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

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

[1]Department of Population Health, NYU 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

Web Appendix.1 Accelerated Smoothing Proximal Gradient Algorithm for JFM

In this section, we present a complete description of an accelerated smoothing proximal gradient (ASPG) algorithm (Chen et al., 2012) to solve Problem (2) for JFM, which briefly introduced in Section 3 of the main text. 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 the proximal operator associated with $\mathcal{P}_{\text{FPR}}$ and $\mathcal{P}_{\text{FNR}}$.

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 and then apply the Nesterov smoothing approximation (Nesterov, 2005). We start with matrix representations of the fairness penalty terms $\mathcal{P}_{\text{FPR}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\text{F}}) = \lambda_{\text{F}} \left\| \mathbf{D}_0 \boldsymbol{\beta} \right\|_1$ and $\mathcal{P}_{\text{FNR}}(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}, \lambda_{\text{F}}) = \lambda_{\text{F}} \left\| \mathbf{D}_1 \boldsymbol{\beta} \right\|_1$, where $\mathbf{D}_y \in \mathbb{R}^{K(K-1)/2 \times pK}$ is defined as below. Similarly, we use the matrix representation of the similarity penalty $\mathcal{P}_{\text{Sim}}(\boldsymbol{\beta}; \lambda_{\text{Sim}}) = \lambda_{\text{Sim}} \left\| \mathbf{F} \boldsymbol{\beta} \right\|_1$ with $\mathbf{F}$ defined as below.

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

Here, $\bar{\mathbf{X}}_{ky} = \frac{1}{|S_{ky}|} \sum_{i \in S_{ky}} \mathbf{X}_i$ is the average predictor vector for group k with true outcome y, $\mathbf{I}_p$ is the p-dimensional identity matrix. The matrix form of the fairness penalty term and the similarity penalty term is therefore defined as:

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

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

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; -\sum_{k=1}^{K} \frac{1}{n_k} \ell(\boldsymbol{\beta}_k; \mathbf{X}_k, \mathbf{y}_k) + \|\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta}\|_1 + \sum_{k=1}^{K} \lambda_{\mathrm{Sp}_k} \|\boldsymbol{\beta}_k\|_1, \tag{A.1}$$

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 L_1 norm is the L_∞ norm, we have

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

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

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

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 A.1: *For any $\mu > 0$, the Nesterov smooth approximation satisfies the following inequalities:*

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

Proof: See Web Appendix 4.

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 A.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:*

$$\nabla f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}) = \mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}}^T \boldsymbol{\alpha}^*, \tag{A.3}$$

where $\boldsymbol{\alpha}^* = \operatorname{argmax}\left\{\boldsymbol{\alpha}^T \mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta} - \frac{\mu}{2}\|\boldsymbol{\alpha}\|_2^2 : \|\boldsymbol{\alpha}\|_\infty \leqslant 1\right\}$. 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 Web Appendix 4.

The following proposition yields $\boldsymbol{\alpha}^*$ in Proposition A.2, which is essential to compute the gradient $\nabla f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}})$.

Proposition A.3: *For any $\mu > 0$, we have*

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

where $\mathcal{S}_\infty(\cdot)$ is the projection onto the unit L_∞ ball such that $[\mathcal{S}_\infty(\boldsymbol{x})]_i = x_i \mathbb{I}_{[-1,1]}(x_i) + \mathbb{I}_{(1,\infty)}(x_i) - \mathbb{I}_{(-\infty,-1)}(x_i)$, where $\mathbb{I}$ is the indicator function.

Proof: See Web Appendix 4.

Computational Remark: The computational complexities of $\mathbf{D}^T_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\alpha}^*$ and $\mathbf{D}_{\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}} \boldsymbol{\beta}$ is quadratic in p, making them each computationally intensive when p is large. However, the computations can be substituted by a series of scalar multiplications and vector addition, reducing the complexities to be linear in p. Details are provided in Web Appendix Web Appendix.5.

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 (A.1) becomes

$$\underset{\boldsymbol{\beta}}{\text{minimize}}\ \tilde{F}(\boldsymbol{\beta}) = -\sum_{k=1}^{K} \frac{1}{n_k} \ell(\boldsymbol{\beta}_k; \mathbf{X}_k, \mathbf{y}_k) + f_\mu(\boldsymbol{\beta}; \lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}) + \sum_{k=1}^{K} \lambda_{\mathrm{Sp}_k} \|\boldsymbol{\beta}_k\|_1, \tag{A.4}$$

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., 2007).

Algorithm 1 (can be found in the main text) 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 (A.3). 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 A.1 *Let* $\{\boldsymbol{\beta}^{(t)} : t = 1, 2, \cdots\}$ *be a sequence generated by Algorithm 1. Then for any* $t \geqslant 1$,

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

where $\boldsymbol{\beta}^*$ *is a global minimizer of problem* (A.4).

Proof: See Web Appendix 4.

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* $t \geqslant 1$,

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

where $\boldsymbol{\beta}^*$ *and* $\boldsymbol{\beta}^{**}$ *are global minimizers of problem* (A.4) *and problem (2), respectively, and* L *is the Lipschitz constant of* $\tilde{F}$ *presented in Lemma A.1.*

Proof: See Web Appendix 4.

Given the desired accuracy $\delta > 0$ for the approximation, we set $\mu = \frac{2\delta}{pK}$. Then, we have $F(\boldsymbol{\beta}^{(t)}) - F(\boldsymbol{\beta}^{**}) \leqslant \delta + \frac{2L\|\boldsymbol{\beta}^{0} - \boldsymbol{\beta}^{*}\|_2^2}{t^2}$. 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 Theorem 3.1, we present the rate of convergence of the algorithm in the following proposition.

PROPOSITION A.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 Web Appendix 4.

Finally, we provide the time complexity of a single iteration of Algorithm 1. PROPOSITION A.4: *Time complexity of a single iteration of Algorithm 1 is $\mathcal{O}((n + K^2)pK)$.*

Proof: See Web Appendix 4.

Web Appendix.2 Accelerated Smoothing Proximal Gradient Algorithm for JFM with Group Lasso Similarity Term

With the group lasso-like similarity term, which encourages simultaneous selection but not encourage the estimated values to be similar, the JFM formulation is

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; -\sum_{k=1}^{K} \frac{1}{n_k} \ell(\boldsymbol{\beta}_k; \mathbf{X}_k, \mathbf{y}_k) + \|\mathbf{D}_{\lambda_\mathrm{F}}\|_1 + \sum_{k=1}^{K} \lambda_{\mathrm{Sp}_k} \|\boldsymbol{\beta}_k\|_1 + \lambda_{\mathrm{Sim}} \sum_{j=1}^{p} \sqrt{K} \|\boldsymbol{\beta}_{[j]}\|_2,$$

where $\boldsymbol{\beta}_{[j]} = (\beta_{j+1}, \cdots, \beta_{j+K})$ is the vector of coefficients associated covariate j. We can apply the same Nesterov approximation technique on $\|\mathbf{D}_{\lambda_\mathrm{F}}\|_1$ and other two terms are the same as in sparse group lasso, which can be handled via a well-known proximal operator. The pseudocode of the algorithm for solving the problem is given in Algorithm 2. Note $\mathcal{S}$ is the soft-thresholding operator and $(x)_+ = \max\{0, x\}$.

Web Appendix.3 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 can easily handle. However, CVXPY is equipped with a general quadratic optimization solver and 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:

$$\underset{\boldsymbol{\beta}}{\text{minimize}} \; -\ell(\boldsymbol{\beta}; \mathbf{X}, \mathbf{y}) + \lambda_{\mathrm{F}_0} \sum_{j<k} |(\bar{\mathbf{X}}_{j0} - \bar{\mathbf{X}}_{k0})\boldsymbol{\beta}| + \lambda_{\mathrm{F}_1} \sum_{j<k} |(\bar{\mathbf{X}}_{j1} - \bar{\mathbf{X}}_{k1})\boldsymbol{\beta}| + \lambda_{\mathrm{Sp}} \|\boldsymbol{\beta}\|_1. \quad \text{(A.5)}$$

Algorithm 2 Accelerated Smoothing Proximal Gradient Method for JFM-Group

```
Input: Data $\mathbf{X}_k, \mathbf{y}_k$ for $k = 1, \cdots, K$, hyperparameters $\lambda_{\mathrm{F}}, \lambda_{\mathrm{Sim}}, \lambda_{\mathrm{Sp}}$, $\epsilon$, $\mu$
Output: $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1, \cdots, \hat{\boldsymbol{\beta}}_K)$ solving the Joint Fairness optimization problem.
Initialize: $\boldsymbol{\beta}^{(0)} = \mathbf{0},\ \boldsymbol{\gamma}^{(0)} = \mathbf{0},\ s^{(0)} = 1$
Compute $L = \frac{1}{4} \max \left\{ \lambda_{\max}(\mathbf{X}_k^T \mathbf{X}_k) : k = 1, \cdots, K \right\} + \mu^{-1} \|\mathbf{D}_{\lambda_{\mathrm{F}}}\|_2^2$
for $t \geqslant 1$ do
    $\boldsymbol{\alpha}^{(t)} = \boldsymbol{\gamma}^{(t-1)} - L^{-1} \left( -\nabla \ell \left( \boldsymbol{\gamma}^{(t-1)} \right) + \nabla f_\mu \left( \boldsymbol{\gamma}^{(t-1)} \right) \right)$
    for $j = 1, \cdots, p$ do
        $\boldsymbol{\beta}_{[j]}^{(t)} = \left( 1 - \frac{L^{-1} \lambda_{\mathrm{Sim}} \sqrt{K}}{\|\mathcal{S}(\boldsymbol{\alpha}_{[j]}^{(t)}; L^{-1} \lambda_{\mathrm{Sp}})\|_2} \right)_+ \mathcal{S} \left( \boldsymbol{\alpha}_{[j]}^{(t)}; L^{-1} \lambda_{\mathrm{Sp}} \right)$
    if $\|\boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)}\|_2 \leqslant \epsilon$ break
    $s^{(t)} = \frac{1 + \sqrt{1 + 4 {s^{(t-1)}}^2}}{2}$
    $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\beta}^{(t)} + \left( \frac{s^{(t-1)} - 1}{s^{(t)}} \right) \left( \boldsymbol{\beta}^{(t)} - \boldsymbol{\beta}^{(t-1)} \right)$
    $t \leftarrow t + 1$
end for
$\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(t)}$.
```

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

$$\mathbf{D}_{\lambda_{\mathrm{F}}} = \begin{pmatrix} \lambda_{\mathrm{F}_0}(\bar{\mathbf{X}}_{10} - \bar{\mathbf{X}}_{20}) \\ \vdots \\ \lambda_{\mathrm{F}_0}(\bar{\mathbf{X}}_{K-1\ 0} - \bar{\mathbf{X}}_{K0}) \\ \lambda_{\mathrm{F}_1}(\bar{\mathbf{X}}_{11} - \bar{\mathbf{X}}_{21}) \\ \vdots \\ \lambda_{\mathrm{F}_1}(\bar{\mathbf{X}}_{K-1\ 1} - \bar{\mathbf{X}}_{K1}) \end{pmatrix}.$$

