

---

# Debiasing representations by removing unwanted variation due to protected attributes

---

Amanda Bower<sup>\*1</sup> Laura Niss<sup>\*1</sup> Yuekai Sun<sup>\*1</sup> Alexander Vargo<sup>\*1</sup>

## Abstract

We propose a regression-based approach to removing implicit biases in representations. On tasks where the protected attribute is observed, the method is statistically more efficient than known approaches. Further, we show that this approach leads to debiased representations that satisfy a first order approximation of conditional parity. Finally, we demonstrate the efficacy of the proposed approach by reducing racial bias in recidivism risk scores.

## 1. Introduction

In practice, the use of algorithms does not remove all human bias from decision making. There are numerous examples of algorithms with outcomes that are unfair to members of different protected classes (see e.g. Angwin et al. (2016), Steel & Angwin (2010)). In recent years, there has been a flurry of work aimed at correcting this issue. Starting with (Dwork et al., 2011), this work has generally fallen into four categories:

1. Mathematical or statistical definitions of fairness (e.g. Friedler et al. (2016), Ritov et al. (2017)).
2. Algorithms that are modeled to ensure fairness (e.g. Joseph et al. (2016)).
3. Methods of preprocessing data in order to remove inherent bias so that algorithms trained on the debiased data will be fair (e.g. Zemel et al. (2013), Feldman et al. (2015)).
4. Methods of debiasing the outcomes of existing algorithms (a postprocessing step; e.g. Hardt et al. (2016)).

This work falls into the third category of preprocessing. We introduce a factor model prevalent in genetics applications to model the contributions of the protected and permissible attributes to the representation. By treating the variation that is present in the data due to protected attributes (e.g. race)

as unwanted, we devise a method to remove this unwanted variation based on the factor model (and thus debias the data). We show that under certain idealized conditions, the debiased representation is conditionally uncorrelated with the protected attributes. In other words, it satisfies a first order approximation of conditional parity (Ritov et al., 2017) in these cases.

### 1.1. Motivating example

We use ProPublica’s COMPAS dataset and COMPAS risk recidivism scores as an example throughout. More information can be found from (Angwin et al., 2016) and the Practitioners Guide to COMPAS<sup>1</sup>. Much has been written questioning the fairness of these scores with respect to race, with concerns about the disparate false negative and false positive rates between African-Americans and Caucasians.

**Notation:** We denote matrices by uppercase greek or Latin characters and vectors by lowercase characters. A (single) subscript on a matrix indexes its rows (unless otherwise stated). A random matrix  $X \in \mathbb{R}^{n \times d}$  is distributed according to a *matrix-variate normal* distribution with mean  $M \in \mathbb{R}^{n \times d}$ , row covariance  $\Sigma_r \in \mathbb{R}^{n \times n}$ , and column covariance  $\Sigma_c \in \mathbb{R}^{d \times d}$ , which we denote by  $X \sim \text{MN}(M, \Sigma_r, \Sigma_c)$ .

## 2. Related Work

The factor model that motivates the proposed approach (Equation 3.1) is widely used in genetics and was first introduced by (Leek & Storey, 2008) to represent wanted and unwanted variation in gene expression data. It was further exploited by (Gagnon-Bartsch et al., 2013) to develop the removing unwanted variation (RUV) family of methods.

RUV methods rely on knowledge of a set of control genes: genes whose variation in their expression levels are solely attributed to variation in  $Z$ , for example, genes unaffected by the treatments. Formally, a set of controls is a set of indices  $\mathcal{I} \subset [d]$  such that  $B_{\mathcal{I}} = 0$ . Thus

$$Y_{\mathcal{I}} = X A_{\mathcal{I}}^T + E_{\mathcal{I}},$$

where  $Y_{\mathcal{I}}$  and  $E_{\mathcal{I}}$  consist of subsets of the *columns* of  $Y$  and

---

<sup>\*</sup>Equal contribution <sup>1</sup>University of Michigan, Michigan, USA.

