Bayesian Modeling of Intersectional Fairness: The Variance of Bias*

James R. Foulds†‡ Rashidul Islam†‡ Kamrun Naher Keya†‡ Shimei Pan†

Abstract

Intersectionality is a framework that analyzes how interlocking systems of power and oppression affect individuals along overlapping dimensions including race, gender, sexual orientation, class, and disability. Intersectionality theory therefore implies it is important that fairness in artificial intelligence systems be protected with regard to multi-dimensional protected attributes. However, the measurement of fairness becomes statistically challenging in the multi-dimensional setting due to data sparsity, which increases rapidly in the number of dimensions, and in the values per dimension. We present a Bayesian probabilistic modeling approach for the reliable, data-efficient estimation of fairness with multi-dimensional protected attributes, which we apply to two existing intersectional fairness metrics. Experimental results on census data and the COMPAS criminal justice recidivism dataset demonstrate the utility of our methodology, and show that Bayesian methods are valuable for the modeling and measurement of fairness in intersectional contexts.

1 Introduction

With the rising influence of machine learning algorithms on many important aspects of our daily lives, there are growing concerns that biases inherent in data can lead the behavior of these algorithms to discriminate against certain populations [1, 3, 5, 6, 11, 22, 23]. In recent years, substantial research effort has been devoted to the development and enforcement of mathematical definitions of bias and fairness in machine learning algorithms [11, 14, 21, 17].

In this work, our guiding principle for fairness is *intersectionality*, the core theoretical framework underlying the third-wave feminist movement [10, 8]. Intersectionality theory states that racism, sexism, and other social systems which harm marginalized groups have interlocking effects, such that the lived experience of, e.g., Black women, is very different than that of, e.g., white women. We therefore focus on fairness scenarios where there are *multiple protected attributes*, such as gender, race, and sexual orientation.

Protected attributes	gender	gender, nationality	gender, nationality, race
Median # instances	14,719	5,195	172
Minimum # instances	9,216	963	5

Table 1: Number of instances at each intersection of the protected attributes' values, UCI Adult census dataset.

While fairness methods have been extended to multiple protected attributes [17, 15, 13], data sparsity rapidly becomes an issue as the number of dimensions (and their number of distinct values) increases due to the curse of dimensionality, leading to uncertainty in the measurement of fairness. For example, Table 1 shows how the number of instances per value at the intersections of the protected attributes, and especially the minimum of these counts, decreases as more protected attributes are introduced, on the UCI Adult census dataset [20]. It may be difficult, e.g., to estimate the overall behavior of a classifier on North American indigenous women who are immigrants to the USA, due to a lack of recorded data on such individuals. Minority groups will have relatively little observed data if observations are i.i.d., and may be even further underrepresented if the data collection process is biased toward non-minorities. To detect intersectional discrimination, we need to measure the system's behavior on potentially small intersectional groups, which is unreliable to estimate in the resulting "small N" regime [25].

The goal of this work, therefore, is to address the challenge of *reliably modeling and measuring fairness in an intersectional context, despite data sparsity.* While small data uncertainty [2, 25], intersectionality [7, 13], and multiple attribute definitions [17, 15, 13] have been studied, we are first to consider them concurrently.

*This work was performed under the following financial assistance award: 60NANB18D227 from U.S. Department of Commerce, National Institute of Standards and Technology. This material is based upon work supported by the National Science Foundation under Grant No. IIS 1850023. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

†Department of Information Systems, UMBC, Baltimore, USA

‡Equal Contribution

The majority of the research on fairness in AI to date has focused on the development of *learning algorithms which enforce fairness metrics* [11, 28, 14, 6, 21, 4]. In contrast, **we instead focus on accurately *measuring* the unfairness of a system or dataset**. Fairness measurement is crucial when engineering AI systems for deployment [26]. It is essential for determining whether disparities in system behavior meet legal thresholds for discrimination [25]. And it is integral to investigative reporting on disparate behavior of existing AI systems, which promotes awareness and can ultimately lead to the rectification of algorithm injustice [1, 7, 24]. We propose practical techniques to mitigate *variance* (i.e. uncertainty) in the estimation of algorithmic *bias* in datasets with multiple protected attributes.[1] Our primary contributions include:

1. We propose a *Bayesian probabilistic modeling framework* for reliably estimating fairness and its uncertainty in the data-sparse intersectional regime.

2. We instantiate our proposed framework with a *novel hierarchical extension of Bayesian logistic regression* which is potentially an appropriate choice for this setting, and three other statistical models, each with a different bias and variance trade-off. We further propose a Bayesian model averaging approach which leverages all of the models together.

3. We study the behavior of our Bayesian models on criminal justice, census, and synthetic data. Our results *demonstrate the importance of the Bayesian modeling approach in an intersectional context.*

The paper is structured as follows. We begin by discussing intersectionality theory, which motivates our multi-dimensional approach to fairness, and describe two intersectional fairness metrics from the literature [13, 17]. Next, we propose Bayesian probabilistic models for estimating these (and other) fairness metrics in the multi-dimensional fairness regime. We then empirically study the behavior of the models in estimating the intersectional fairness metrics, and showcase their real-world application with a case study on the COMPAS recidivism dataset. Finally, we conclude with a discussion of the practical implications of our work.

