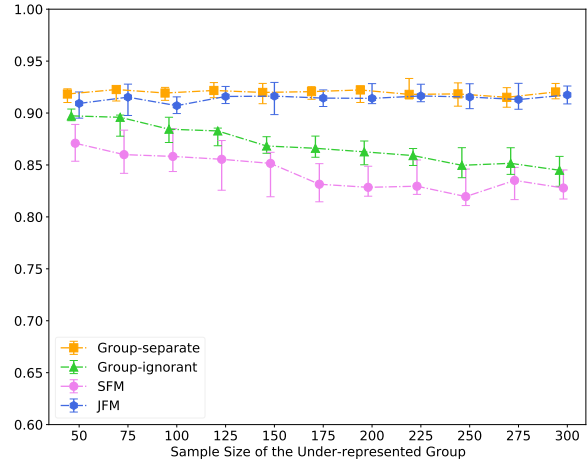
(b) Average of TPR and TNR of the Over-represented Group
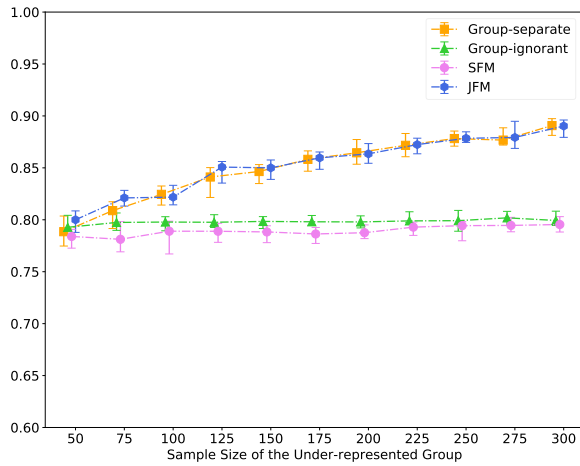
(c) Overall Average of TPR and TNR
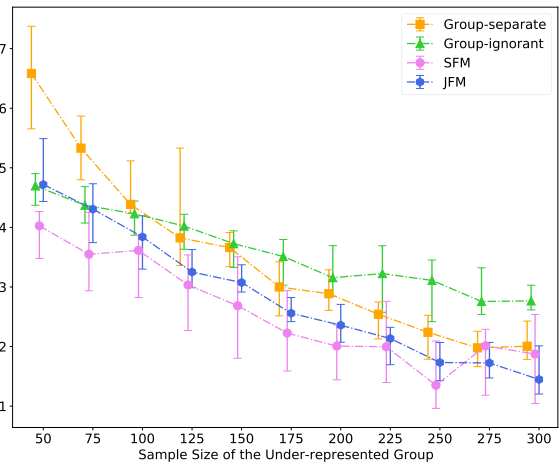
(d) Disparity of TPR and TNR

Figure S.7: Experimental Results for Scenario 3 (TPR + TNR)
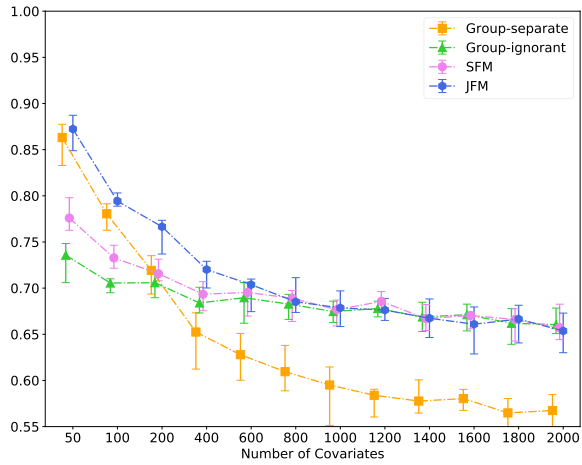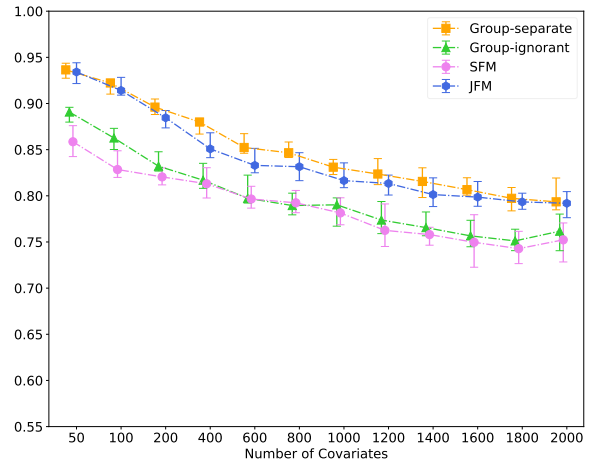
55