<sup>1</sup>[http://www.northpointeinc.com/files/technical\\_documents/FieldGuide2.081412.pdf](http://www.northpointeinc.com/files/technical_documents/FieldGuide2.081412.pdf)

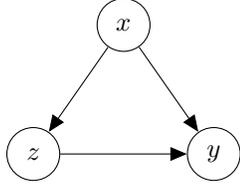


Figure 1. The model (3.1) and (3.2). permissible attributes  $E$ , which suggests estimating  $A_{\mathcal{I}}^T$  by linear regression. This is precisely the “transpose” of the method that we advocate. The task of debiasing features was studied by (Lum & Johndrow, 2016), but there are a few differences between their approach and the proposed one. First, their approach creates new features that satisfy unconditional rather than conditional parity. Second, their approach entails estimating the conditional distributions of the features. This is hard in general, especially if the features are high-dimensional.

### 3. Adjusting for protected attributes

Consider the widely adopted model for matrix-variate data:

$$Y = \begin{matrix} (n \times d) \\ \end{matrix} = \begin{matrix} X \\ (n \times k) \end{matrix} \begin{matrix} A^T \\ (k \times d) \end{matrix} + \begin{matrix} Z \\ (n \times l) \end{matrix} \begin{matrix} B^T \\ (l \times d) \end{matrix} + \begin{matrix} E \\ (n \times d) \end{matrix}. \quad (3.1)$$

The rows of  $Y$  are representations, the rows of  $X$  (resp.  $Z$ ) are protected attributes (resp. permissible attributes) of the samples, and the rows of  $E$  are error terms that represent idiosyncratic variation in the representations. In this paper, we assume  $k, l \ll d$ , but this low-dimensionality is not requisite for the use of the algorithms presented here.

In practice,  $Y$  is usually observed,  $X$  is sometimes observed, and  $Z$  is unobserved. For example, in (Bolukbasi et al., 2016), the representations are embeddings of words in the vocabulary, and the protected attribute is the gender bias of (the embeddings of) words. The rows of  $Z$  are unobserved factor loadings that represent the “good” variation in the word embeddings. In analogy to the framework in (Friedler et al., 2016), the rows of  $Z$  are points in the construct space while the rows of  $Y$  are points in the observed space. We emphasize that like the construct space,  $Z$  is unobserved.

We highlight that we permit non-trivial correlation between the protected and permissible attributes. In other words, we allow the protected attribute to *confound* the relationship between the permissible attribute and the representation (see Figure 1 for a graphical representation of the dependencies between the rows of  $Y$ ,  $X$ , and  $Z$ ). This complicates the task of debiasing the representations. To keep things simple, we assume the regression of  $Z$  on  $X$  is linear:

$$Z = \begin{matrix} (n \times l) \\ \end{matrix} = \begin{matrix} X \\ (n \times k) \end{matrix} \begin{matrix} \Gamma^T \\ (k \times l) \end{matrix} + \begin{matrix} W \\ (n \times l) \end{matrix}. \quad (3.2)$$

The rows of  $W$  are error terms that represent variation in the permissible attributes not attributed to variation in the

protected attributes. We specify the distributions of  $X$ ,  $E$ , and  $W$ , in Sections 3.2 and 3.3.

Our goal is to obtain debiased representations  $Y_{\text{db}}$  such that the debiased representations are uncorrelated with the protected attributes conditioned on the permissible attributes:

$$\text{Cov}[[Y_{\text{db}}]_i, x_i | z_i] = 0. \quad (3.3)$$

This is implied by *conditional parity*:  $[Y_{\text{db}}]_i \perp x_i | z_i$ , and we consider (3.3) as a first-order approximation of conditional parity. An ideal debiased representation is the variation in the representation attributed to the permissible attributes  $ZB^T$ , but this is typically unobservable in practice.

#### 3.0.1. COMPAS EXAMPLE

Under this model, each row of  $Y$  corresponds to a person’s data for recidivism prediction. In our experiments, this includes age, juvenile and adult felony and misdemeanor counts, and whether the offense was a misdemeanor or felony. In this case,  $X$  is a vector, and each component indicates the person’s race. We restrict to Caucasians and African-Americans in our experiments. A person’s true propensity to choose to commit a crime,  $Z$ , is unknown.