2 Background and Motivation: AI Fairness and Intersectionality

Intersectionality is a critical lens for analyzing how unfair processes in society, such as sexism and systemic racism, affect certain groups. The term was original introduced by [10], who studied how the combined harms of such *systems of oppression* affect Black women, who are simultaneously affected by sexism, racism, and other related disadvantages [27, 9]. In its more general form, advanced by [8] and others, intersectionality theory posits that *individuals at the intersection of multiple protected categories*, along lines of gender, race, social class, disability, and so on, are harmed by overlapping systems of oppression.

In an AI fairness context, this implies that fairness should be enforced at the intersections of multiple protected attributes [7, 13]. Here, we consider several existing fairness definitions which are appropriate in an intersectional context.

2.1 Differential Fairness Differential fairness [13] is a definition specifically motivated by intersectionality, which aims to ensure equitable treatment by an algorithm for *all intersecting subgroups* of a set of protected categories. We use the notation of [19] for all definitions we consider. Let $M(x)$ be an algorithmic mechanism which takes an individual's data x and assigns them an outcome y, e.g. whether or not the individual was awarded a loan. Let $S_1, \ldots, S_p$ be discrete-valued protected attributes, $A = S_1 \times S_2 \times \ldots \times S_p$, and Θ be the set of plausible distributions θ that may generate x (typically assumed to be a single distribution, $\Theta = \{\theta\}$).

Definition 2.1. *(Differential Fairness) A mechanism $M(x)$ is ϵ-*differentially fair (DF) *with respect to (A, Θ) if for all $\theta \in \Theta$ with $x \sim \theta$, and $y \in Range(M)$,*

$$e^{-\epsilon} \leq \frac{P_{M,\theta}(M(x) = y|s_i, \theta)}{P_{M,\theta}(M(x) = y|s_j, \theta)} \leq e^{\epsilon} \ , \tag{2.1}$$

for all $(s_i, s_j) \in A \times A$ where $P(s_i|\theta) > 0$, $P(s_j|\theta) > 0$.

In Definition 2.1, s_i, $s_j \in A$ are tuples of *all* protected attribute values, e.g. gender, race, and nationality. If all of the $P_{M,\theta}(M(x) = y|s, \theta)$ probabilities are equal for each group s, across all outcomes y and distributions θ, $\epsilon = 0$, otherwise $\epsilon > 0$. [13] proved that this definition guarantees fairness protections for all subsets of the protected attributes, e.g. if all intersections of gender and race are protected (e.g. Black women), then gender (e.g. women) and race (e.g. white people) are separately protected, a property which is consistent with the ethical principles of intersectionality theory. [13] proved privacy and economic guarantees which determine the fairness consequences of a particular ϵ. They further proposed a variant definition which only considers the increase in unfairness by the algorithm, over the unfairness in the original data.

[1] Algorithmic bias is not to be confused with statistical bias, not withstanding the pun in the title of this paper.

Definition 2.2. *(DF Bias Amplification) A mechanism $M(x)$ satisfies* $(\epsilon_2 - \epsilon_1)$-DF bias amplification *with respect to* $(A, \Theta, D, \mathcal{M})$ *if it is* ϵ_2*-DF and D is a labeled dataset which is* ϵ_1*-DF w.r.t. a model $\mathcal{M}$ which was trained on D to estimate $P(y|s)$ in the data.*

2.2 Subgroup Fairness [17] proposed multi-attribute fairness definitions which aim to prevent *fairness gerrymandering* at the intersections of protected groups, as later empirically validated by [18].

Definition 2.3. *(Statistical Parity Subgroup Fairness) Let $\mathcal{G}$ be a collection of protected group indicators $g : A \to \{0, 1\}$, where $g(s) = 1$ designates that an individual with protected attributes s is in group g. Assume that the mechanism $M(x)$ is binary, i.e. $y \in \{0, 1\}$.*

Then $M(x)$ is γ-statistical parity subgroup fair *(SF) with respect to θ and $\mathcal{G}$ if for every $g \in \mathcal{G}$,*

$$(2.2)\quad \begin{aligned} &|P_{M,\theta}(M(x) = 1|\theta) - P_{M,\theta}(M(x) = 1)|g(x) = 1, \theta)| \\ &\times P_\theta(g(x) = 1|\theta) \le \gamma \ . \end{aligned}$$

Since we are interested in fairness applications where intersectional ethics are to be upheld, we focus on the case where, similarly to DF, $\mathcal{G}$ contains all possible assignments of the protected attributes s (presumed to be enumerable). [17] and [15] proposed further related multi-attribute definitions regarding false positive rates and calibration, respectively. Our methods can also be applied to these definitions, but it is beyond our scope.

2.3 Empirical Fairness Estimation The central challenge for measuring fairness in an intersectional context, either via ϵ-DF, γ-SF, or related notions, is to estimate $M(x)$'s marginal behavior $P_{M,\theta}(y|s, \theta)$ for each (y, s) pair, with potentially little data for each of these. The simplest method to do this is to use the empirical data distribution. E.g., for the ϵ-DF criterion, assuming discrete outcomes and protected attributes, $P_{Data}(y|s) = \frac{N_{y,s}}{N_s}$, where $N_{y,s}$ and N_s are empirical counts of their subscripted values in the dataset. **Empirical differential fairness (EDF)** [13] corresponds to verifying that for any y, s_i, s_j,

$$(2.3)\qquad e^{-\epsilon} \le \frac{N_{y,s_i}}{N_{s_i}} \frac{N_{s_j}}{N_{y,s_j}} \le e^{\epsilon} \ ,$$

whenever $N_{s_i} > 0$ and $N_{s_j} > 0$. However, in the intersectional setting, the counts $N_{y,s}$ at the intersection of the values of the protected attributes become rapidly smaller as the dimensionality and cardinality of protected attributes increase (cf. Table 1). In this case, the conditional probabilities in Equations 2.1 and 2.2, and hence the fairness metrics, will generally have high uncertainty (or variance, from a frequentist perspective) [25]. The $N_{y,s}$ counts may even be 0, which can make the estimate of ϵ in Equation 2.3 infinite/undefined.[2]