Algorithm 3 Accelerated Smoothing Proximal Gradient Method for SFM

```
Input: Data $\mathbf{X}, \mathbf{y}$, hyperparameters $\lambda_{\mathrm{F}_0}, \lambda_{\mathrm{F}_1}, \lambda_{\mathrm{Sp}}$, $\epsilon$, $\mu$
Output: $\hat{\boldsymbol{\beta}}$ solving the Single Fairness optimization problem (A.5).
Initialize: $\boldsymbol{\beta}^{(0)} = \mathbf{0}, \ \boldsymbol{\gamma}^{(0)} = \mathbf{0}, \ s^{(0)} = 1$
Compute $L = \frac{1}{4}\lambda_{\max}(\mathbf{X}^T\mathbf{X}) + \mu^{-1}\|\mathbf{D}_{\lambda_{\mathrm{F}}}\|_2^2$
for $m \geqslant 1$ do
    $\boldsymbol{\alpha}^{(m)} = \boldsymbol{\gamma}^{(m-1)} - L^{-1}\left(-\nabla \ell\left(\boldsymbol{\gamma}^{(m-1)}\right) + \nabla f_\mu\left(\boldsymbol{\gamma}^{(m-1)}\right)\right)$
    $\boldsymbol{\beta}^{(m)} = \mathcal{S}\left(\boldsymbol{\alpha}^{(m)}; L^{-1}\lambda_{\mathrm{Sp}}\right)$
    if $\|\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}\|_2 \leqslant \epsilon$ break
    $s^{(m)} = \frac{1+\sqrt{1+4s^{(m-1)2}}}{2}$
    $\boldsymbol{\gamma}^{(m)} = \boldsymbol{\beta}^{(m)} + \left(\frac{s^{(m-1)}-1}{s^{(m)}}\right)\left(\boldsymbol{\beta}^{(m)} - \boldsymbol{\beta}^{(m-1)}\right)$
    $m \leftarrow m + 1$
end for
$\hat{\boldsymbol{\beta}} \leftarrow \boldsymbol{\beta}^{(m)}$.
```

We can easily verify the Nesterov smooth approximation can be applied to approximate $\|\mathbf{D}_{\lambda_{\mathrm{F}}}\boldsymbol{\beta}\|_1$ and Propositions A.1, A.2, and A.3 hold. Therefore, Algorithm 3 solves the SFM optimization problem.

Web Appendix.4 Proofs

Proof of Proposition 2.1

Let $\sigma(x) = \frac{1}{1+\exp(-x)}$ be a sigmoid function. Since σ is differentiable, by the mean value theorem, we have

$$\sigma(x_j) - \sigma(x_k) = \sigma'(z)(x_j - x_k),$$

for any x_j and x_k, where $z = \alpha x_j + (1-\alpha)x_k$, $\alpha \in [0,1]$. Since $0 \leqslant \sigma'(z) = \sigma(z)(1-\sigma(z)) \leqslant \frac{1}{4}$, we have

$$-\frac{1}{4}(x_j - x_k) \leqslant \sigma(x_j) - \sigma(x_k) \leqslant \frac{1}{4}(x_j - x_k).$$

By letting $x_j = \mathbf{x}_{jy}\boldsymbol{\beta}_j$, we obtain

$$-\frac{1}{4}\iint \left(\mathbf{x}_{jy}\boldsymbol{\beta}_j - \mathbf{x}_{ky}\boldsymbol{\beta}_k\right) dp(\mathbf{x}_{jy}, \mathbf{x}_{ky}) \leqslant \iint \left(\sigma(\mathbf{x}_{jy}\boldsymbol{\beta}_j) - \sigma(\mathbf{x}_{ky}\boldsymbol{\beta}_k)\right) dp(\mathbf{x}_{jy}, \mathbf{x}_{ky})$$
$$\leqslant \frac{1}{4}\iint \left(\mathbf{x}_{jy}\boldsymbol{\beta}_j - \mathbf{x}_{ky}\boldsymbol{\beta}_k\right) dp(\mathbf{x}_{jy}, \mathbf{x}_{ky}),$$

where p is the joint probability density function of $\mathbf{X}_{jy}$ and $\mathbf{X}_{ky}$ which represent $\mathbf{X}_j|Y = y$ and $\mathbf{X}_k|Y = y$, respectively. Therefore,

$$-\frac{1}{4}\left(\mathbb{E}[\mathbf{X}_{jy}\boldsymbol{\beta}_j - \mathbf{X}_{ky}\boldsymbol{\beta}_k]\right) \leqslant \mathbb{E}[\sigma(\mathbf{X}_{jy}\boldsymbol{\beta}_j) - \sigma(\mathbf{X}_{ky}\boldsymbol{\beta}_k)] \leqslant \frac{1}{4}\left(\mathbb{E}[\mathbf{X}_{jy}\boldsymbol{\beta}_j - \mathbf{X}_{ky}\boldsymbol{\beta}_k]\right).$$

By the linearity of the expectation, we can rewrite the inequality as follows:

$$\left|\mathbb{E}[\sigma(\mathbf{X}_{jy}\boldsymbol{\beta}_j)] - \mathbb{E}[\sigma(\mathbf{X}_{ky}\boldsymbol{\beta}_k)]\right| \leqslant \frac{1}{4}\left|\mathbb{E}[\mathbf{X}_{jy}\boldsymbol{\beta}_j] - \mathbb{E}[\mathbf{X}_{ky}\boldsymbol{\beta}_k]\right|,$$

which is equivalent to the statement given in Proposition 2.1.

Proof of Proposition A.1

Note that the proof of Propositions A.1, A.2, and A.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

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

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

Proof of Proposition A.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 $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_\infty \leqslant 1\}$, which is given by

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

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

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

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 $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_\infty \leqslant 1\}$ 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,

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

Since $\{\boldsymbol{\alpha} : \|\boldsymbol{\alpha}\|_\infty \leqslant 1\}$ 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 $\|\boldsymbol{\alpha}\|_\infty \leqslant 1$. 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

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

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

Proof of Proposition A.3

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

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

which can be rewritten as the following minimization problem

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

It is equivalent to

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

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.

Proof of Lemma A.1

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

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

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

Proof of Theorem 3.1 We can easily verify the inequality by applying Proposition A.1 and Lemma A.1, and using $\tilde{F}(\boldsymbol{\beta}^*) \leqslant \tilde{F}(\boldsymbol{\beta}^{**})$ as below:

$$\begin{aligned} F(\boldsymbol{\beta}^{(t)}) - F(\boldsymbol{\beta}^{**}) &= \left(F(\boldsymbol{\beta}^{(t)}) - \tilde{F}(\boldsymbol{\beta}^{(t)})\right) + \left(\tilde{F}(\boldsymbol{\beta}^{(t)}) - \tilde{F}(\boldsymbol{\beta}^*)\right) + \left(\tilde{F}(\boldsymbol{\beta}^*) - F(\boldsymbol{\beta}^{**})\right) \\ &\leqslant \frac{\mu p K}{2} + \frac{2L\|\boldsymbol{\beta}^{(0)} - \boldsymbol{\beta}^*\|_2^2}{t^2} + 0. \end{aligned}$$

Proof of Proposition A.4

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

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

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

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

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

Proof of Proposition A.5

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.

Proof of Theorem 4.1

To prove the theorem, it is sufficient to show that $\mathcal{V}_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) \to \mathcal{V}(\mathbf{u}_1, \cdots, \mathbf{u}_K)$ as $\min_{k=1,\cdots,K} n_k \to \infty$, where $\mathcal{V}_n(\mathbf{u}_1, \cdots, \mathbf{u}_K)$ is defined in (A.6).

From Theorem 4.1, we can re-write $\mathcal{V}(\mathbf{u}_1, \cdots, \mathbf{u}_K)$ as

$$\mathcal{V}(\mathbf{u}_1, \cdots, \mathbf{u}_K) = g(\mathbf{u}_1, \cdots, \mathbf{u}_K) + h(\mathbf{u}_1, \cdots, \mathbf{u}_K)$$

where

$$g(\mathbf{u}_1, \cdots, \mathbf{u}_K) = \sum_{k=1}^K \mathbf{u}_k^T \mathbf{W}_k + \frac{1}{2}\sum_{k=1}^K \mathbf{u}_k^T \mathbf{C}_k \mathbf{u}_k$$

and

$$\begin{aligned} h(\mathbf{u}_1, \cdots, \mathbf{u}_K) = {} & \lambda_{\mathrm{F}}^{(0)} \sum_{j<k} \sum_{y\in\{0,1\}} \mathcal{T}(\bar{\mathbf{X}}_{jy}\mathbf{u}_j - \bar{\mathbf{X}}_{ky}\mathbf{u}_k, \bar{\mathbf{X}}_{jy}\boldsymbol{\beta}_j - \bar{\mathbf{X}}_{ky}\boldsymbol{\beta}_k) \\ & + \lambda_{\mathrm{Sim}}^{(0)} \sum_{j<k}\sum_{l=1}^p \mathcal{T}(u_{jl} - u_{kl}, \beta_{jl} - \beta_{kl}) + \lambda_{\mathrm{Sp}}^{(0)} \sum_{k=1}^K \sum_{l=1}^p \mathcal{T}(u_{kl}, \beta_{kl}). \end{aligned}$$

Let

$$\mathcal{V}_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) = g_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) + h_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) \tag{A.6}$$

where

$$g_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) = -\sum_{k=1}^K \left\{ \ell\left(\boldsymbol{\beta}_k + \frac{\mathbf{u}_k}{\sqrt{n}}\right) - \ell(\boldsymbol{\beta}_k)\right\},$$

and

$$\begin{aligned} h_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) =& \lambda_{\mathrm{F}} \sum_{j<k} \sum_{y \in \{0,1\}} \left\{ \left| \bar{\mathbf{X}}_{jy} \left(\boldsymbol{\beta}_j + \frac{\mathbf{u}_j}{\sqrt{n}} \right) - \bar{\mathbf{X}}_{ky} \left(\boldsymbol{\beta}_k + \frac{\mathbf{u}_k}{\sqrt{n}} \right) \right| - \left| \bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j - \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k \right| \right\} \\ &+ \lambda_{\mathrm{Sim}} \sum_{j<k} \sum_{y \in \{0,1\}} \left\{ \left| \left(\beta_{jl} + \frac{u_{jl}}{\sqrt{n}} \right) - \left(\beta_{kl} + \frac{u_{kl}}{\sqrt{n}} \right) \right| - |\beta_{jl} - \beta_{kl}| \right\} \\ &+ \lambda_{\mathrm{Sp}} \sum_{k=1}^{K} \sum_{l=1}^{p} \left\{ \left| \beta_{kl} + \frac{u_{kl}}{\sqrt{n}} \right| - |\beta_{kl}| \right\}. \end{aligned} \tag{A.7}$$

We first show $g_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) \to g(\mathbf{u}_1, \cdots, \mathbf{u}_K)$, that is,

$$\ell \left(\boldsymbol{\beta}_k + \frac{\mathbf{u}_k}{\sqrt{n}} \right) - \ell(\boldsymbol{\beta}_k) \to \mathbf{u}_k^T \mathbf{W}_k + \frac{1}{2} \mathbf{u}_k^T \mathbf{C}_k \mathbf{u}_k. \tag{A.8}$$

Following the arguments of Viallon et al. (2013), we apply Taylor series expansion on the left side of (A.8) which yields