#### 3.1. Homogeneous subgroups

The proposed approach relies crucially on knowledge of homogeneous subgroups: groups of samples in which the variation in their representations is mostly attributed to variation in their protected attributes. Formally, we presume knowledge of sets of indices  $\mathcal{I}_1, \dots, \mathcal{I}_G \subset [n]$  such that  $H_g Z_{\mathcal{I}_g} \approx 0$ , where  $H_g = I_{|\mathcal{I}_g|} - \frac{1}{|\mathcal{I}_g|} \mathbf{1}_{|\mathcal{I}_g|} \mathbf{1}_{|\mathcal{I}_g|}^T$  is the centering matrix, for any  $g \in [G]$ . In other words,

$$H_g Y_{\mathcal{I}_g} \approx H_g X_{\mathcal{I}_g} A^T + H_g E.$$

Ideally,  $H_g Z_{\mathcal{I}_g}$  exactly vanishes. This ideal situation arises when the samples in the  $g$ -th group share permissible attributes:  $Z_{\mathcal{I}_g} = \mathbf{1}_{|\mathcal{I}_g|} z_g^T$  for some  $z_g \in \mathbb{R}^l$ .

Intuitively, homogeneous subgroups are groups of samples in which we expect a machine learning algorithm that only discriminates by the permissible attributes to treat similarly. For example, in (Bolukbasi et al., 2016), the homogeneous subgroups are pairs of words that differ only in their gender bias: (*waiter*, *waitress*), (*king*, *queen*).

##### 3.1.1. COMPAS EXAMPLE

In Section 4, we take the homogeneous groups to be people who either did not recidivate within two years or people who did recidivate within two years and were charged with the same degree of felony or misdemeanor. Although  $Z$  is unknown, we expect subjects who go on to commit similar crimes or those who do not recidivate to have similar  $Z$

regardless of race. We emphasize that the homogeneous subgroups are not defined by having similar attributes in  $Y$ .

### 3.2. Adjustment when the protected attribute is unobserved

We now show that the approach proposed by Bolukbasi et al. (2016) produces debiased representations that satisfy (3.3). When the protected attributes are not observed, it is generally not possible to attribute variation in the representations to variation in the protected and permissible attributes. Thus, Bolukbasi et al. (2016) settle on removing the variation in the representations in the subspace spanned by the protected attributes. In other words, we debias the representations by projecting them onto the orthocomplement of  $\mathcal{R}(A)$ .

Formally, let  $Q_g \in \mathbb{R}^{|\mathcal{I}_g| \times (|\mathcal{I}_g|-1)}$  be a subunitary matrix such that  $\mathcal{R}(Q_g)$  coincides with  $\mathcal{R}(H_g)$ . Under (3.4),

$$Q_g^T Y_{\mathcal{I}_g} \approx Q_g^T X_{\mathcal{I}_g} + Q_g^T E,$$

which implies  $\text{Cov}[Q_g^T Y_{\mathcal{I}_g}] \approx \Sigma_E + AA^T$ . This is a factor model, which allows us to consistently estimate  $A$  by factor analysis under mild conditions. We impose classical sufficient conditions for identifiability of  $A$  (Anderson & Rubin, 1956):

1. Let  $A_{-i}$  be the  $(d-1) \times k$  submatrix of  $A$  consisting of all but the  $i$ -th row of  $A$ . For any  $i \in [n]$ , there are two disjoint submatrices of  $A_{-i}$  of rank  $k$ .
2.  $A^T \Sigma_E^{-1} A$  is diagonal, and the diagonal entries are distinct, positive, and arranged in decreasing order.

We remark that the additional assumptions we imposed in this section are a tad stronger than necessary: the assumptions actually imply identifiability of  $A$ , but we only wish to estimate  $\mathcal{R}(A)$ .