3 Model-Based Fairness Estimation

Instead of using empirical probabilities, in this paper we propose to generalize beyond the training set by learning $P_{M,\theta}(y|s, \theta)$ via a *probabilistic model*. This approach has several advantages. First, by exploiting structure in the distributions, e.g. if the mechanism's behavior on *women* is informative of its behavior on *Black women*, we can accurately model all of the conditional probabilities with fewer parameters than empirical frequencies, thereby reducing variance in estimation. Second, we can use a Bayesian approach to manage uncertainty in the estimation, and to report this uncertainty to an analyst.

A simple baseline, proposed by [13] to address the zero count issue, is to put a Dirichlet prior on the probabilities in Equation 2.3. Estimating ϵ*-DF* via the posterior predictive distribution of the resulting Dirichlet-multinomial, the criterion for any y, s_i, s_j is

$$(3.4)\qquad e^{-\epsilon} \le \frac{N_{y,s_i} + \alpha}{N_{s_i} + |\mathcal{Y}|\alpha} \frac{N_{s_j} + |\mathcal{Y}|\alpha}{N_{y,s_j} + \alpha} \le e^{\epsilon} \ ,$$

where scalar α is each entry of the parameter of a symmetric Dirichlet prior with concentration parameter $|\mathcal{Y}|\alpha$, $\mathcal{Y} = \text{Range}(M)$. [13] refer to this as **smoothed EDF**. This can also be used for γ-SF.

More generally, in this work we propose to estimate $P_{M,\theta}(y|s, \theta)$, and hence the fairness metrics, via **a probabilistic classifier** which predicts the outcome y given protected attribute values $s \in A$, trained on $\mathcal{D}_s$. The complexity of the model determines the trade-off between (statistical) bias and variance in the estimation.[3] For instance, ordered from high statistical bias to high variance, we could consider *naive Bayes*, *logistic regression*, or *deep neural networks*.

For most typical models and datasets, to manage uncertainty in the data-sparse intersectional regime, we recommend that the probabilistic classifier be trained via **fully Bayesian inference**. Fully accounting for

[2]Note that [17] prove large-sample generalization guarantees for empirical estimates of γ-SF. As we shall see, this does not imply that empirical estimates of γ will be accurate for small-to-moderately sized datasets. Nevertheless, since SF downweights small groups (the second term of Equation 2.2) and uses an additive formulation of fairness (compared to DF's multiplicative formulation), it is expected that empirical estimates will be somewhat more stable for SF than DF.

[3]Here, *statistical bias* is not to be confused with unfairness.

Algorithm 1 Bayesian estimation of differential fairness and its uncertainty (and similarly for γ-SF).

Input: Development set $\mathcal{D} = \{(x_i, y_i)\}$, mechanism $M(x)$, protected attributes A

Output: $\hat{\epsilon}_{data}$, $\hat{\epsilon}_{M(x)}$, boxplots of posterior uncertainty in ϵ_{data}, $\epsilon_{M(x)}$, $\epsilon_{M(x)} - \epsilon_{data}$

- Apply $M(x)$ to $x_i \in \mathcal{D}$, obtain mechanism labels y'_i
- Fit Bayesian classifier $p_1(y|s, \bar{\theta}_1)$ on $\mathcal{D}_s = \{(s_i, y_i)\}$
- Fit Bayesian classifier $p_2(y'|s, \bar{\theta}_2)$ on $\mathcal{D}'_s = \{(s_i, y'_i)\}$
- Estimate $\hat{\epsilon}_{data}$ via Eqn.2.1 with posterior predictive $p_1(y|s)$
- Estimate $\hat{\epsilon}_{M(x)}$ via Eqn.2.1 with posterior predictive $p_2(y'|s)$
- Plot posterior uncertainty in ϵ_{data}, $\epsilon_{M(x)}$, $\epsilon_{M(x)} - \epsilon_{data}$

parameter uncertainty, a single best estimate of the conditional distributions $\hat{\theta}$ to compute ϵ or γ is the posterior predictive distribution, $\hat{\theta} = P_{Model}(y|s, \mathcal{D}_s) = \int_{\bar{\theta}} P_{Model}(y|s, \bar{\theta}) P_{Model}(\bar{\theta}|\mathcal{D}_s)$, for model parameters $\bar{\theta}$. This can be approximated by, e.g., averaging $P_{Model}(y|s, \bar{\theta})$ over MCMC samples of $\bar{\theta}$ or a variational posterior. We then report uncertainty in ϵ by plotting the posterior distribution over ϵ based on posterior samples of $\bar{\theta}$, and similarly for γ-SF. Our overall approach to the Bayesian modeling of intersectional fairness metrics is shown in pseudocode in Algorithm 1.