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

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

On the other hand, to show $h_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) \to h(\mathbf{u}_1, \cdots, \mathbf{u}_K)$, we follow the arguments in Theorem 2 of Knight and Fu (2000). For the first term of (A.7), we have

$$\begin{aligned} &\lambda_{\mathrm{F}}^{(n)} \sum_{j<k} \sum_{y \in \{0,1\}} \left\{ \left| \bar{\mathbf{X}}_{jy} \left(\boldsymbol{\beta}_j + \frac{\mathbf{u}_j}{\sqrt{n}} \right) - \bar{\mathbf{X}}_{ky} \left(\boldsymbol{\beta}_k + \frac{\mathbf{u}_k}{\sqrt{n}} \right) \right| - \left| \bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j - \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k \right| \right\} \\ &\quad = \lambda_{\mathrm{F}}^{(n)} \sum_{j<k} \sum_{y \in \{0,1\}} \left\{ \left| \bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j - \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k + \frac{\bar{\mathbf{X}}_{jy} \mathbf{u}_j - \bar{\mathbf{X}}_{ky} \mathbf{u}_k}{\sqrt{n}} \right| - \left| \bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j - \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k \right| \right\} \\ &\quad \to \lambda_{\mathrm{F}}^{(0)} \sum_{j<k} \sum_{y \in \{0,1\}} \left\{ (\bar{\mathbf{X}}_{jy} \mathbf{u}_j - \bar{\mathbf{X}}_{ky} \mathbf{u}_k) \mathrm{sign}(\bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j - \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k) \mathbb{I}(\bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j \neq \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k) \right. \\ &\qquad\qquad \left. + |\bar{\mathbf{X}}_{jy} \mathbf{u}_j - \bar{\mathbf{X}}_{ky} \mathbf{u}_k| \mathbb{I}(\bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j = \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k) \right\}, \end{aligned}$$

as $n \to \infty$. Similarly,

$$\begin{aligned}
\lambda_{\text{Sim}}^{(n)} \sum_{j<k} \sum_{l=1}^{p} & \left\{ \left| \left(\beta_{jl} + \frac{u_{jl}}{\sqrt{n}} \right) - \left(\beta_{kl} + \frac{u_{kl}}{\sqrt{n}} \right) \right| - |\beta_{jl} - \beta_{kl}| \right\} \\
&= \lambda_{\text{Sim}}^{(n)} \sum_{j<k} \sum_{l=1}^{p} \left\{ \left| \beta_{jl} - \beta_{kl} + \frac{u_{jl} - u_{kl}}{\sqrt{n}} \right| - |\beta_{jl} - \beta_{kl}| \right\} \\
&\to \lambda_{\text{Sim}}^{(0)} \sum_{j<k} \sum_{l=1}^{p} \left\{ (u_{jl} - u_{kl}) \text{sign}(\beta_{jl} - \beta_{kl}) \mathbb{I}(\beta_{jl} \neq \beta_{kl}) + |u_{jl} - u_{kl}| \mathbb{I}(\beta_{jl} = \beta_{kl}) \right\},
\end{aligned}$$

as $n \to \infty$. We also have

$$\lambda_{\text{Sp}}^{(n)} \sum_{k=1}^{K} \sum_{l=1}^{p} \left\{ \left| \beta_{kl} + \frac{u_{kl}}{\sqrt{n}} \right| - |\beta_{kl}| \right\} \to \lambda_{\text{Sp}}^{(0)} \sum_{k=1}^{K} \sum_{l=1}^{p} \{ u_{kl} \text{sign}(\beta_{kl}) \mathbb{I}(\beta_{kl} \neq 0) + |u_{kl}| \mathbb{I}(\beta_{kl} = 0) \},$$

as $n \to \infty$.

We showed that $g_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) \to g(\mathbf{u}_1, \cdots, \mathbf{u}_K)$ and $h_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) \to h(\mathbf{u}_1, \cdots, \mathbf{u}_K)$ as $n \to \infty$. Thus, $\mathcal{V}_n(\mathbf{u}_1, \cdots, \mathbf{u}_K) \to \mathcal{V}(\mathbf{u}_1, \cdots, \mathbf{u}_K)$ as $n \to \infty$ as desired.

Note: Theorem 4.1 is proved for the JFM with L_1 penalization. For a model defined with L_2 penalization, we can simply modify $h(\mathbf{u}_1, \cdots, \mathbf{u}_K)$ as below.

$$\begin{aligned}
h(\mathbf{u}_1, \cdots, \mathbf{u}_K) &= \lambda_{\text{F}}^{(0)} \sum_{j<k} \sum_{y \in \{0,1\}} (\bar{\mathbf{X}}_{jy} \mathbf{u}_j - \bar{\mathbf{X}}_{ky} \mathbf{u}_k) \text{sign}(\bar{\mathbf{X}}_{jy} \boldsymbol{\beta}_j - \bar{\mathbf{X}}_{ky} \boldsymbol{\beta}_k) |\bar{\mathbf{X}}_{jy} \mathbf{u}_j - \bar{\mathbf{X}}_{ky} \mathbf{u}_k| \\
+ \lambda_{\text{Sim}}^{(0)} \sum_{j<k} \sum_{l=1}^{p} & \{ (u_{jl} - u_{kl}) \text{sign}(\beta_{jl} - \beta_{kl}) |u_{jl} - u_{kl}| \} + \lambda_{\text{Sp}}^{(0)} \sum_{k=1}^{K} \sum_{l=1}^{p} \{ u_{kl} \text{sign}(\beta_{kl}) |u_{kl}| \}.
\end{aligned}$$

Following Knight and Fu (2000) and the arguments above, we can show $\sqrt{n}$-consistency of the estimates obtained from a model with L_1 penalization. The consistency of estimates obtained form a model utilizing mixture of L_1 and L_2 penalization can be proved similarly.

Web Appendix.5 Computational Remark

Although $\mathbf{D}_{\lambda_{\text{F}}, \lambda_{\text{Sim}}}^{T} \boldsymbol{\alpha}^*$ and $\mathbf{D}_{\lambda_{\text{F}}, \lambda_{\text{Sim}}} \boldsymbol{\beta}$ in Proposition A.2 and A.3 seem computationally expensive due to the high-dimensionality of $\mathbf{D}_{\lambda_{\text{F}}, \lambda_{\text{Sim}}} \in \mathbb{R}^{(p+1)K(K-1) \times pK}$, we can reduce the complexity because of their structure.

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

$$\mathbf{D}_{\lambda_{\text{F}}, \lambda_{\text{Sim}}}^{T} \boldsymbol{\alpha}^* = \lambda_{\text{F}} \mathbf{D}_0 \alpha_1^* + \lambda_{\text{F}} \mathbf{D}_1 \alpha_2^* + \lambda_{\text{Sim}} \mathbf{A}^*, \tag{A.9}$$

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. (A.9) 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

$$\mathbf{D}_{\lambda_{\mathrm{F}},\lambda_{\mathrm{Sim}}}\boldsymbol{\beta} = \begin{pmatrix} \lambda_{\mathrm{F}}\mathbf{D}_0\boldsymbol{\beta} \\ \lambda_{\mathrm{F}}\mathbf{D}_1\boldsymbol{\beta} \\ \lambda_{\mathrm{Sim}}\mathbf{F}\boldsymbol{\beta} \end{pmatrix}. \tag{A.10}$$

Here, $\mathbf{F}$ is a sparse matrix consists of identity matrices and thus $\mathbf{F}\boldsymbol{\beta}$ can be computed without matrix multiplication by

$$\mathbf{F}\boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_{K-1} - \boldsymbol{\beta}_K \end{pmatrix}. \tag{A.11}$$

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 (A.11) only requires a series of vector subtraction operations, it is more efficient than multiplying large matrices. Note that the complexity of (A.10) is also $\mathcal{O}(pK^2)$ because $\mathbf{D}_0\boldsymbol{\beta}$ and $\mathbf{D}_1\boldsymbol{\beta}$ both require $\mathcal{O}(pK)$ computations.

Web Appendix.6 Computational Analysis

[Web Figure 1 about here.]

Figure 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 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 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 the problem is ill-posed ($p > n$) and requires more iterations for convergence.

Web Appendix.7 Details of Choice of Hyperparameters

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). Web Figures 2, 3, and 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.

[Web Figure 2 about here.]

[Web Figure 3 about here.]

[Web Figure 4 about here.]

Web Appendix.8 Additional Plots for Simulation Study

[Web Figure 5 about here.]

[Web Figure 6 about here.]

[Web Figure 7 about here.]

[Web Figure 8 about here.]

Web Appendix.9 Additional Simulation Scenarios

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

- **Scenario 4 (Sensitivity with respect to baseline prevalence):** The outcome prevalence ranged from $10 - 50\%$ for group 2, and fixed at 50% for group 1. The sample sizes were set at 500 and 200 for group 1 and 2 respectively. The number of features was $p = 100$, with half of the features with non-zero coefficients were shared.

- **Scenario 1B (Sensitivity with respect to model difference):** 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 $p = 100$.
- **Scenario 2B (Sensitivity with respect to sample size):** The samples size of the over-represented group ranged from 500 to $2,500$ with the sample size of the under-represented group fixed at 200.
- **Scenario 3B (Sensitivity with respect to dimensionality):** The number of features p ranged from 50 to 2,000. Everything is the same with the Scenario 3, except that for each p, 30% of the features had non-zero coefficients.
- **Scenario 4B (Sensitivity with respect to baseline prevalence):** 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%.

As same as with the Section 5, we evaluated the methods on independent testing datasets under the same setups with large sample sizes (both $1,000$). AUC was used to evaluate the predictive performance of each model.

Figure 9 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 group fixed. In Figure 9(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 9(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 9(c) and 9(d), the JFM achieves overall satisfactory AUCs and parity between groups with varying sample sizes of the under-represented group. Web Figure 8(a) through 8(d) compares the average of TPR and TNR and disparity in TPR and TNR differences of the four methods.

[Web Figure 9 about here.]

[Web Figure 10 about here.]

[Web Figure 11 about here.]

[Web Figure 12 about here.]

[Web Figure 13 about here.]

Web Appendix.10 Variable Selection Performances

[Web Figure 14 about here.]

Web Appendix.11 Comparison of the Fusion and Group Similarity Penalties

[Web Figure 15 about here.]

Web Appendix.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 ($\geqslant$ 4.8.3)

- Cython ($\geqslant$ 0.29.8)

- scipy ($\geqslant$ 1.6.2)

- numpy ($\geqslant$ 1.17.0)

- pandas ($\geqslant$ 1.2.4)

- matplotlib ($\geqslant$ 3.1.1)

- scikit-learn ($\geqslant$ 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 1 (a) and (b), respectively.

References

Bechavod, Y. and Ligett, K. (2017). Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

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

Chen, X., Lin, Q., Kim, S., Carbonell, J. G., Xing, E. P., et al. (2012). Smoothing proximal gradient method for general structured sparse regression. *The Annals of Applied Statistics*, 6(2):719–752.