In light of the preceding development, here is a natural approach to adjustment when the protected attribute is unobserved:

1. estimate  $A$  by factor analysis:  
 $\arg \min \left\{ \frac{1}{2} \sum_{g=1}^G \|H_g Y_{\mathcal{I}_g} - XA\|_F^2 \right\};$
2. debias  $Y$  by projection onto  $\mathcal{R}(A)^\perp$ :  
 $Y_{\text{db}} = Y(I - P_{\mathcal{R}(A)}),$

which gives

$$\text{Cov}[[Y_{\text{db}}]_i, x_i | z_i] = \text{Cov}[P_{\mathcal{R}(A)^\perp}(Bz_i + e_i) | z_i] = 0.$$

Note that when  $B \subset \mathcal{R}(A)$  the debiased representations will be non-informative because they only contain noise.

### 3.3. Adjustment if the protected attribute is observed

If the protected attribute is observed, it is straightforward to debias the representations. The main challenge here is

estimating  $A$ . Once we have a good estimator  $\hat{A}$ , we debias the representations by subtracting  $X\hat{A}^T$ . We summarize the approach in Algorithm 1.

---

**Algorithm 1** Adjustment if the protected attr. is observed

---

**Input:** representations  $Y \in \mathbb{R}^{n \times d}$ , protected attributes  $X \in \mathbb{R}^{n \times k}$  and groups  $\mathcal{I}_1, \dots, \mathcal{I}_G \subset [n]$

**Estimate  $A$  by regression:**

$$\hat{A}^T \in \arg \min \left\{ \frac{1}{2} \sum_{g=1}^G \|Y_g - X_g A^T\|_F^2 \right\},$$

where  $Y_g = Y_{\mathcal{I}_g} - \mathbf{1}_{|\mathcal{I}_g|} \left( \frac{1}{|\mathcal{I}_g|} \mathbf{1}_{|\mathcal{I}_g|}^T Y_{\mathcal{I}_g} \right)$  and  $X_g$  is defined similarly.

**Debias  $Y$ :** subtract the variation in  $Y$  attributed to  $X$  from  $Y$ :  $Y_{\text{db}} = Y - X\hat{A}^T$ .

---

To study the properties of Algorithm 1, we impose the following assumptions on the distributions of  $X$ ,  $E$ , and  $W$ :

$$\begin{aligned} X &\sim \text{MN}(0, I_n, \Sigma_x), \\ E | (X, Z) &\sim \text{MN}(0, I_n, \Sigma_\epsilon). \end{aligned} \quad (3.4)$$

**Proposition 3.1.** Let  $Z_g$  and  $E_g$  be defined similarly as  $Y_g$  and  $X_g$ . Under conditions (3.1), (3.2), and (3.4),

$$\hat{A}^T - A^T | (X, Z) \sim \text{MN}(T \sum_{g=1}^G X_g^T Z_g B^T, T, \Sigma_\epsilon)$$

where  $T = (\sum_{g=1}^G X_g^T X_g)^\dagger$ .

The (conditional) bias in the OLS estimator of  $A$  depends on the similarity of the permissible attributes in homogeneous subgroups. If  $Z_g = 0$  for all  $g \in G$ , then  $\hat{A}$  is a (conditionally) unbiased estimator of  $A$ .

**Proposition 3.2.** Under conditions (3.1), (3.2), and (3.4), we have

$$\text{Cov}[y_i - \hat{A}x_i, x_i | z_i] = -B \text{Cov}[\tilde{Z} \tilde{X} (\tilde{X}^T \tilde{X})^\dagger x_i, x_i | z_i].$$

We see that if  $\hat{A} = A$  or  $\tilde{Z} = 0$  then the debiased  $y_i$  is uncorrelated with the protected attributes  $x_i$ .

## 4. Experiments: Debiased representations for recidivism risk scores

We empirically demonstrate the efficacy of Algorithm 1 for reducing racial bias in recidivism risk scores based on data ProPublica<sup>2</sup> used in their investigation of COMPAS scores. We fit our own models to the raw and debiased data. Although simple, the scores output by our model perform comparably to the proprietary COMPAS scores (see Figure 2 and Table 3).