3.1 Hierarchical Logistic Regression As a compromise between statistical bias and variance in this setting, we propose a novel hierarchical extension of logistic regression (*HLR*), where the "prior" on $\gamma = \text{logit}(P(y = 1|s))$ is a Gaussian around the prediction of a jointly trained logistic regression, allowing deviations justified by sufficient data. Let $\vec{\mathbf{s}}_j$ be an encoding of protected attribute values s_j with a binary indicator for each attribute's value, with integer j indexing each possible value of s, and β_i be a regression coefficient for each entry of the $\vec{\mathbf{s}}_j$'s. The model's generative process is:

- $\sigma_2 \sim \text{Exponential}(\lambda)$
- $\beta_i \sim \text{Normal}(\mu, \sigma_1), \quad c \sim \text{Normal}(\mu, \sigma_1)$
- $\gamma_j \sim \text{Normal}(\beta^\intercal \vec{\mathbf{s}}_j + c, \sigma_2)$
- $P(y = 1|s_j) = \sigma(\gamma_j))$,

where λ and σ_1 are prior hyperparameters, and σ_2 encodes the extent of deviations from logistic regression.

4 Bayesian Model Averaging Ensemble

A potential concern with the above approach is that different probabilistic models will lead to different estimates in the measurement of ϵ-DF and γ-SF. Consistently with our Bayesian methodology, rather than performing model selection we can account for uncertainty over models by combining them using Bayesian model averaging [16]. Suppose there are K candidate models. We estimate the posterior distribution of ϵ (similarly γ) in the ensemble given dataset $\mathcal{D}$ via:

$$(4.5) \qquad P(\epsilon|\mathcal{D}) = \sum_{k=1}^{K} P(\epsilon|M_k, \mathcal{D}) P(M_k|\mathcal{D}) \ .$$

Assuming a uniform prior over models, $P(M_k|\mathcal{D}) \propto \prod_{(y,s)\in\mathcal{D}} P(y|s, M_k)$, the conditional marginal likelihood. The distribution $P(\epsilon|M_k, \mathcal{D})$ is estimated via MCMC or variational inference over the posterior over the model parameters $P(\bar{\theta}_k|M_k, \mathcal{D})$, with each $\bar{\theta}_k$ corresponding to an ϵ (or γ). Finally, we obtain a gold-standard estimate $\hat{\epsilon}$ or $\hat{\gamma}$ by simulating from the ensemble to estimate the posterior predictive distributions $p(y|s, \mathcal{D})$, and plugging these into Equations 2.1 or 2.2.

5 Experimental Results

The goals of our experiments were to compare our proposed Bayesian modeling approach for estimating intersectional fairness to point estimation and to empirical measurement, to evaluate the performance of different models and of model averaging, to study the effect of uncertainty/variance in intersectional fairness estimation, and to illustrate the practical application of our methods. We performed all experiments on two datasets:

- The Adult 1994 U.S. census income data from the UCI repository [20]. This dataset consists of 14 attributes regarding work, relationships, and demographics for 48,842 individuals, who are labeled according to whether their income exceeds \$50,000 per year, We select *race*, *gender*, and *nationality* as the protected attributes. As most instances have *U.S. nationality*, we treat *nationality* as binary between *U.S.* and *non-U.S.* (We also consider the case where all 40 categories in the *nationality* attribute are used in Table 2.) *Gender* is also coded as binary. The *race* attribute had 5 values. For Adult, we set $M(x)$ to be a logistic regression model, since it has an appropriate level of model complexity for this data regime. We trained the model on half of the training set (which was held out from the $P_{M,\theta}(y|s, \theta)$ models).
- The COMPAS dataset regarding a system that is used to predict criminal recidivism, and which has been criticized as potentially biased [1]. We used *race* and *gender* as protected attributes. *Gender* was coded as binary while *race* had 6 categories.

	Negative Cross-entropy				Total Variation Distance			
Adult Dataset (Nationality Attribute Binarized to {U.S., non-U.S.})								
Models	Actual-labeled test set (full training set)		$M(x)$-relabeled test set (held-out training subset)		Actual-labeled test set (full training set)		$M(x)$-relabeled test set (held-out training subset)	
	PE	FB	PE	FB	PE	FB	PE	FB
EDF	-0.4201	-0.4139	-0.3715	**-0.3687**	0.0393	0.0380	0.0605	0.0576
NB	-0.4254	-0.4237	-0.3870	-0.3801	0.0618	0.0611	0.0802	0.0792
LR	-0.4411	-0.4148	-0.4047	-0.3752	0.0823	0.0389	0.1024	0.0586
DNN	-0.4179	-0.4128	-0.3794	-0.3753	0.0547	0.0405	0.0722	0.0629
HLR	X	**-0.4106**	X	-0.3713	X	**0.0328**	X	**0.0544**
Ensemble	-0.4109		-0.3709		0.0329		0.0556	
Adult Dataset (Nationality Attribute not Binarized)								
Models	Actual-labeled test set (full training set)		$M(x)$-relabeled test set (held-out training subset)		Actual-labeled test set (full training set)		$M(x)$-relabeled test set (held-out training subset)	
	PE	FB	PE	FB	PE	FB	PE	FB
EDF	-0.6416	-0.5754	-0.6282	-0.5095	0.2355	0.1222	0.2153	0.1302
NB	-0.8470	-0.6831	-0.8244	-0.6937	0.6274	0.2565	0.6161	0.3134
LR	-0.8175	-0.5838	-0.7957	-0.5175	0.6020	0.1207	0.5814	0.1084
DNN	-0.7624	-0.5759	-0.7262	-0.5092	0.5510	**0.1181**	0.5256	0.1078
HLR	X	**-0.5753**	X	**-0.5090**	X	0.1188	X	**0.1077**
Ensemble	-0.5778		-0.5093		0.1195		0.1103	

Table 2: Comparison of predictive performance of intersectional fairness models with respect to average negative cross-entropy (higher is better) and total variation distance (lower is better) per intersection on the test set, on Adult (top) and on Adult without binarizing the "*nationality*" attribute (bottom), a high data sparsity scenario. Here, PE = point estimate, FB = fully Bayesian estimate using the posterior predictive distribution. EDF-FB is the Dirichlet-multinomial model, cf. Equation 3.4. The best performing method is indicated in bold. Results on COMPAS are given in the Supplementary Materials.