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

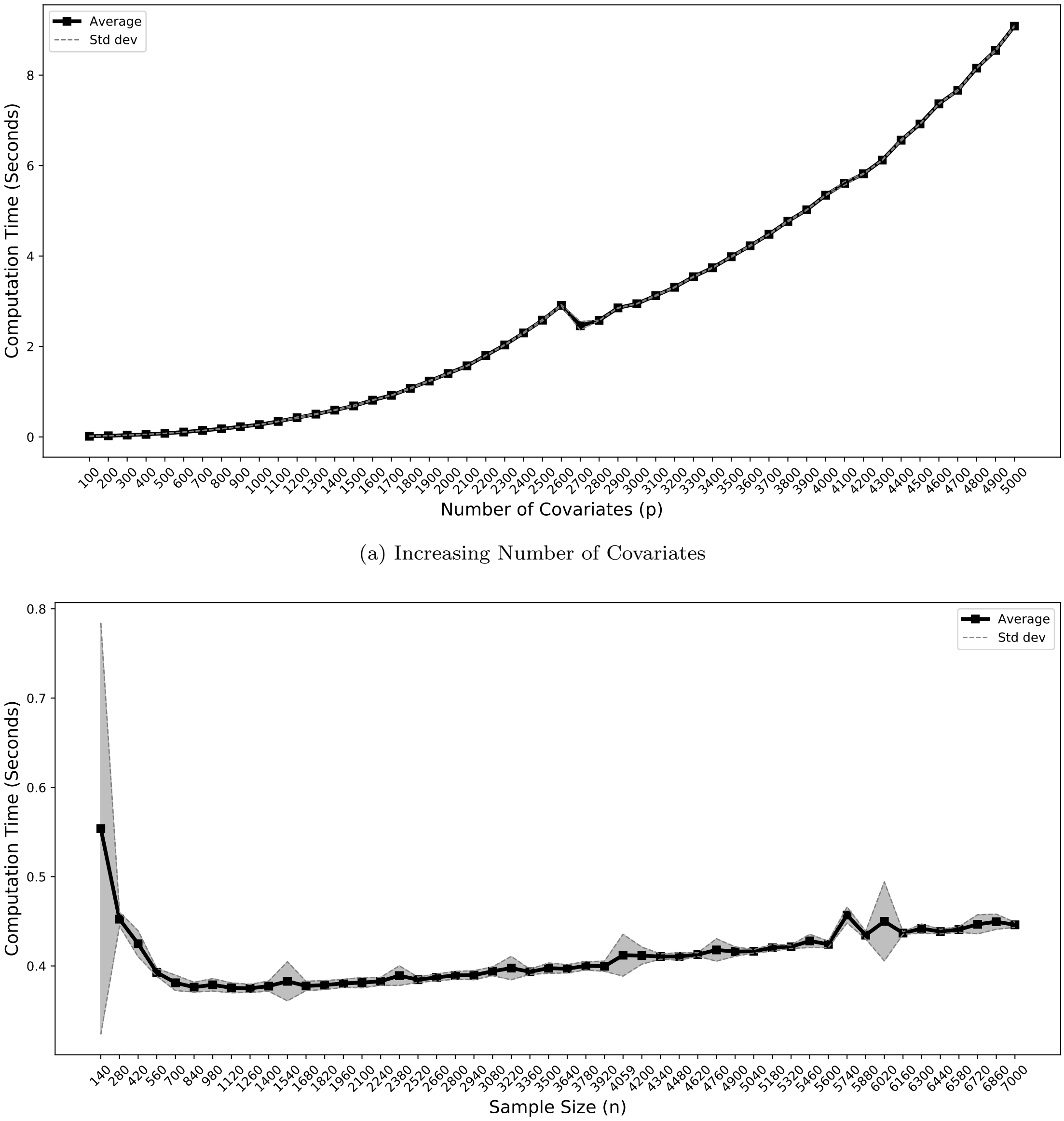
Friedman, J., Hastie, T., Höfling, H., Tibshirani, R., et al. (2007). Pathwise coordinate optimization. *The annals of applied statistics*, 1(2):302–332.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of statistics*, pages 1356–1378.

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

Rockafellar, R. T. (1970). *Convex analysis*. Number 28. Princeton university press.

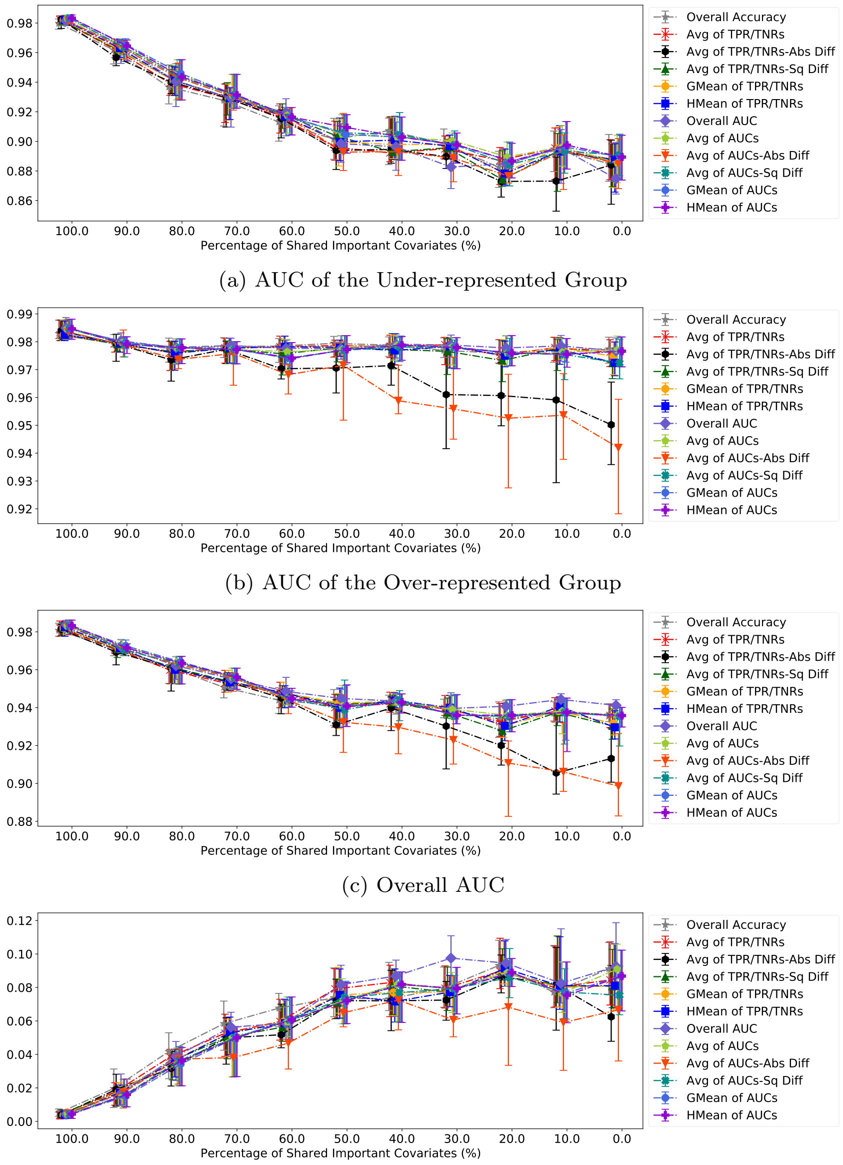
Viallon, V., Lambert-Lacroix, S., Höfling, H., and Picard, F. (2013). Adaptive Generalized Fused-Lasso: Asymptotic Properties and Applications. working paper or preprint.



(a) Increasing Number of Covariates

(b) Increasing Sample Size

Figure 1: Experimental Results for Computational Analysis



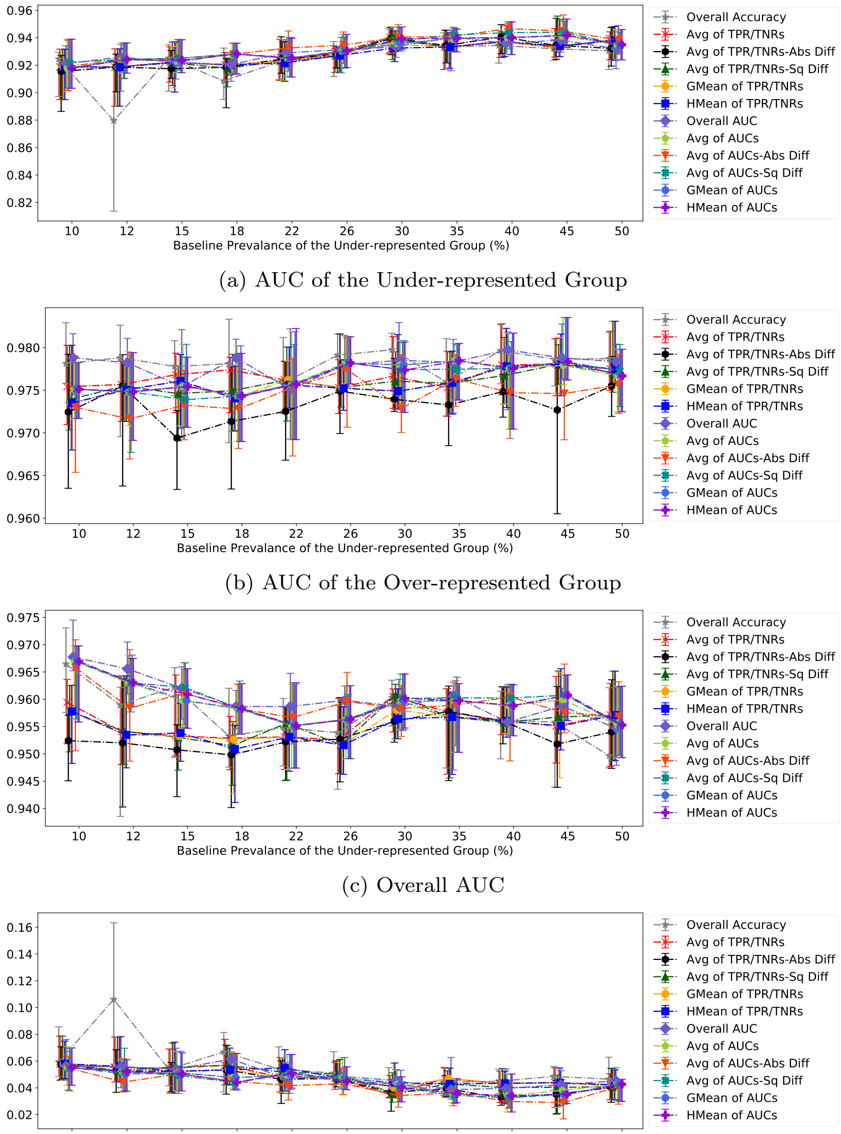
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 2: Experimental Results for Evaluation Metrics on Scenario 1



