Bayesian Modeling of Intersectional Fairness: The Variance of Bias (Supplementary Material)*

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

1 Appendix: Additional Experimental Results

We repeated the predictive performance experiment for intersectional fairness models with respect to average negative cross-entropy and total variation distance per intersection on the COMPAS dataset as shown in Table 1. For COMPAS, the mechanism $M(x)$ is the COMPAS system itself. Although COMPAS is a black box, we observe its assigned class labels y', and our models extrapolate its behavior on intersectional groups. Similar to the experiment on the Adult dataset, typically Bayesian models outperformed the corresponding point estimates although DNN-PE outperformed all models in one case (negative cross-entropy measurements on COMPAS system-relabeled test set). Our HLR-FB and Bayesian Ensemble method once again provided stable performance over different experimental settings.

The HLR-FB model also showed consistently stable behavior, even with a very small number of instances, producing estimates of ϵ and γ which were similar to the final predictions of all models. The variance in the estimates of fairness was substantial for several models, but averaging over bootstrap samples mitigated this.

In Figure 1, we investigated the stability of the estimation of the subgroup fairness γ-SF versus data sparsity, by estimating γ-SF of the logistic regression and the COMPAS algorithm on the Adult and COMPAS datasets, respectively, varying the number of samples. For each number of data instances, we generated 10 bootstrap datasets and reported the average γ-SF for each model. In Figure 3, we show the results of measuring $(\gamma_2 - \gamma_1)$-SF bias amplification, defined similarly to the DF bias amplification metric. We once again average over 10 bootstrap samples, varying the number of data instances for both Adult and COMPAS datasets. The results are similar to the results we obtained for $(\epsilon_2 - \epsilon_1)$-$DF$. For both experiments, HLR-FB was relatively stable in number of instances and performs similarly to the Bayesian ensemble method.

Finally, we conducted case studies with the Adult and COMPAS datasets where we estimated the intersectional fairness metrics, and their uncertainty via the variational posteriors (Figure 4 and Figure 5, respectively). The results on Adult are in line with those on COMPAS discussed in the main paper, and shown in more detail in Figure 5. All models place high posterior probability on substantially high unfairness values ϵ and γ, and they indicate that the direction of bias amplification is almost certainly positive for both metrics (except the γ measurement on Adult dataset where the bias amplification is roughly symmetric about 0).

2 Appendix: Related Work

Bayesian modeling of fairness has been performed by [6] in the context of stop-and-frisk policing. They model risk probabilities within each protected category, and require algorithms (or people, such as police officers) to threshold these probabilities at the same points when determining outcomes. [4] use Bayesian inference on causal graphical models for fairness. Under their *counterfactual fairness* definition, changing protected attributes A, while holding things which are not causally dependent on A constant, will not change the predicted distribution of outcomes. As an alternative to the Bayesian methodology, adversarial methods are another strategy for managing uncertainty in a fairness context. For example, [1] apply this approach to the setting of ensuring fairness given a limited number of observations in which demographic information is available.

In a legal context, and before there was substantial research on fairness in AI, which was not their focus, [5] and [2] studied various frequentist hypothesis testing methods for the 80% rule [3] in the small data regime. These authors pointed out the dangers of determining adverse impact discrimination with small data and without proper statistical care. Although their emphasis was not on intersectionality, AI fairness, or Bayesian methods, these papers are important precursors to our work.

*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

COMPAS Dataset								
	Negative Cross-entropy				Total Variation Distance			
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	-0.6828	-0.6468	-0.6668	-0.6661	0.1155	0.0972	0.0603	0.0601
NB	-0.6729	-0.6457	-0.6657	-0.6642	0.0958	**0.0632**	0.0618	0.0587
LR	-0.6525	-0.6733	-0.6645	-0.6634	0.0829	0.0734	0.0624	0.0609
DNN	-0.6689	-0.6694	**-0.6600**	-0.6659	0.0908	0.1153	0.0622	0.0627
HLR	X	**-0.6429**	X	-0.6625	X	0.0678	X	**0.0585**
Ensemble	-0.6455		-0.6629		0.0762		0.0591	

Table 1: 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 COMPAS. Here, PE = point estimate, FB = fully Bayesian estimate using the posterior predictive distribution. The best performing method is indicated in bold.