<sup>2</sup><https://github.com/propublica/compas-analysis/blob/master/compas-scores-two-years.csv>

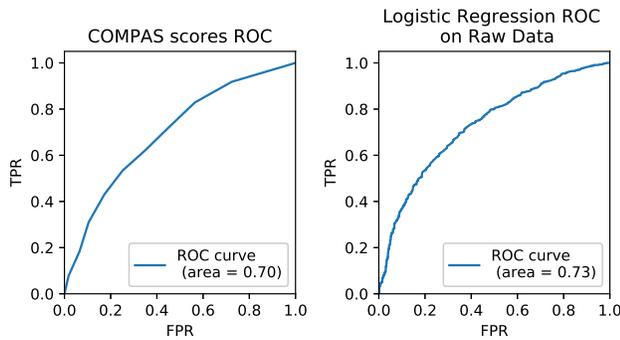


Figure 2.

In particular, we show that logistic regression (LR) trained on debiased data obtained from Algorithm 1 reduces the magnitude of the difference in the false positive rates (FPR) and false negative rates (FNR) between Caucasians and African-Americans (AA) compared to LR trained on raw, potentially biased data. This “fairer” outcome is achieved with a relatively small impact on the percentage of correct predictions. The variables in our LR model are discussed in Section 3.0.1.

We split our data into three pieces: a training set used to estimate  $A$  from Equation (3.1), and a train and test set to evaluate the performance of the learned model. Figure 3 shows the distribution of the probabilities of recidivism for African-Americans according to a logistic model based on the raw and debiased representations. The distribution of the probabilities from the raw representation is skewed to the right. In particular, the right tail of the distribution of probabilities from the raw representation is noticeably heavier than that from the debiased representation.

The ROC curve of the LR model trained on raw data is similar to the ROC curve of COMPAS scores validating the choice of LR as a proxy for COMPAS scores. See Figure 2. For the remaining discussion, we average all results over 30 splits of the data into train and test sets. The average accuracy (the percentage of correct predictions) is 65% for COMPAS. The accuracy for LR trained on raw data and debiased data is comparable, again validating our proxy and justifying the slight loss in accuracy after debiasing in pursuit of fairer outcomes. See Table 3.

Table 1 and Table 2 show the average FPR and FNR for the LR model before and after debiasing. The two tables differ only in the threshold used to declare someone at risk for recidivism based on his or her logistic score; we choose to examine the 50th and 80th quantiles of LR scores since Northpointe specifies that COMPAS scores above the of 50th (respectively 80th) quantile are said to indicate a “Medium” (respectively “High”) risk of recidivism. In Table 1, we see that there is no difference in FPR after debiasing. The

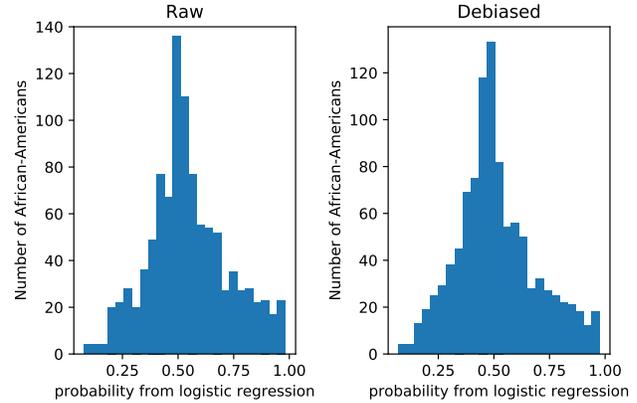


Figure 3. Distribution of recidivism probabilities from raw and debiased representations for African-Americans.

difference in FNR between the races goes from nearly 20% before debiasing to 4% after debiasing. In Table 2, we see FNR are nearly equalized, whereas the magnitude of the difference of FPR between both race groups is improved. However, now Caucasians suffer from disparate impact of FPR instead of African-Americans.