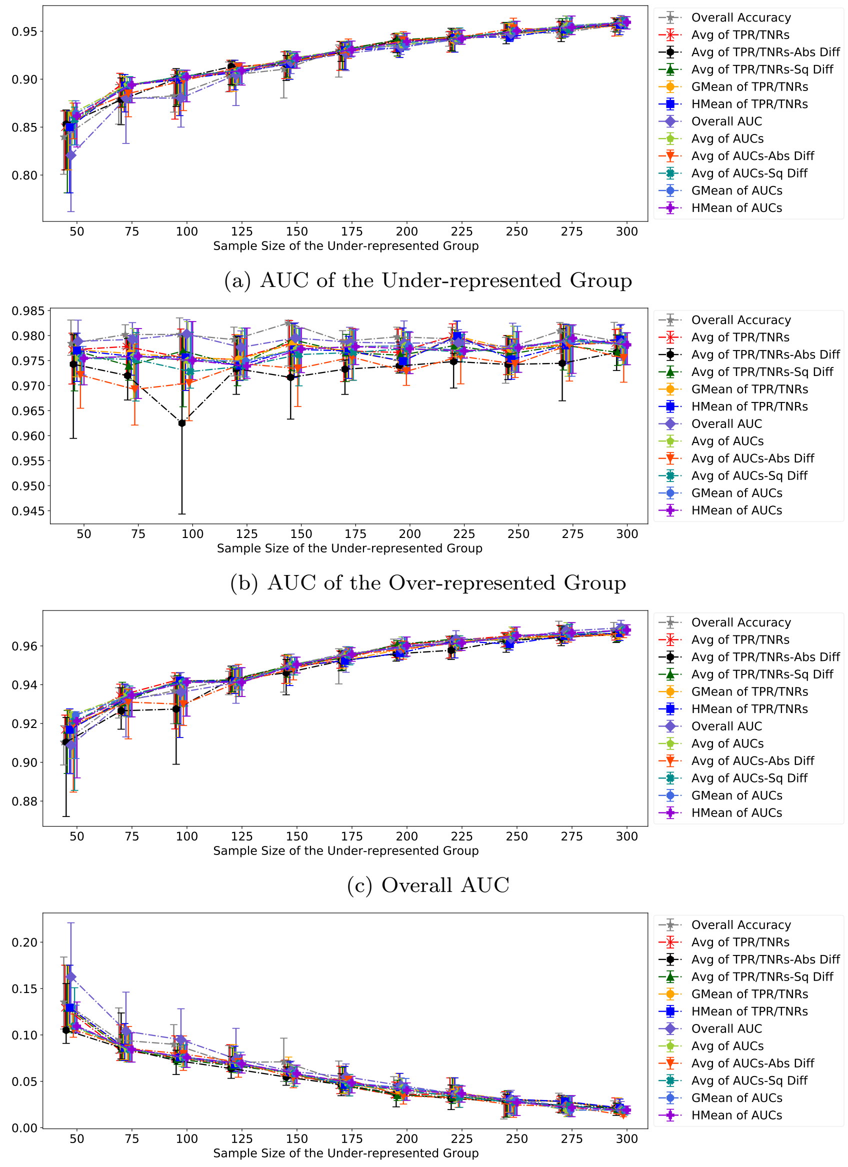
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 3: Experimental Results for Evaluation Metrics on Scenario 2



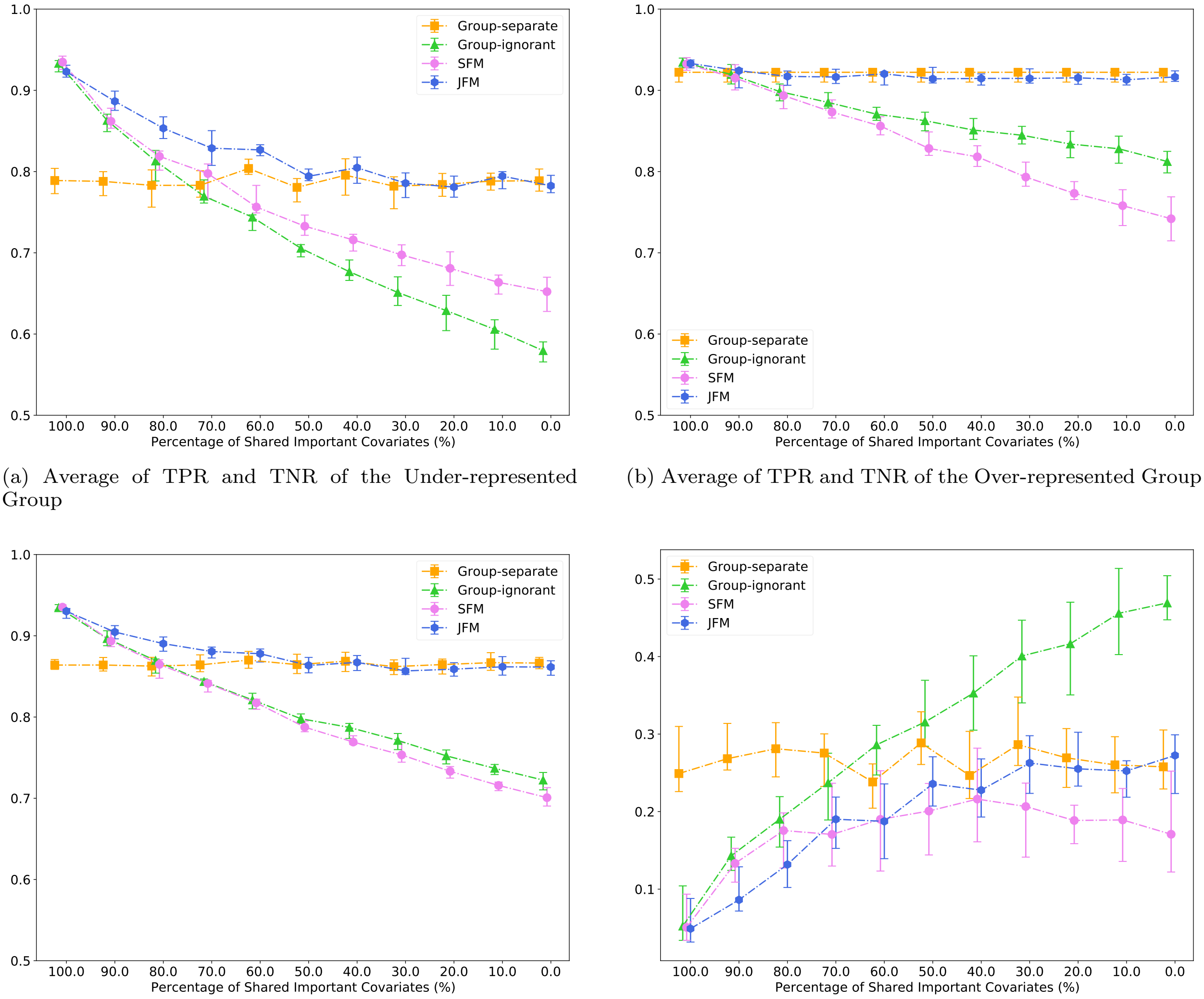
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 4: Experimental Results for Evaluation Metrics on Scenario 3



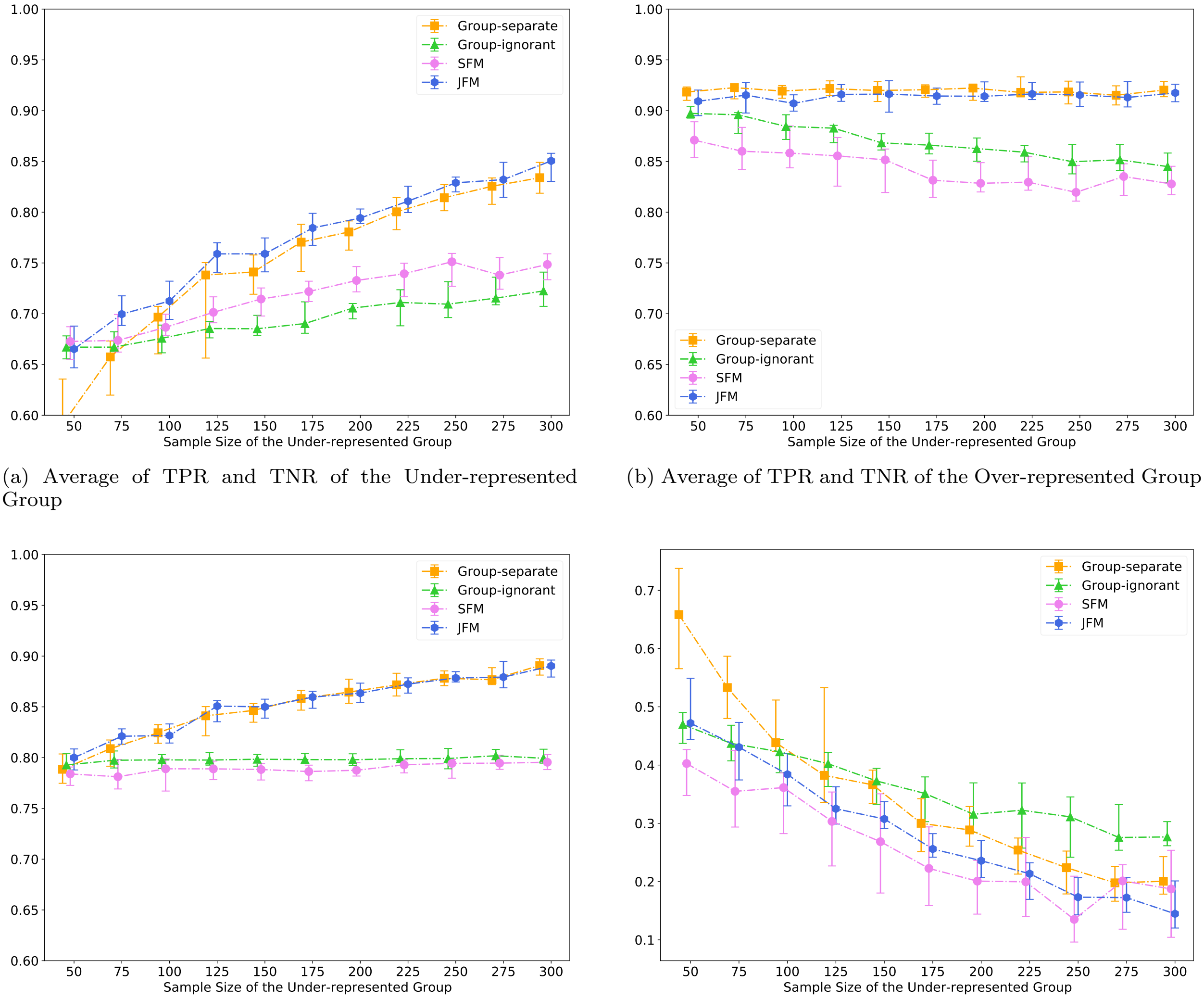
(a) Average of TPR and TNR of the Under-represented Group

(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 5: Experimental Results for Scenario 1 (TPR + TNR)