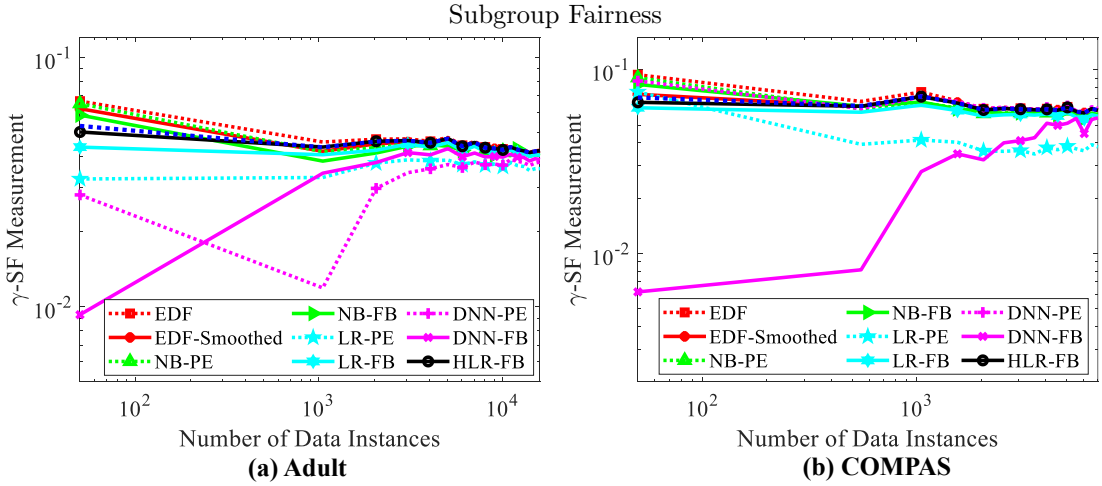

Figure 1: γ-SF measurement of the (a) logistic regression and (b) COMPAS algorithms $M(x)$ using different $P_{M,\theta}(y|s,\theta)$ models on (a) Adult and (b) COMPAS dataset, respectively, with respect to number of data instances with bootstrap data samples. The dotted blue line indicates Bayesian ensemble approach.

References

[1] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Proceedings of 2017 Workshop on Fairness, Accountability, and Transparency in Machine Learning, Halifax, Canada*, 2017.

[2] Michael W Collins and Scott B Morris. Testing for adverse impact when sample size is small. *Journal of Applied Psychology*, 93(2):463, 2008.

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

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

[5] 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.

[6] Camelia Simoiu, Sam Corbett-Davies, Sharad Goel, et al. The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11(3):1193–1216, 2017.