(a) Average of TPR and TNR of the Under-represented Group

(b) Average of TPR and TNR of the Over-represented Group
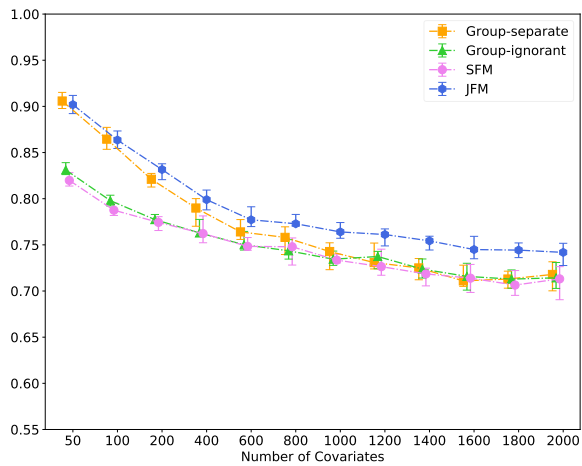
(c) Overall Average of TPR and TNR
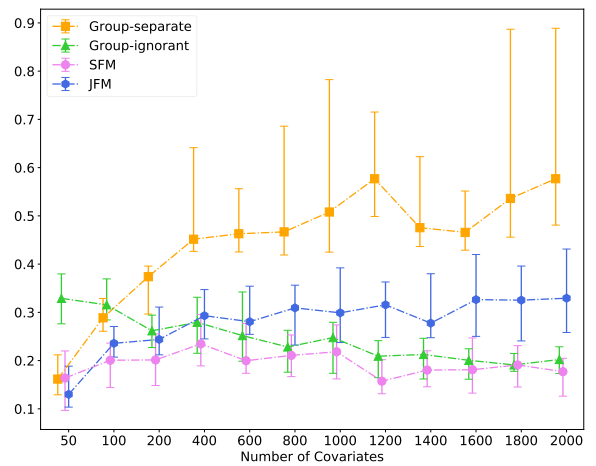
(d) Disparity of TPR and TNR

Figure S.8: Experimental Results for Scenario 4 (TPR + TNR)

S.11 Additional Simulation Scenarios

Here, we present the results for additional simulation scenarios. The datasets are generated in the same way as in Section 5.

- In Scenario 1B, the number of non-zero coefficients of the under-represented group ranged from 20 to 40. The number of shared features fixed at 20, the baseline prevalence were 50% and 30% for the over and under-represented groups, respectively. The sample sizes were set at 500 and 200 for over and under-represented groups. The number of features were $p = 100$.

- In Scenario 2B, the baseline prevalence of the under-represented group ranged from 50% to 90% while the baseline event prevalence of the over-represented group was fixed at 50%.

- In Scenario 3B, samples size of the over-represented group ranged from 500 to 2500 with the sample size of the under-represented group fixed at 200.

- In Scenario 4B, the number of features p ranged from 50 to 2,000. Everything is same with the Scenario 4, except that for each p, 30% of the features had non-zero coefficients.