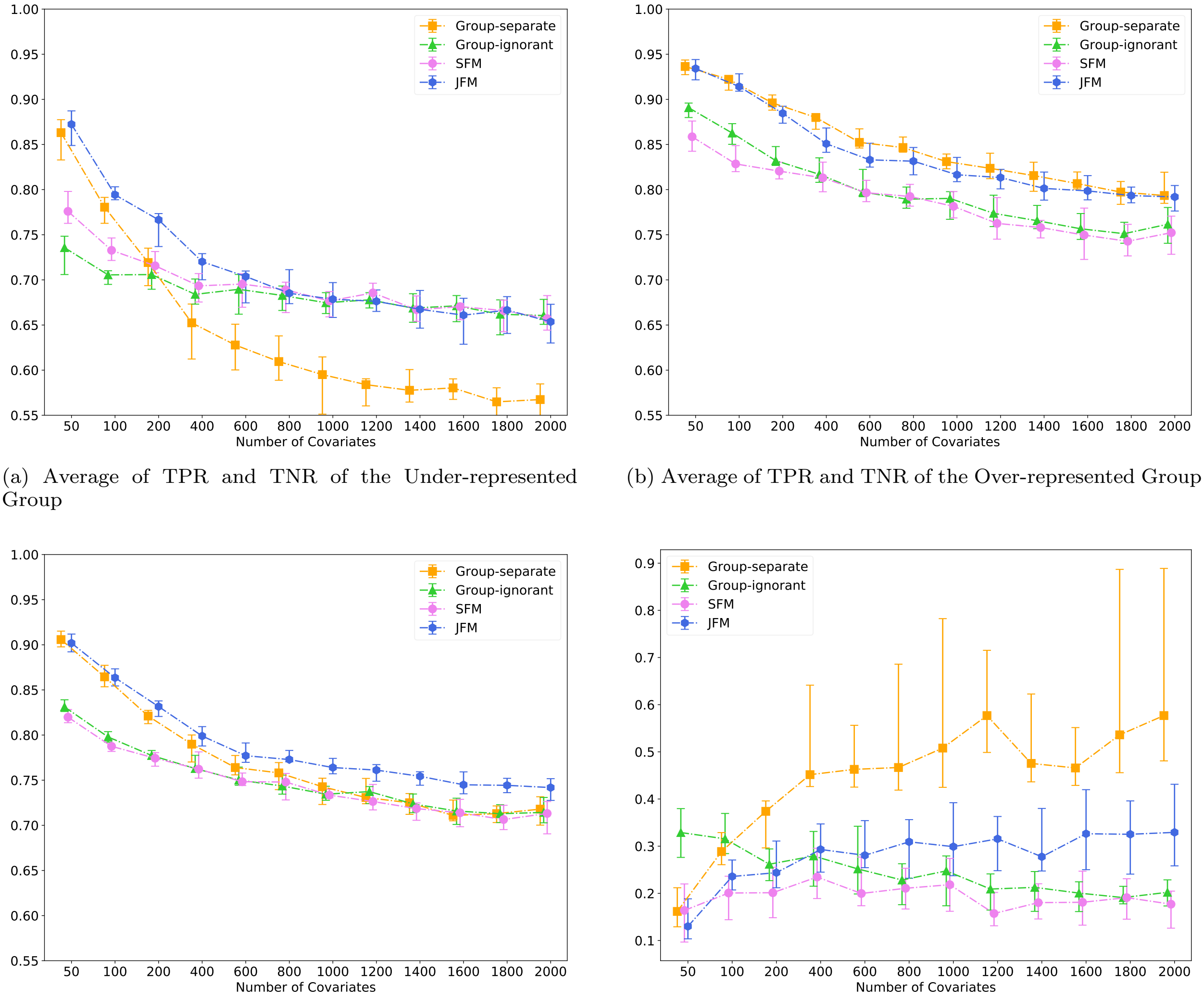
(a) Average of TPR and TNR of the Under-represented Group

(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 6: Experimental Results for Scenario 2 (TPR + TNR)



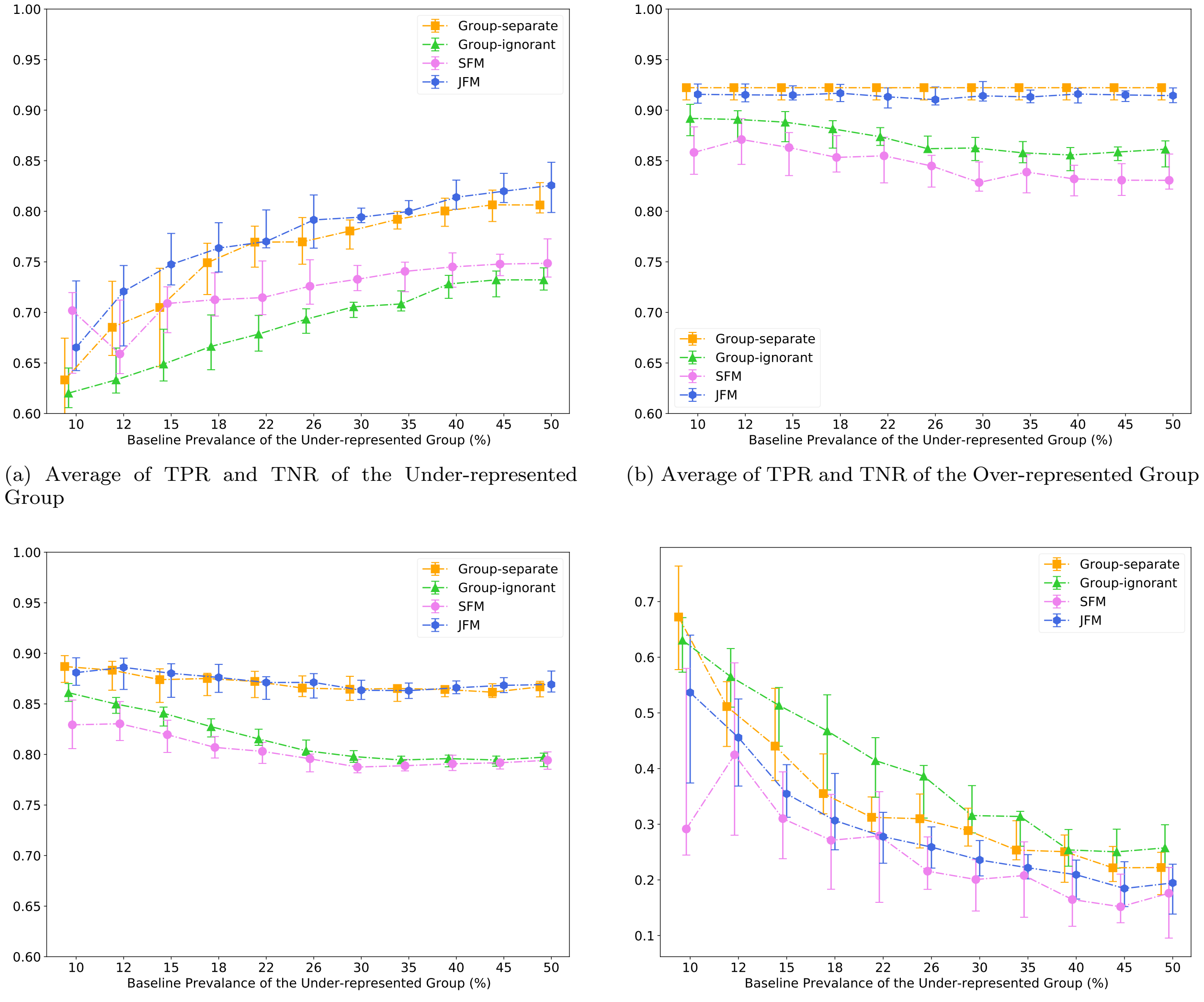
(a) Average of TPR and TNR of the Under-represented Group

(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 7: Experimental Results for Scenario 3 (TPR + TNR)



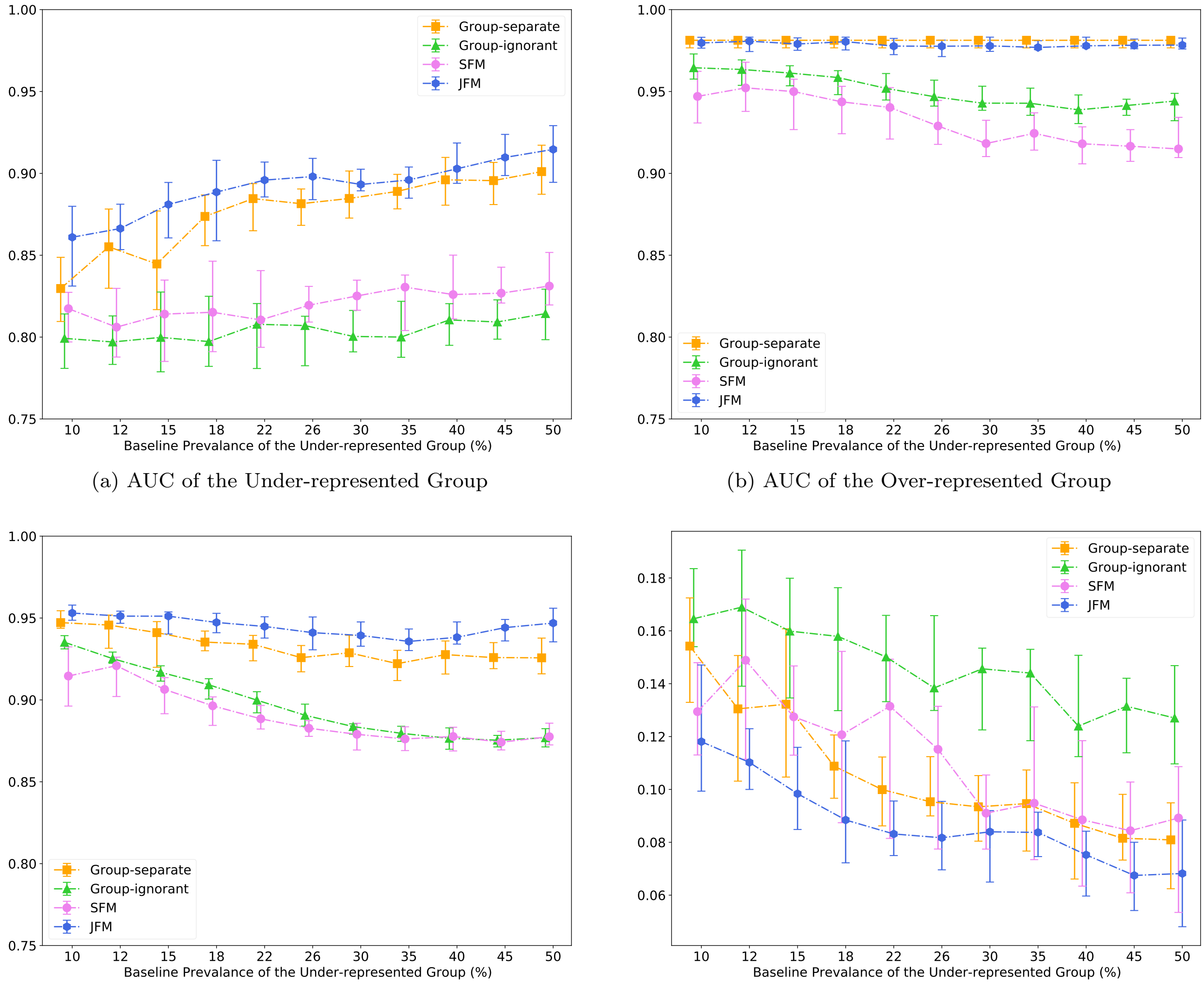
(a) Average of TPR and TNR of the Under-represented Group