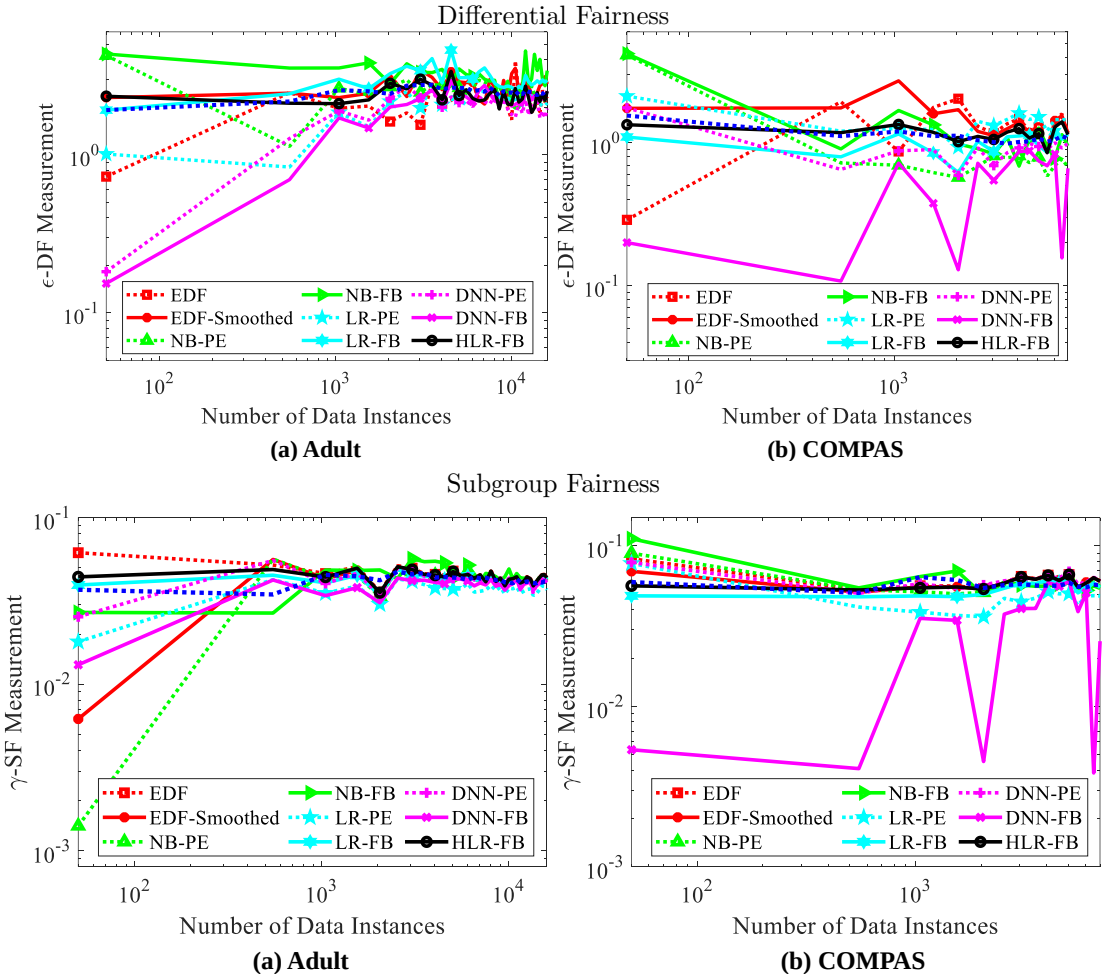

Figure 2: ϵ-DF (top) and γ-SF (bottom) measurement of an algorithm $M(x)$ for (a) logistic regression on the Adult dataset and (b) the COMPAS algorithm, using different $P_{M,\theta}(y|s,\theta)$ models, versus the number of instances, for a randomly chosen bootstrap data sample. For a reference to compare to the other models, we report the average over 10 bootstrap samples for the Bayesian ensemble approach, rather than using a single bootstrap sample, as for the other methods (dotted blue line).