As same as with the Section 5, we evaluated the methods on independent testing datasets under the same setups with large sample sizes (both 1000). AUC was used to evaluate the predictive performance of each model.
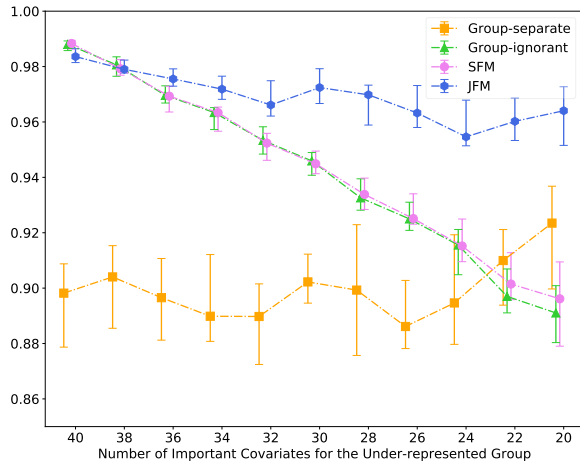
(a) AUC of the Under-represented Group
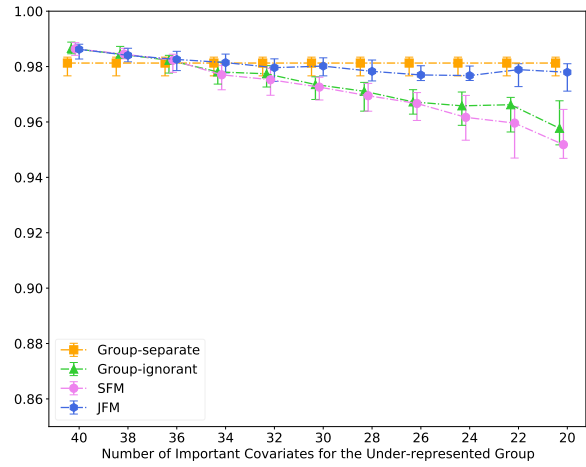
(b) AUC of the Over-represented Group
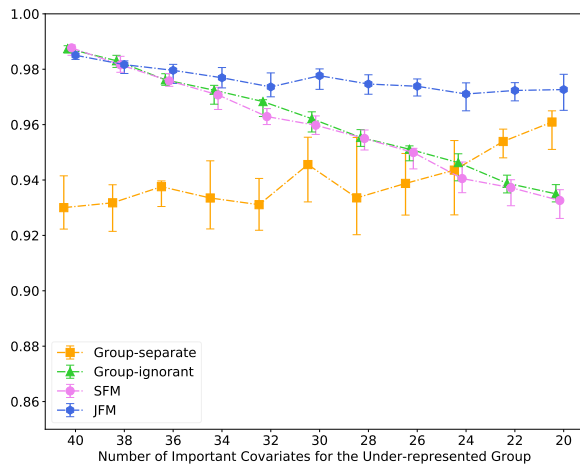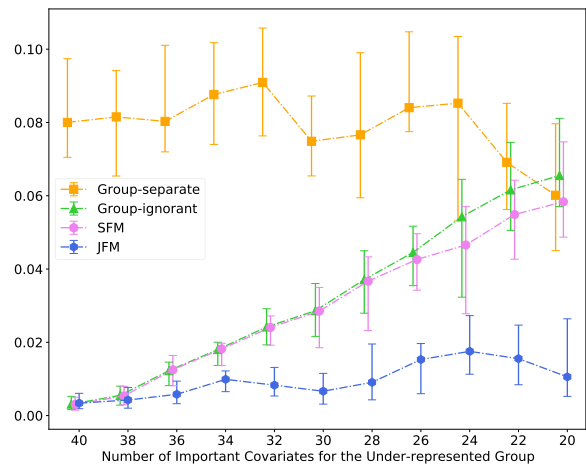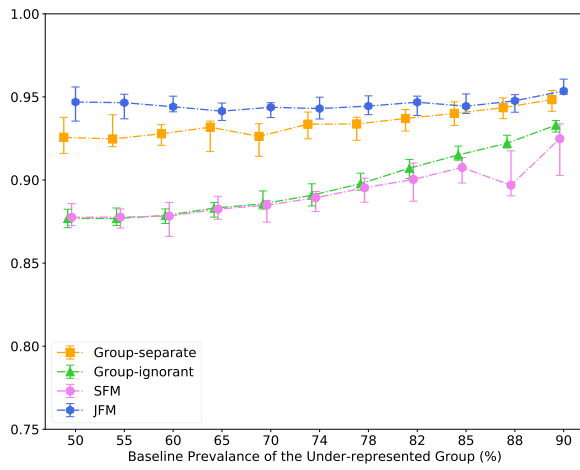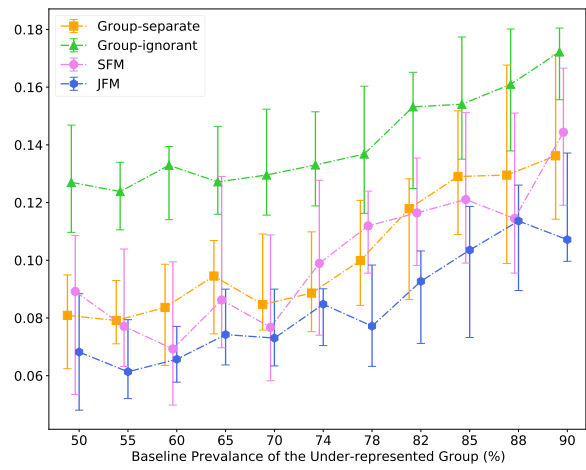
(c) Overall AUC