We used "*actual recidivism*" (within a 2-year period) for $7,214$ individuals, which is binary, as the true label of the data generating process and the COMPAS system's prediction as the labels from $M(x)$. Although COMPAS is a black box, we observe its assigned class labels y', and our models extrapolate its behavior on intersectional groups. Following [1], we merged the "*medium*" and "*high*" labels to make COMPAS scores binary, since the actual labels are binary.

We randomly split 70% and 30% of the dataset as train and test sets using stratified sampling on the intersectional groups. All models were trained using PyMC3, with ADVI used for Bayesian inference. Posterior predictive distributions were estimated by sampling from the variational posterior and averaging the predictions. Important observations are indicated in bold.

5.1 Prediction Performance on Held-Out Data

We first studied the predictive performance for models of $P_{M,\theta}(y|s,\theta)$, which must perform well in order to accurately estimate ϵ-DF and γ-SF: the empirical distribution (EDF), naive Bayes (NB), logistic regression (LR), deep neural networks (DNN), and our hierarchical logistic regression (HLR). For each model, we compare point estimates (PE) (MAP, except for EDF), and fully Bayesian inference via the posterior predictive distribution (FB), as well as a Bayesian model averaging ensemble (Ensemble) of all PE and FB models. Note that the configuration of the DNN architecture is 3 hidden layers, 10 neurons in each layer, "relu" and "sigmoid" activations for the hidden and output layers, respectively. The Dirichlet multinomial model (cf. Eqn. 3.4) is denoted EDF-FB. We trained all models on the training set and reported *negative cross-entropy* and *total variation distance* from the test set's empirical $P(y|s)$ and $P(y'|s)$, averaged over intersections s. Negative cross-entropy is closely related to log-likelihood, and here measures the similarity of the model's conditional distributions to those of the test set. Total variation distance is related to L1 distance, and is also calculated between the empirical $P(y|s)$ and model $P(y'|s)$ distributions.

Results on Adult are shown in Table 2 (COMPAS results are similar, shown in the Supplement). We report results both for y labels in the test data (inequity in society), and for an algorithmic mechanism

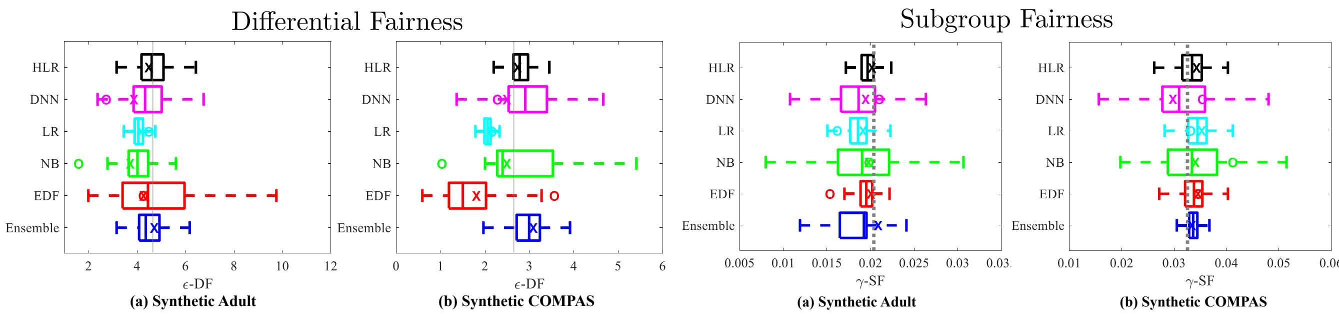


Figure 1: Fairness estimates using variational posteriors, point estimates, and posterior predictive distributions: semi-synthetic versions of (a) Adult and (b) COMPAS datasets, for differential fairness (left) and subgroup fairness (right). The "O" and "X" on top of the box-plots indicate estimates from PE and the posterior predictive distribution of FB models, respectively. The dotted vertical line represents ground truth ϵ or γ. The data were generated using per-group Gaussians and a threshold decision boundary.

$y' = M(x)$. In all cases, we found that **the empirical estimate EDF-PE was outperformed by probabilistic models** in terms of prediction, and that **fully Bayesian inference FB outperformed the corresponding point estimates PE**. Our Bayesian hierarchical logistic regression (**HLR-FB**) method was the best predictor in around 80% of the experimental conditions (dataset × labels vs $M(x)$ × evaluation metric). It also exhibited the **most reliable behavior across data sets**, being the only (non-ensemble) method to outperform EDF-PE in every case. The point estimate (PE) version of HLR performed too poorly to be shown. This may be due to numerical instability in PyMC3. The Bayesian model average (**Ensemble**) was also relatively **stable across datasets**, but HLR outperformed it in all cases.