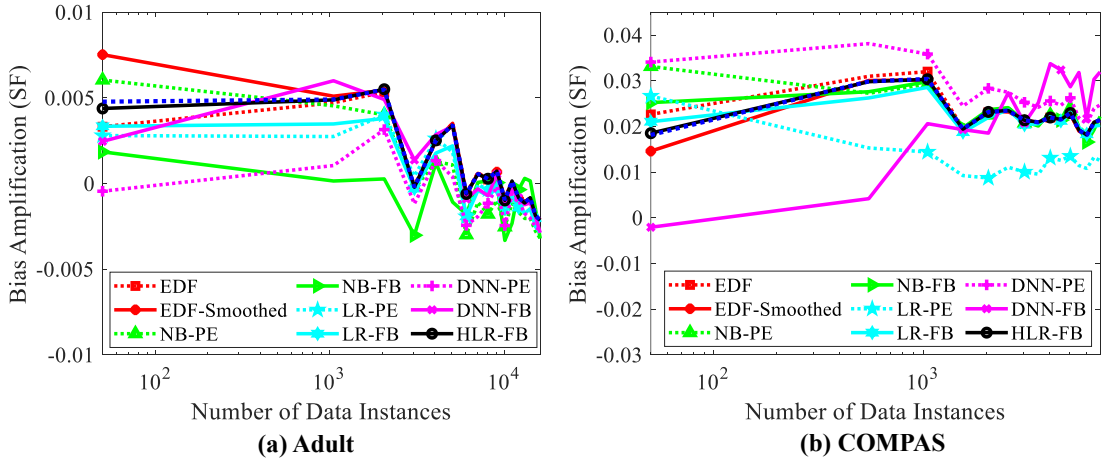

Figure 3: $(\gamma_2 - \gamma_1)$-SF bias amplification by the mechanism $M(x)$ for (a) logistic regression on the Adult dataset and (b) the COMPAS algorithm using different $P_{M,\theta}(y|s,\theta)$ models, with respect to the number of data instances, averaged over 10 bootstrap data samples. The dotted blue line indicates Bayesian ensemble approach.

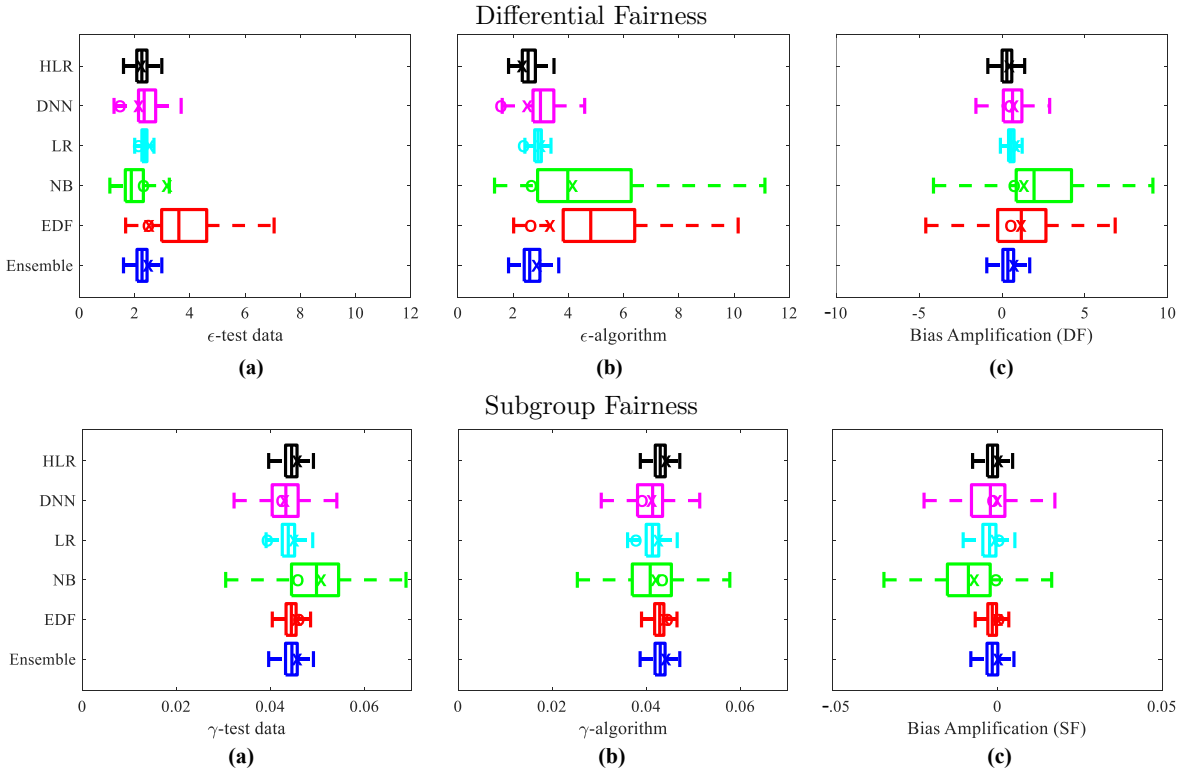

Figure 4: Differential fairness (top) and Subgroup fairness (bottom) estimates using PE and variational posterior distribution of FB to model uncertainty, Adult dataset: Fairness estimates on (a) true label of data, (b) logistic regression $M(x)$-relabeled data, and (c) bias amplification by $M(x)$. The "O" and "X" on top of the box-plots indicate estimates from PE and the posterior predictive distribution of FB models, respectively.