(d) Disparity of AUC

Figure S.9: Experimental Results for Scenario 1B
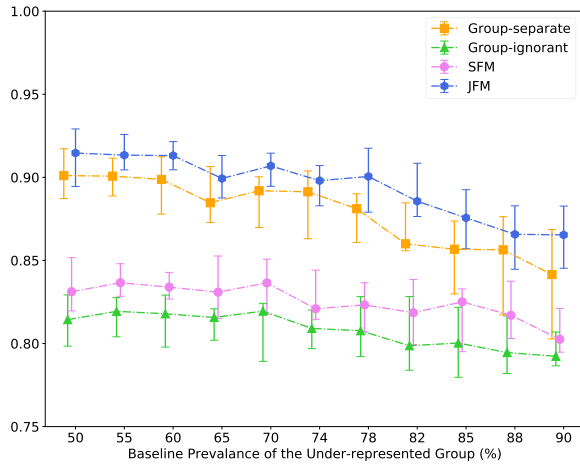
(a) AUC of the Under-represented Group
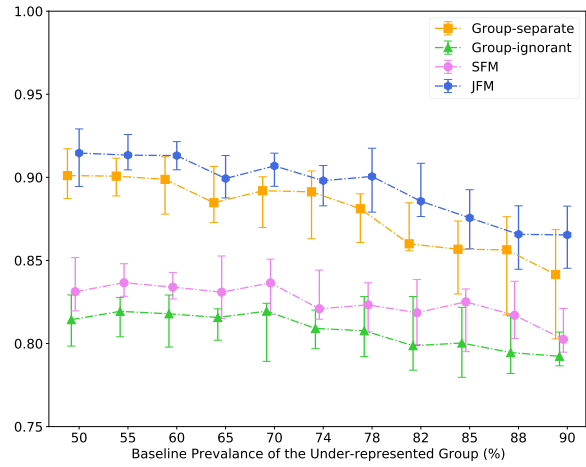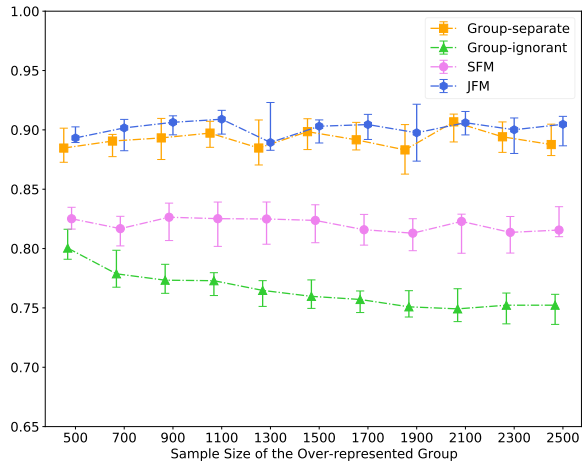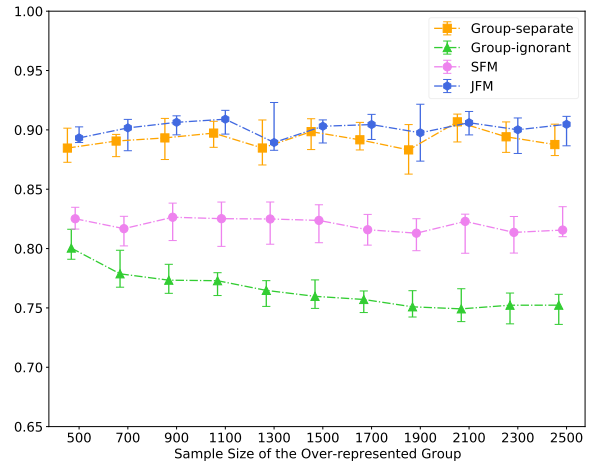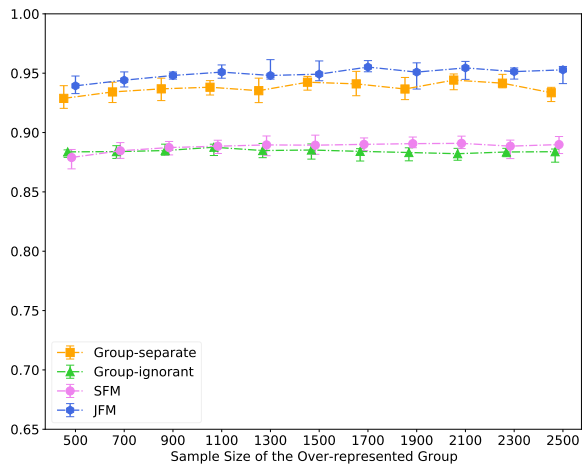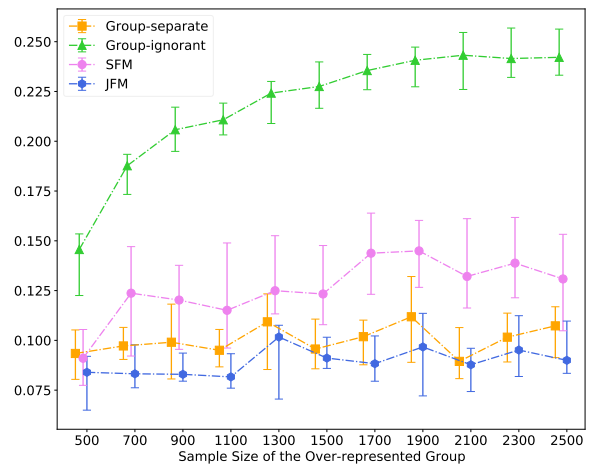
(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure S.10: Experimental Results for Scenario 2B