	LR raw		LR debiased	
	FPR (SE)	FNR (SE)	FPR (SE)	FNR (SE)
Population	0.8 (0.01)	0.68 (0.01)	0.9 (0.01)	0.69 (0.01)
Caucasian	0.05 (0.01)	0.81 (0.02)	0.9 (0.02)	0.72 (0.03)
AA	0.11 (0.01)	0.62 (0.01)	0.9 (0.01)	0.68 (0.02)

Table 1. Average proportion FPR and FNR with standard errors (SE) based on the 80th quantile of LR scores.

	LR raw		LR debiased	
	FPR (SE)	FNR (SE)	FPR (SE)	FNR (SE)
Population	0.32 (0.01)	0.32 (0.01)	0.4 (0.01)	0.34 (0.01)
Caucasian	0.22 (0.02)	0.5 (0.02)	0.42 (0.03)	0.31 (0.02)
AA	0.4 (0.02)	0.23 (0.01)	0.27 (0.02)	0.35 (0.02)

Table 2. Average proportion FPR and FNR with standard errors (SE) based on the 50th quantile of LR scores.

Accuracy	LR Raw (SE)	LR Debiased (SE)	COMPAS (SE)
50 quantile	0.67 (.011)	0.65 (.01)	0.65 (.008)
80 quantile	0.61 (.01)	0.60 (.01)	0.61 (.01)

Table 3. Proportion of correct predictions (with standard errors) by logistic regression and thresholding COMPAS scores

## 5. Summary and discussion

We study a factor model of representations that explicitly models the contributions of the protected and permissible attributes. Based on the model, we propose an approach to debias the representations. We show that under certain conditions, we can guarantee first order conditional parity for the debiased representations.

## References

- Anderson, Theodore W and Rubin, Herman. Statistical inference in factor analysis. In *Proceedings of the third Berkeley symposium on mathematical statistics and probability*, volume 5, pp. 111–150, 1956.
- Angwin, Julia, Larson, Jeff, Mattu, Surya, and Kirchner, Lauren. Machine bias. *ProPublica*, 2016. URL <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.
- Bolukbasi, Tolga, Chang, Kai-Wei, Zou, James Y, Saligrama, Venkatesh, and Kalai, Adam T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, pp. 4349–4357, 2016.
- Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard S. Fairness through awareness. *CoRR*, abs/1104.3913, 2011. URL <http://arxiv.org/abs/1104.3913>.
- Feldman, Michael, Friedler, Sorelle A, Moeller, John, Scheidegger, Carlos, and Venkatasubramanian, Suresh. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.
- Friedler, Sorelle A., Scheidegger, Carlos, and Venkatasubramanian, Suresh. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016. URL <http://arxiv.org/abs/1609.07236>.
- Gagnon-Bartsch, Johann A, Jacob, Laurent, and Speed, Terence P. Removing unwanted variation from high dimensional data with negative controls. *Berkeley: Tech Reports from Dep Stat Univ California*, pp. 1–112, 2013.
- Hardt, Moritz, Price, Eric, and Srebro, Nathan. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016. URL <http://arxiv.org/abs/1610.02413>.
- Joseph, Matthew, Kearns, Michael J., Morgenstern, Jamie, and Roth, Aaron. Fairness in learning: Classic and contextual bandits. *CoRR*, abs/1605.07139, 2016. URL <http://arxiv.org/abs/1605.07139>.
- Leek, Jeffrey T and Storey, John D. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105(48):18718–18723, 2008.
- Lum, Kristian and Johndrow, James. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.
- Ritov, Y., Sun, Y., and Zhao, R. On conditional parity as a notion of non-discrimination in machine learning. *ArXiv*, June 2017. URL <http://arxiv.org/abs/1706.08519>.
- Steel, Emily and Angwin, Juila. On the web’s cutting edge, anonymity in name only. *The Wall Street Journal*, 2010. URL <https://www.wsj.com/articles/SB10001424052748703294904575385532109190198>.
- Zemel, Rich, Wu, Yu, Swersky, Kevin, Pitassi, Toni, and Dwork, Cynthia. Learning fair representations. In *International Conference on Machine Learning*, pp. 325–333, 2013.