To study the predictive performance of the models in the very sparse regime, we repeated the experiment for Adult in the case where all values of the nationality attribute were used, instead of treating this attribute as binary (157 intersectional groups after dropping groups which occur at most once). We found that the improvement of fully Bayesian methods FB over their corresponding PE was greatly magnified in this regime.

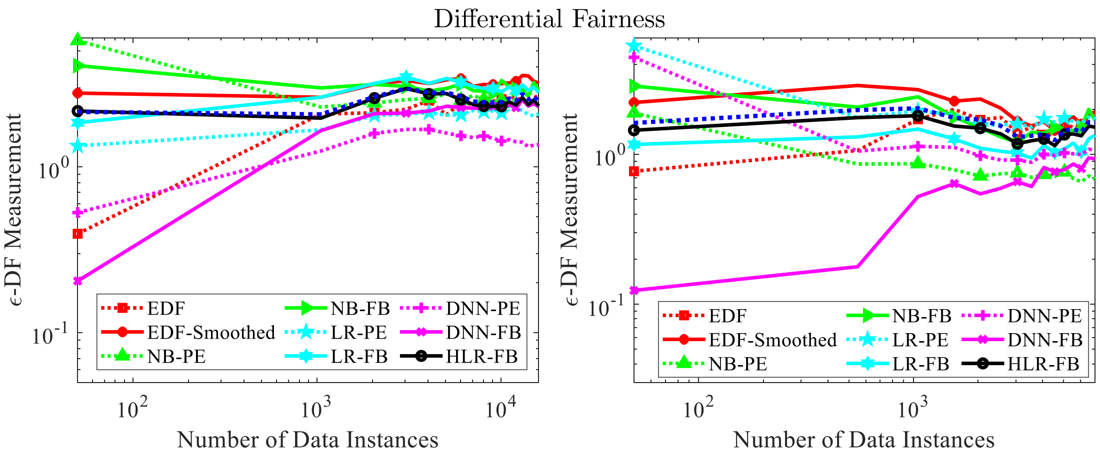
5.2 Fairness Metrics on Semi-Synthetic Data

In this section, we compare all the models with respect to the deviation of their fairness estimates from the ground truth. Since we cannot compute ground truth fairness metrics without knowing the true data distribution θ, we design these experiments on semi-synthetic versions of the Adult and COMPAS datasets. We use the same number of protected attribute values, and instances per intersectional group as for the test datasets, but where the class probabilities are determined by a Gaussian model with a threshold decision boundary.

In our Gaussian threshold model, suppose that x is a "risk score" encoding the untrustworthiness of an individual, generated from a Gaussian given the individual's protected attributes s. The mechanism $M(x) = x \geq t$ assign a "high risk of recidivism" label if the individual's risk score exceeds threshold t. We generate the binary class labels for our semi-synthetic COMPAS and Adult datasets by drawing the same number of instances x per intersectional group as in the original data, and assigning class labels using $M(x)$. We generate the data via $P(x|s) = N(x; \mu = w_s \times \sum_d s_d, \sigma = 1)$, where s_d is the dth protected attribute value for the individual encoded as an integer, $w_s \in (0, 1)$ is a group-specific weight, and $t = 2.5$. We chose $w_s = P_{data}(y = 1|s)$ plus a small constant, thereby making the synthetic data $P(y = 1|s)$'s have some association with the empirical $P_{data}(y = 1|s)$'s. The overall process creates semi-synthetic data with correlations between intersectional groups, and reasonable ground truth values of ϵ and γ.

Figure 1 shows DF and SF estimates for all models on both semi-synthetic datasets, with the gray dotted vertical lines on top of the plots indicating the ground truth ϵ-DF and γ-SF. Fully Bayesian inference allow us to encode uncertainty in the fairness metrics (box-plots) as well as a "best" estimate using the posterior predictive distribution ("X"). Point estimates (PE) via MAP or the empirical distribution (for EDF), are indicated as "O".

Our first finding is that although **empirical estimates** (the "O" for EDF) and **Dirichlet-smoothed estimates** ("X" for EDF) of ϵ and γ were in some cases accurate, in others they **deviated substantially from the true values** (ϵ for Synthetic COMPAS, γ for synthetic Adult). The **Bayesian estimates** of ϵ and γ, using the posterior predictive ("X"), were **closer to the**



(a) Adult **(b) COMPAS**

Figure 2: ϵ-DF measurement of $M(x)$ for (a) logistic regression on the Adult dataset and (b) the COMPAS algorithm, using different $P_{M,\theta}(y|s,\theta)$ models, versus number of instances, averaged over 10 bootstrap samples. The dotted blue line indicates Bayesian ensemble approach. Results for γ-SF are given in the Supplementary Materials.