59

(a) AUC of the Under-represented Group

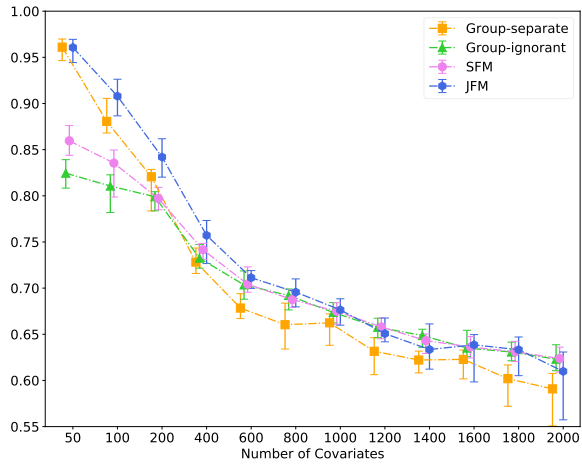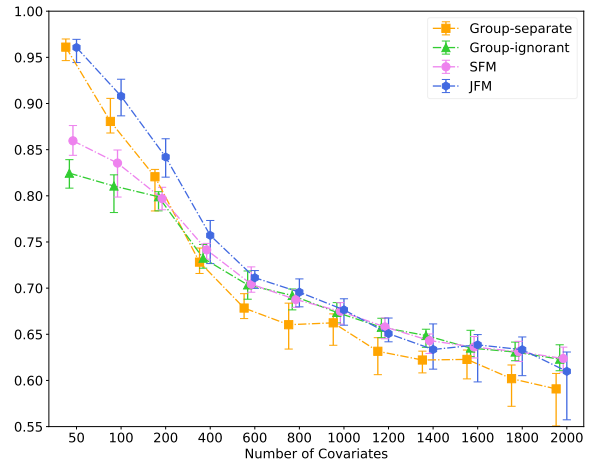(b) AUC of the Over-represented Group