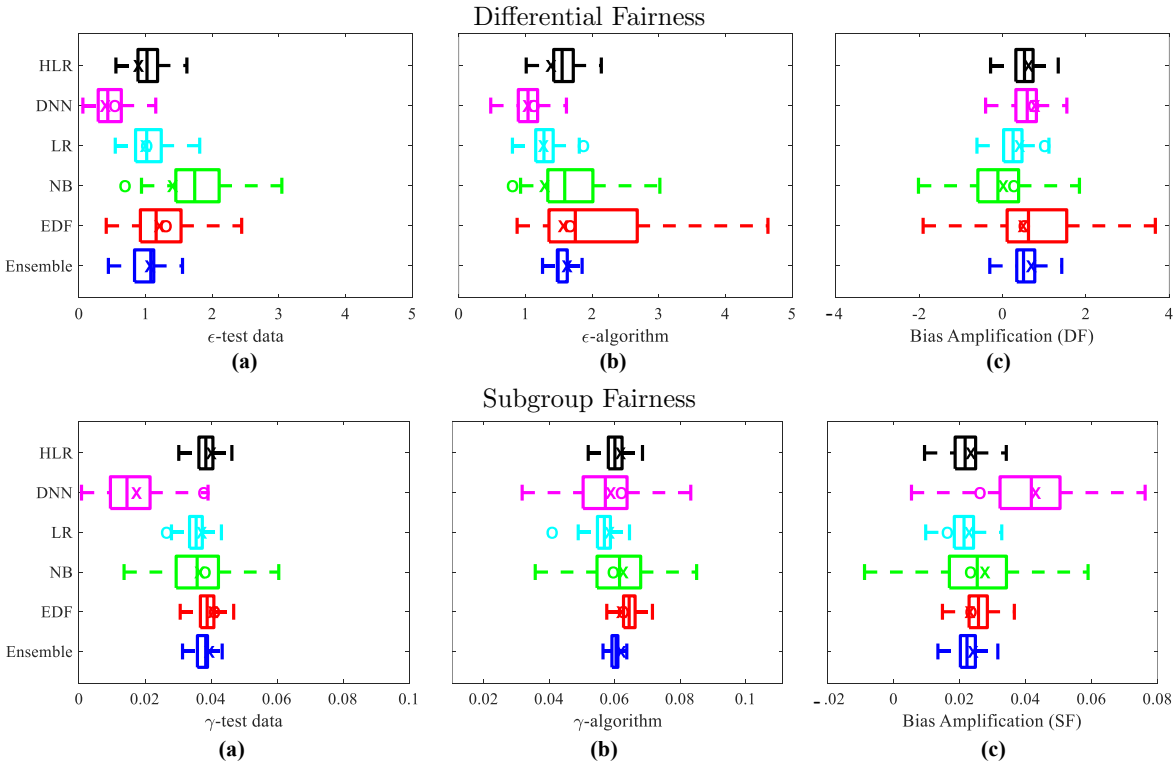

Figure 5: ϵ-DF (top) and γ-SF (bottom) estimates using PE and FB variational posterior distributions, on the COMPAS dataset: (a) fairness estimates on true recidivism label of data, (b) fairness estimates on COMPAS system $M(x)$-relabeled data, and (c) bias amplification by the COMPAS system, for both DF ($\epsilon_{(b)} - \epsilon_{(a)}$) and SF ($\gamma_{(b)} - \gamma_{(a)}$). The "O" and "X" on top of the box-plots indicates estimates from PE models and the posterior predictive distribution of FB models, respectively.