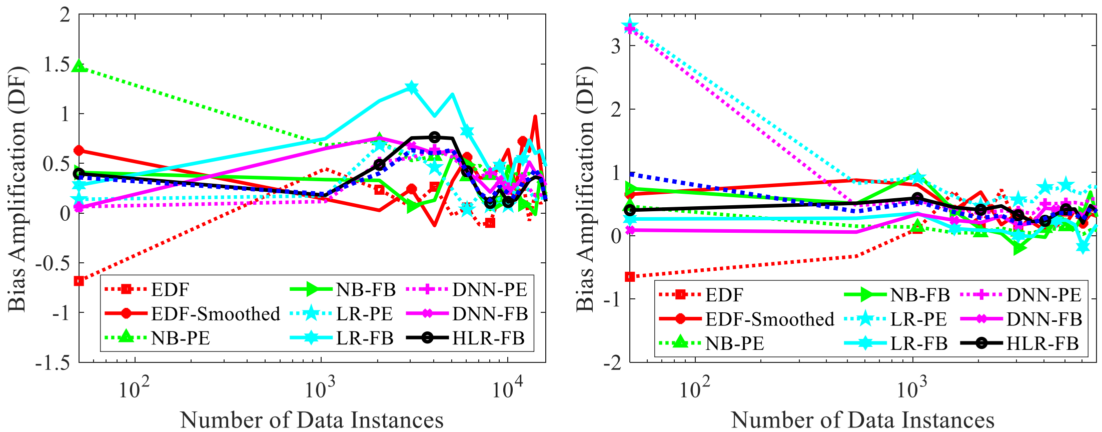
ground truth compared to PE ("O") methods. Note that the **posterior median** ϵ (calculated as an average over the ϵ's corresponding to each posterior sample of $\bar{\theta}$) was sometimes quite **far from the ground truth** (and similarly for γ), and from the posterior predictive estimates "X" (calculated via an average over $p(y|s,\bar{\theta})$'s to compute a single ϵ). Our **HLR approach performed the best**, in that its fairness estimates from the posterior predictive were overall the closest to the ground truth for both ϵ-DF and γ-SF.

5.3 Stability of Estimation vs Data Sparsity We now turn to the study of intersectional fairness estimation on the real datasets. We first investigated the stability of the estimation of the fairness metrics versus data sparsity, by estimating ϵ from bootstrap samples of the datasets, varying the number of samples (Figure 2). For each number of data instances, we generated 10 bootstrap datasets and reported the average ϵ-DF for each model on the Adult and COMPAS datasets. Results for γ-SF were similar, given in the Supplement.

The estimates of ϵ differed greatly between models in the small-data regime, and the models converged to relatively similar estimates as the amount of data increased (similar results were observed for γ). The empirical (EDF) estimates were very noisy with little data, compared to most other models. Except for deep neural networks (DNN), **Bayesian models** (solid lines) were typically found to **converge more quickly in the amount of data** to the consensus full-data estimates, compared to point estimates (dashed lines). The Bayesian DNN (DNN-FB)'s estimates of ϵ deviated substantially from the full data estimates in the low data regime, likely indicating poor performance. This may be due to the overparameterization of the model, and/or convergence issues. The proposed **HLR-FB** model was relatively **stable in the number of instances**, and produced estimates of the fairness metrics which were similar to all models' full-data estimates, even when the number of instances was very small. The Ensemble method also exhibited this behavior.

To analyze the models' performance in the very sparse data regime in more detail, we compared their small data fairness metric estimates, calculated at the left end of the curves in Figure 2 (1% of the data), with the full data "ground truth" (the right end of the curves). We approximate the ground truth ϵ and γ as the median of all bootstrap samples for all the models, where the size of the bootstrap samples is the size of the full dataset, and we report the average L1 distance from the small data estimates and the approximate ground truth (Table 3). We found that **in the sparse data regime**, the **fully Bayesian models** (**FB**) had **smaller deviations** from the full data "**ground truth**" estimates. Our **HLR-FB model performed the best** in this sparse data regime in terms of deviations and stability while the Bayesian model averaging **Ensemble method also showed stable performance** across the different metrics.

In Figure 3, we further studied the impact of dataset size on the **DF bias amplification metric** (Definition 2.2). Since this is calculated as the difference of two noisy estimates of ϵ-DF, the relative noise was higher. The methods differed in the estimated direction of the bias amplification (increase or decrease) in the small data regime, but all pointed to a positive increase with



(a) Adult **(b) COMPAS**

Figure 3: $(\epsilon_2 - \epsilon_1)$-$DF$ bias amplification measurement of algorithm $M(x)$ for (a) logistic regression on the Adult dataset and (b) the COMPAS algorithm, using different $P_{M,\theta}(y|s,\theta)$ models, versus number of data instances, averaged over 10 bootstrap samples. The dotted blue line indicates Bayesian ensemble approach.

Models	1% of Adult Dataset				1% of COMPAS Dataset			
	ϵ-DF		γ-SF		ϵ-DF		γ-SF	
	PE	FB	PE	FB	PE	FB	PE	FB
EDF	2.295	0.0934	0.027	0.021	1.524	0.923	0.037	0.019
NB	3.083	1.387	0.025	0.019	0.593	0.560	0.026	0.022
LR	1.352	0.844	0.015	**0.012**	4.010	0.171	0.034	0.014
DNN	2.160	2.248	0.013	0.032	3.151	1.172	0.031	0.051
HLR	X	**0.089**	X	**0.012**	X	**0.153**	X	**0.013**
Ensemble	0.5488		0.014		1.013		0.016	

Table 3: L1 Deviations of ϵ-DF and γ-SF measurements with 1% of Adult and COMPAS dataset from full dataset "ground truth" estimates, to show the effect of data sparsity (lower is better).

the full data, on both datasets. The **HLR-FB** and **Ensemble** methods were once again the **most stable when given little data**. We report results on a bias amplification version of γ-SF in the Supplement. Note that in these experiments, averaging over bootstraps improves the stability of high variance estimation methods, and the estimates may differ in individual bootstrap samples. Setting aside pure Bayesian or frequentist ideologies, to estimate the fairness metrics in practice it may be useful to **use bootstrap averaging in conjunction with Bayesian models**.