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



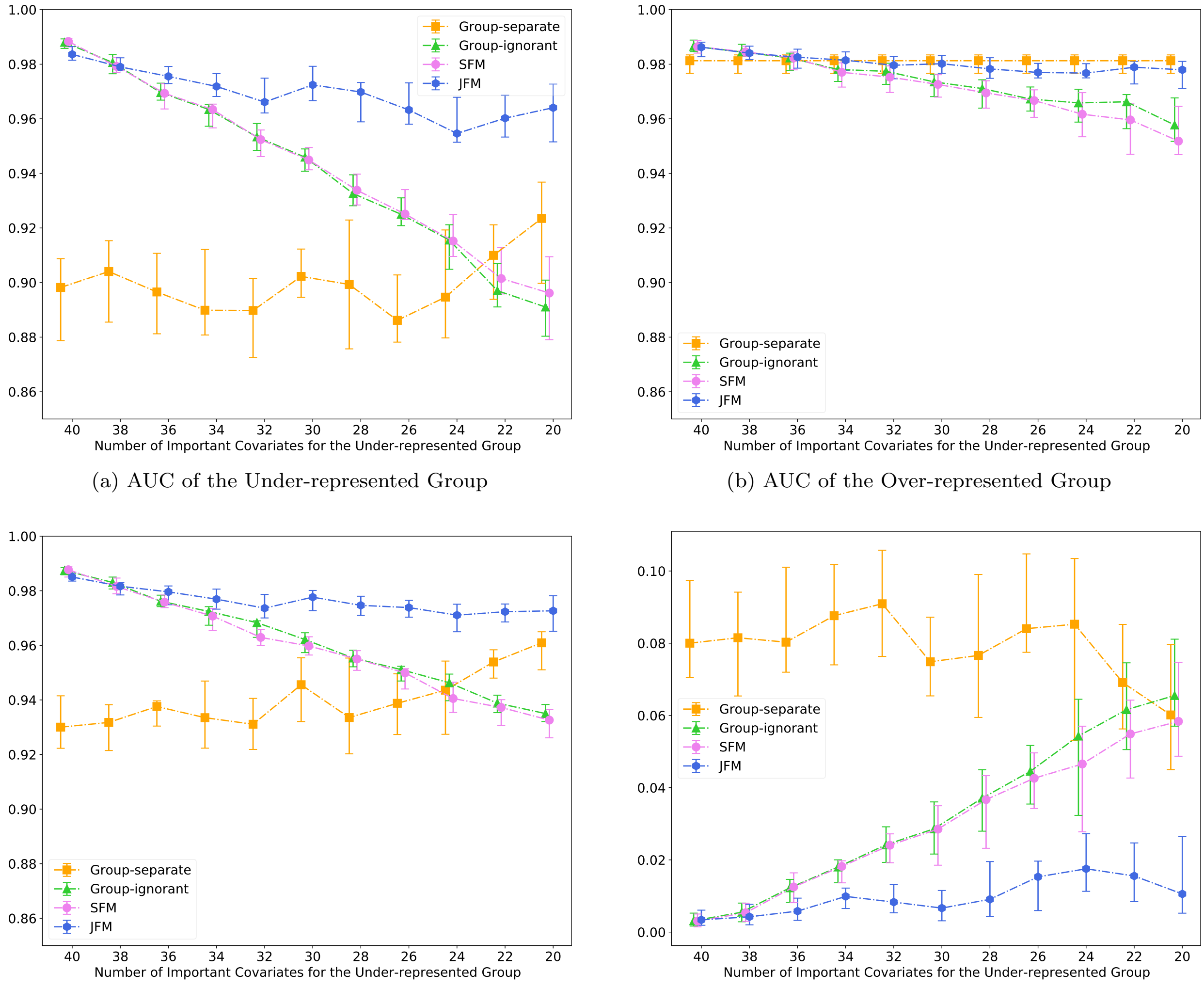
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 9: Experimental Results for Scenario 4



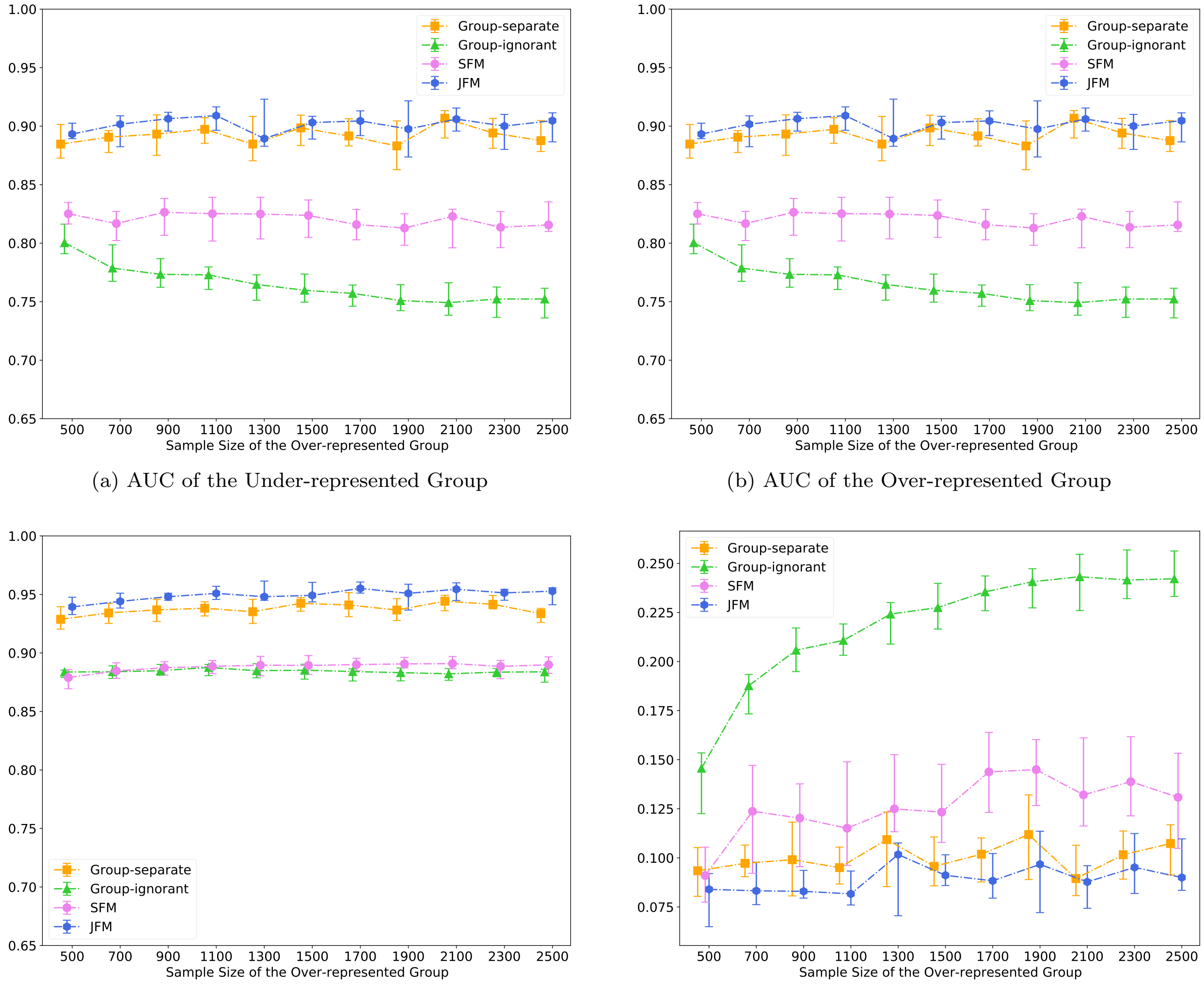
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 10: Experimental Results for Scenario 1B



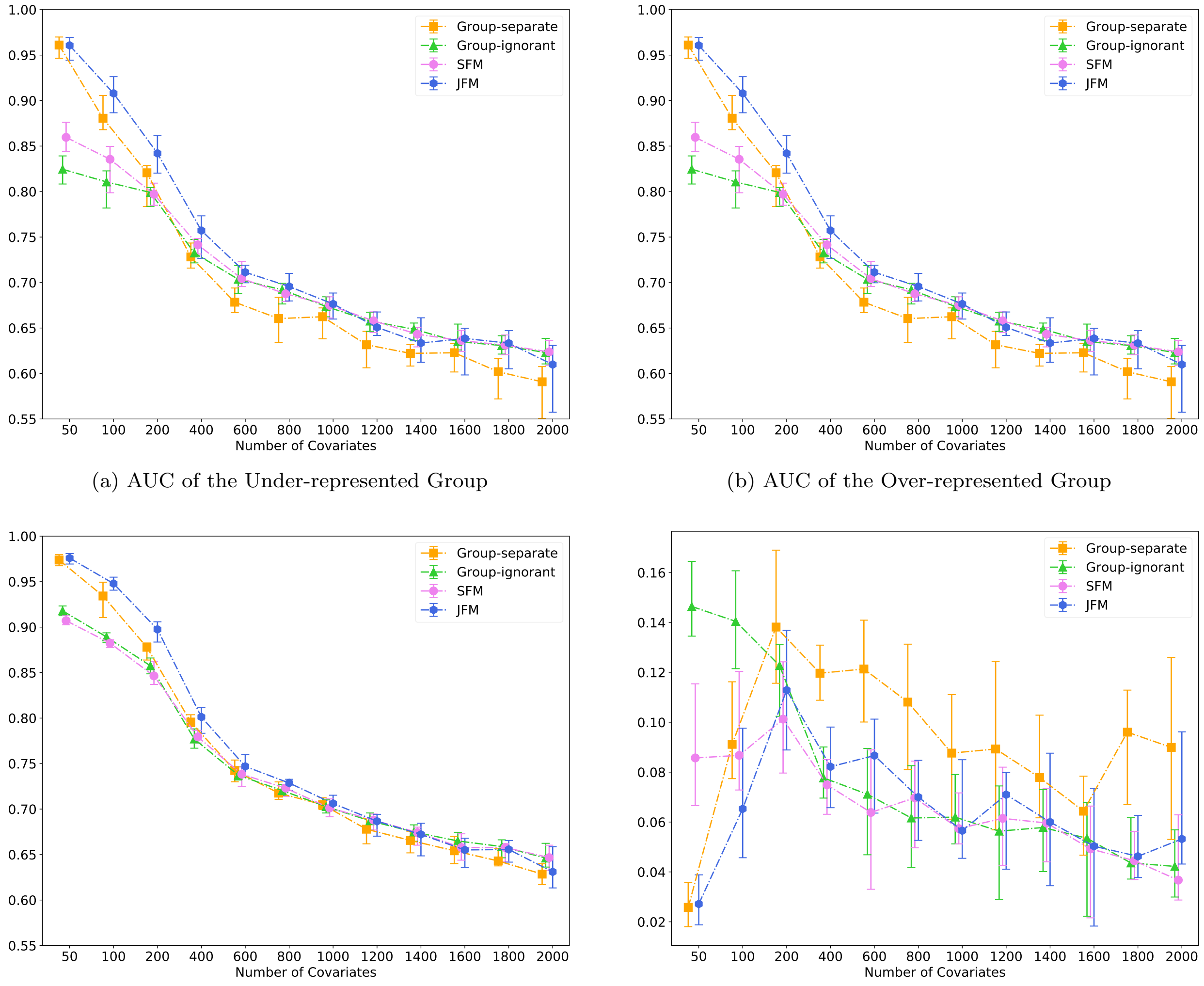
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 11: Experimental Results for Scenario 2B