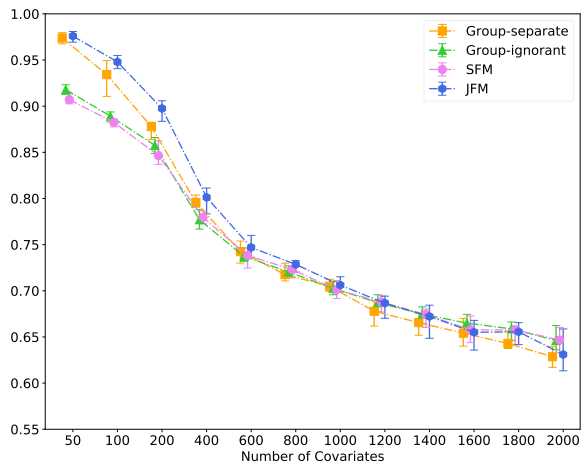(c) Overall AUC

(d) Disparity of AUC

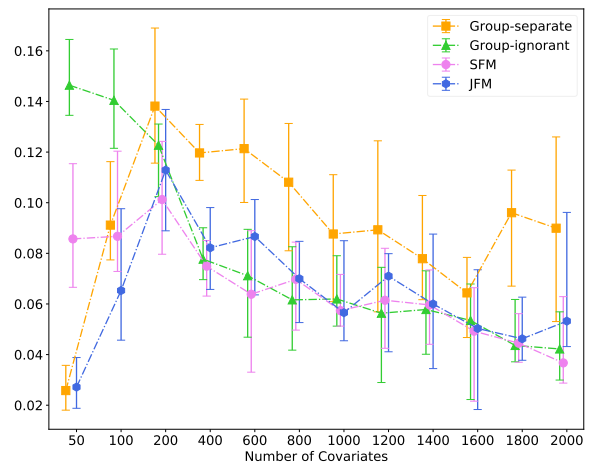Figure S.11: Experimental Results for Scenario 3B

(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure S.12: Experimental Results for Scenario 4B

S.12 Python Implementation

We provide a Python implementation to reproduce the simulation study results. The codes will be available at `https://github.com/hyungrok-do/joint-fairness-model`.

Dependencies:

- anaconda3 ($\geq$ 4.8.3)

- cython ($\geq$ 0.29.8)

- scipy ($\geq$ 1.6.2)

- numpy ($\geq$ 1.17.0)

- pandas ($\geq$ 1.2.4)

- matplotlib ($\geq$ 3.1.1)

- scikit-learn ($\geq$ 0.24.1)

Install: Users have to compile the enclosed cython source code (tested on Windows 10, macOS Catalina 10.15.7, and Red Hat Enterprise Linux 8.2.) After unzipping or cloning the git, type

```
python setup.py build_ext --inplace.
```

Reproducing the results: We provide shell/slurm scripts to run the repeated experiments to reproduce the results. For the results of scenarios 1 through 4 and the supplementary results scenarios 1B through 4B, use `run-simulation.sh` or `run-simulation.s`. To draw the plots, run `visualization-simulation-results.py`.

For the experiments for validation measures, execute `run-validation-measure.sh` or `run-validation-measure.s`. To draw the plots, run `visualization-validation-measures.py`.

Executing `experiment-computation-time-p.py` and `experiment-computation-time-n.py` will produce the Figure S.1 (a) and (b), respectively.

References – Supplementary

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Xi Chen, Qihang Lin, Seyoung Kim, Jaime G Carbonell, Eric P Xing, et al. Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752, 2012.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

Keith Knight and Wenjiang Fu. Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378, 2000.

R Tyrrell Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1970.

Vivian Viallon, Sophie Lambert-Lacroix, Holger Höfling, and Franck Picard. Adaptive generalized fused-lasso: Asymptotic properties and applications. 2013. doi: hal-00813281f.