5.4 Case Study on COMPAS As a practical case study, we estimated the intersectional fairness metrics, and their uncertainty via the variational posteriors, on COMPAS (posterior boxplots in the Supplement). To interpret ϵ-DF, note that the 80% rule, used as a legal standard for evidence of disparate impact discrimination [12], finds evidence of discrimination if $\epsilon \geq -\log 0.8 = 0.2231$.[4] All models put most of their posterior density, and their posterior predictive estimates, on values higher than this for true recidivism, the COMPAS system, and its bias amplification. The most reliable model, HLR-FB, predicts that the DF bias amplification of COMPAS is around 0.5-DF, with lower and upper posterior quartiles at around 0.35 and 0.65, respectively. This shows strong evidence that COMPAS increases the bias beyond the inequities in the data. A similar study on Adult is given in the Supplement.

6 Discussion: Practical Recommendations

We showed that fully Bayesian models provide more reliable estimates of intersectional fairness metrics than empirical estimates and point estimates. Our HLR-FB model provides stable estimates compared to other methods, particularly in the very sparse data setting. We found that a Bayesian model averaging ensemble also improves stability in estimation, but it did not outperform HLR-FB on its own. We therefore recommend the use of HLR-FB as a reliable intersectional fairness estimation method with sparse multi-attribute data.

7 Conclusion

We have proposed Bayesian modeling approaches to reliably estimate fairness and its uncertainty in the sparse data regime which arises from multi-attribute intersectional fairness definitions. Our empirical results show the benefits of the probabilistic model-based approach in this setting compared to empirical probabil-

[4] DF calculates ratios of probabilities for all y. Strictly, the 80% rule is calculated on the favorable outcome only.

ity estimates, especially when using Bayesian inference. We proposed a Bayesian hierarchical logistic regression model which provides stable estimates of fairness metrics with sparse intersectional data, and we applied our methods to study the bias in the COMPAS recidivism predictor and a model trained on census data. We plan to develop extensions to model continuous protected attributes, and more sophisticated latent variable models.

References

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica, May*, 23, 2016.

[2] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE, 2019.

[3] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

[4] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression. *4th Annual Workshop on Fairness, Accountability, and Transparency in Machine Learning.*, 2017.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *In Sociological Methods and Research*, 1050:28, 2018.

[6] Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kalai. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in NeurIPS*, 2016.

[7] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[8] Patricia Hill Collins. *Black feminist thought: Knowledge, consciousness, and the politics of empowerment (2nd ed.).* Routledge, 2002 [1990].

[9] Combahee River Collective. A black feminist statement. In Zillah Eisenstein, editor, *Capitalist Patriarchy and the Case for Socialist Feminism.* Monthly Review Press, New York, 1978.

[10] Kimberlé Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *U. Chi. Legal F.*, pages 139–167, 1989.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proc. of the 3rd ITCS*, pages 214–226, 2012.

[12] Equal Employment Opportunity Commission. Guidelines on employee selection procedures. *C.F.R.*, 29. Part 1607, 1978.

[13] J. R. Foulds, R. Islam, K. Keya, and S. Pan. An intersectional definition of fairness. *36th IEEE International Conference on Data Engineering (ICDE) (accepted, in press), arXiv:1807.08362 [CS.LG]*, 2020.

[14] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in NeurIPS*, pages 3315–3323, 2016.

[15] Ursula Hebert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (Computationally-identifiable) masses. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th ICML, PMLR 80*, pages 1944–1953, 2018.

[16] Jennifer Hoeting, David Madigan, Adrian Raftery, and Chris Volinsky. Bayesian model averaging: a tutorial. *Statistical science*, pages 382–401, 1999.

[17] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th ICML, PMLR 80*, pages 2569–2577, 2018.

[18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 100–109, 2019.

[19] Daniel Kifer and Ashwin Machanavajjhala. Pufferfish: A framework for mathematical privacy definitions. *ACM TODS*, 39(1):3, 2014.

[20] Ron Kohavi. Scaling up the accuracy of naive-Bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second SIGKDD*, pages 202–207, 1996.

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in NeurIPS*, 2017.

[22] Cecilia Munoz, Megan Smith, and DJ Patil. *Big data: A report on algorithmic systems, opportunity, and civil rights.* Exec. Office of the President, 2016.

[23] Safiya Umoja Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, 2018.

[24] Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *AAAI/ACM Conf. on AI Ethics and Society*, 2019.

[25] Philip L Roth, Philip Bobko, and Fred S Switzer III. Modeling the behavior of the 4/5ths rule for determining adverse impact: Reasons for caution. *Journal of Applied Psychology*, 91(3):507, 2006.

[26] Till Speicher, Hoda Heidari, Nina Grgic-Hlaca, Krishna P Gummadi, Adish Singla, Adrian Weller, and Muhammad Bilal Zafar. A unified approach to quantifying algorithmic unfairness: Measuring individual & group unfairness via inequality indices. In *Proc. of the 24th ACM SIGKDD*, pages 2239–2248. ACM, 2018.

[27] Sojourner Truth. Ain't I a woman?, 1851. Speech delivered at Women's Rights Convention, Akron, Ohio.

[28] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning (ICML)*, pages 325–333, 2013.