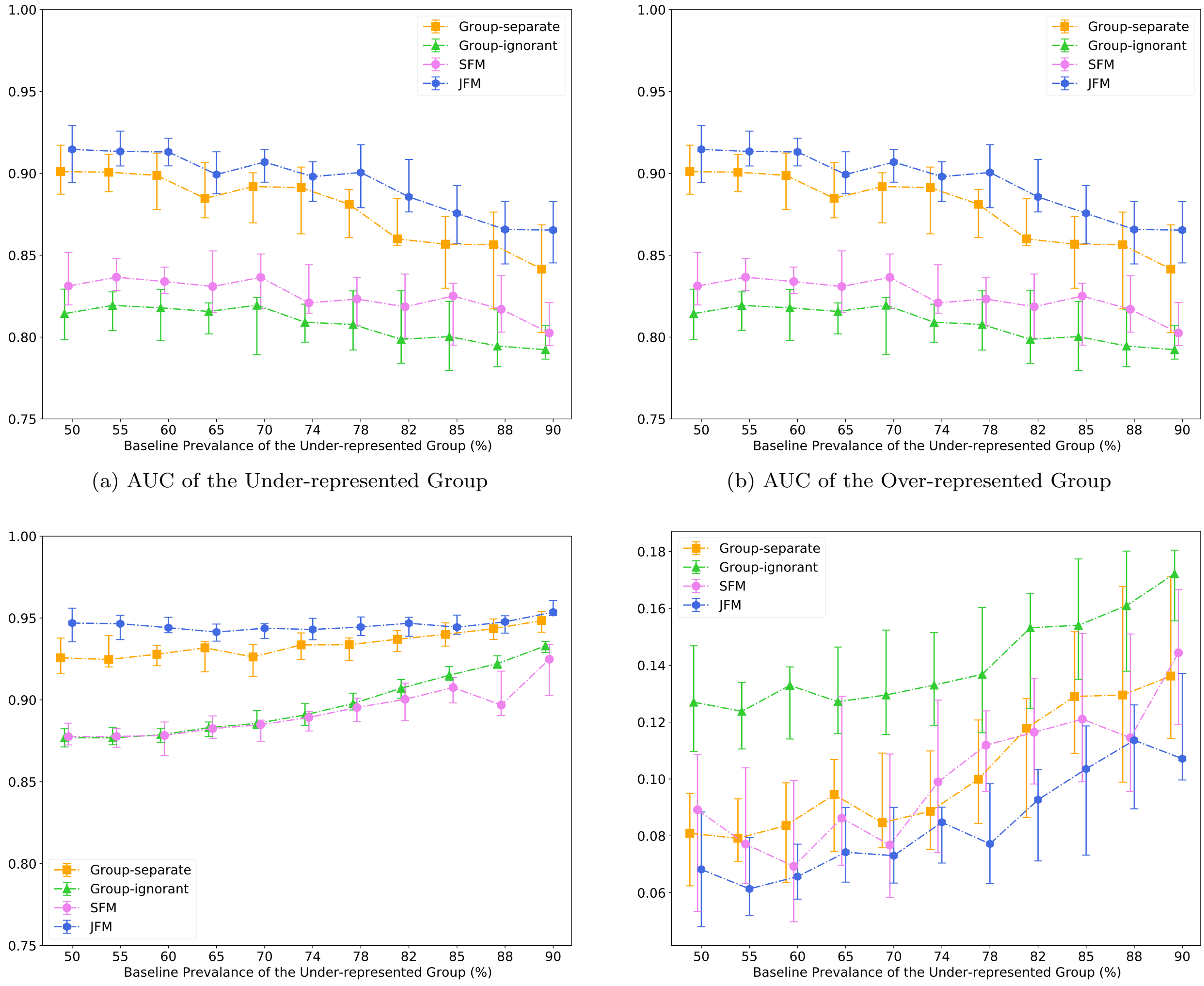
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 12: Experimental Results for Scenario 3B



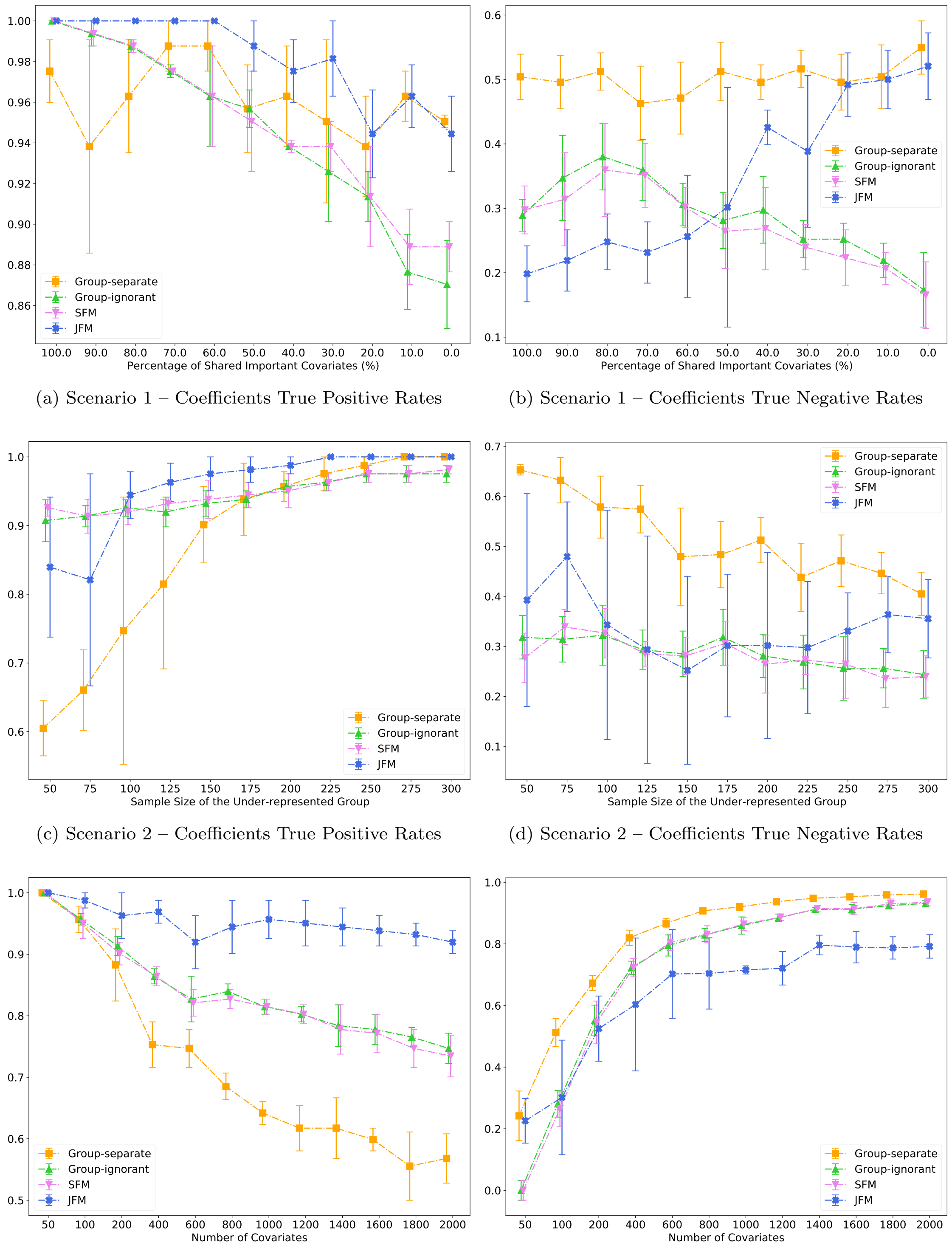
(a) AUC of the Under-represented Group

(b) AUC of the Over-represented Group

(c) Overall AUC

(d) Disparity of AUC

Figure 13: Experimental Results for Scenario 4B



(a) Scenario 1 – Coefficients True Positive Rates

(b) Scenario 1 – Coefficients True Negative Rates

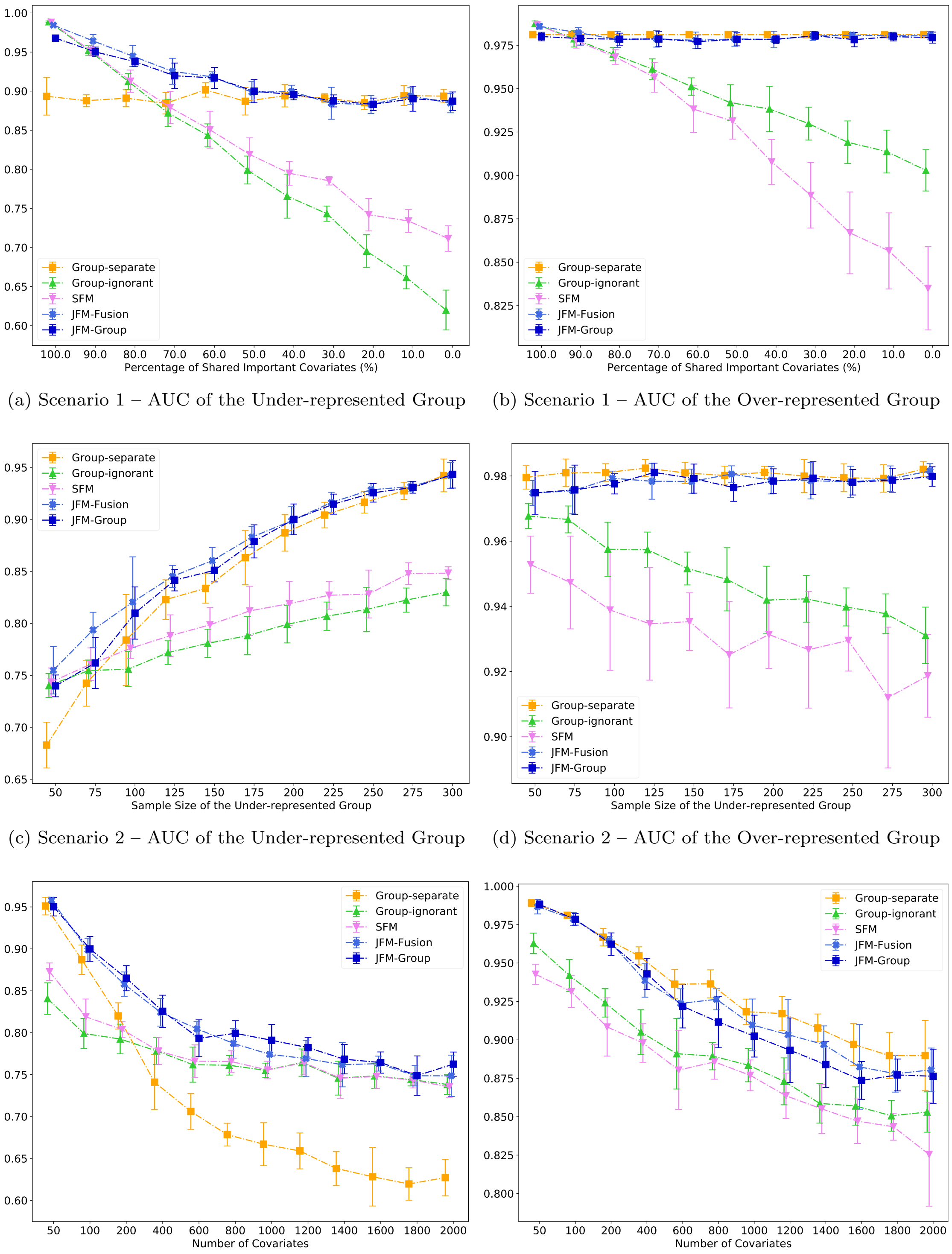
(c) Scenario 2 – Coefficients True Positive Rates

(d) Scenario 2 – Coefficients True Negative Rates

(e) Scenario 3 – Coefficients True Positive Rates

(f) Scenario 3 – Coefficients True Negative Rates

Figure 14: Variable Selection Performances



(a) Scenario 1 – AUC of the Under-represented Group

(b) Scenario 1 – AUC of the Over-represented Group

(c) Scenario 2 – AUC of the Under-represented Group

(d) Scenario 2 – AUC of the Over-represented Group

(e) Scenario 3 – AUC of the Under-represented Group

(f) Scenario 3 – AUC of the Over-represented Group

Figure 15: Experimental Results for JFM-Fusion / JFM-Group