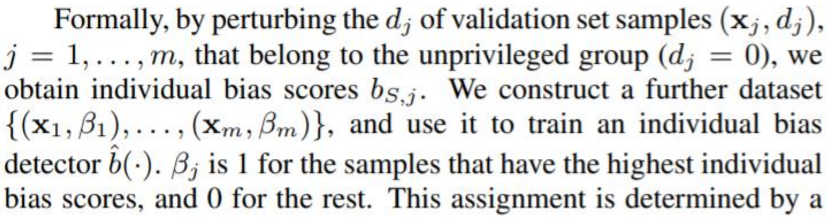
Pranay K. Lohia, Karthikeyan Natesan Ramamurthy, Manish Bhide, Diptikalyan Saha, Kush R. Varshney, Ruchir Puri. *BIAS MITIGATION POST-PROCESSING FOR INDIVIDUAL AND GROUP FAIRNESS*. Brighton, UK: Institute of Electrical and Electronics Engineers, 2018. Report. Link: https://arxiv.org/pdf/1812.06135.pdf

As mentioned in the annotated bibliography, this article dealt with debiasing through post-processing. This article specifically targets individual bias, but uses this to improve both individual and group fairness

Individual Bias Detector Algorithm

They chose samples from parts of the data which were most likely to have individual bias. This again would allow for generalizing to include group fairness.

- Use a validation partition so some samples will have individual bias, others will not
- Learn some classifier/detector for bias from this validation set
- Need some classification or anomaly detection algorithm that provides score outputs. This article used a logistic regression, which would be familiar to us as well



Formally, by perturbing the d_j of validation set samples $(\mathbf{x}_j, d_j)$, $j = 1, \ldots, m$, that belong to the unprivileged group ($d_j = 0$), we obtain individual bias scores $b_{S,j}$. We construct a further dataset $\{(\mathbf{x}_1, \beta_1), \ldots, (\mathbf{x}_m, \beta_m)\}$, and use it to train an individual bias detector $\hat{b}(\cdot)$. β_j is 1 for the samples that have the highest individual bias scores, and 0 for the rest. This assignment is determined by a

I had a lot of trouble following this paragraph so correct me where I'm wrong. First of all, I had trouble figuring out what they meant by perturbing, and definitions were not much help, but I think it means to make some slight change or adjustment to a sample or system.

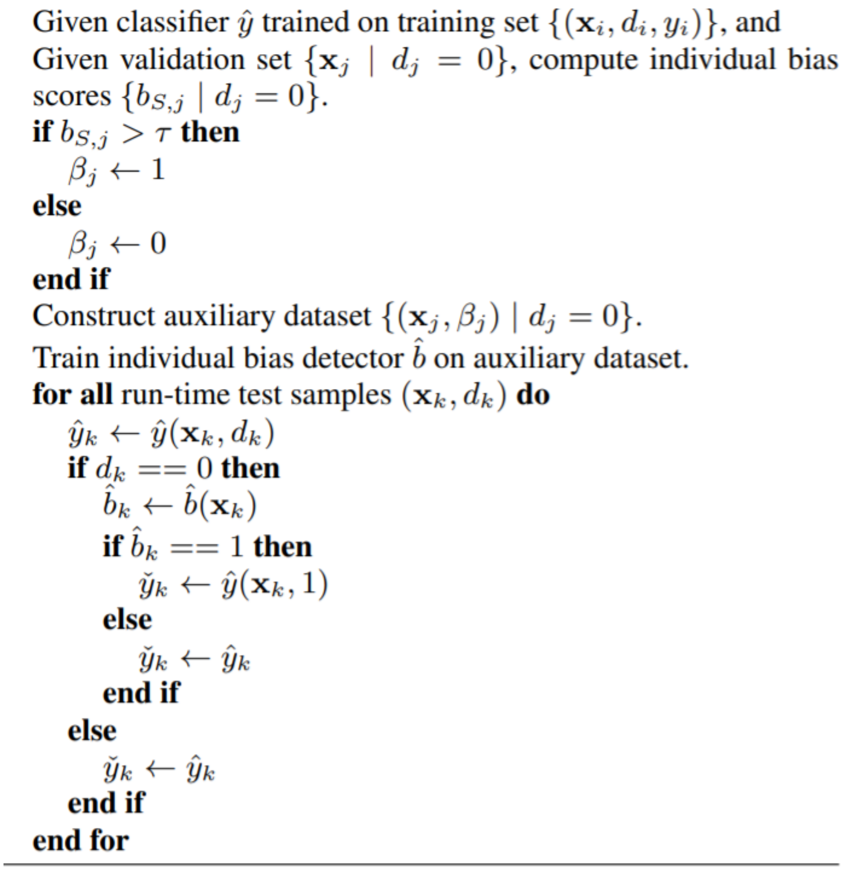
- X references the sample space
- D references privilege, when d_j = 0 this means it is the unprivileged group, d_j = 1 is the privileged

I am assuming that these bias scores are the same as they are in statistics classes, which is defined: expected value of our estimator - our parameter,
= E[**Γ**] - **γ**

- There is some threshold τ which is used to determine whether a sample is categorized as biased or not. This is "based on the disparate impact constraint: from the validation set

Main Algorithm

- Beta hat references the individual bias detector
- Each sample from unprivileged group which is determined to have individual bias is assigned the outcome it would receive had it been in the privileged group
- All other samples are left unchanged



```
Given classifier $\hat{y}$ trained on training set $\{(\mathbf{x}_i, d_i, y_i)\}$, and
Given validation set $\{\mathbf{x}_j \mid d_j = 0\}$, compute individual bias
scores $\{b_{S,j} \mid d_j = 0\}$.
if $b_{S,j} > \tau$ then
    $\beta_j \leftarrow 1$
else
    $\beta_j \leftarrow 0$
end if
Construct auxiliary dataset $\{(\mathbf{x}_j, \beta_j) \mid d_j = 0\}$.
Train individual bias detector $\hat{b}$ on auxiliary dataset.
for all run-time test samples $(\mathbf{x}_k, d_k)$ do
    $\hat{y}_k \leftarrow \hat{y}(\mathbf{x}_k, d_k)$
    if $d_k == 0$ then
        $\hat{b}_k \leftarrow \hat{b}(\mathbf{x}_k)$
        if $\hat{b}_k == 1$ then
            $\check{y}_k \leftarrow \hat{y}(\mathbf{x}_k, 1)$
        else
            $\check{y}_k \leftarrow \hat{y}_k$
        end if
    else
        $\check{y}_k \leftarrow \hat{y}_k$
    end if
end for
```

Closing Notes

Overall they were successful in improving fairness with their methods. An added benefit to this method was its “black box” feature, meaning that you can apply this method without access to the training process and it can be handled in run-time environments. I think that this could be a viable method, although it would still require a deeper dive and maybe some more eyes to fully grasp this.

Amanda Bower, Laura Niss, Yuekai Sun, Alexander Vargo. *DEBIASING REPRESENTATIONS BY REMOVING UNWANTED VARIATION DUE TO PROTECTED ATTRIBUTES.* Proceedings of the 35th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

Link: https://www.fatml.org/media/documents/debiasing_representations.pdf

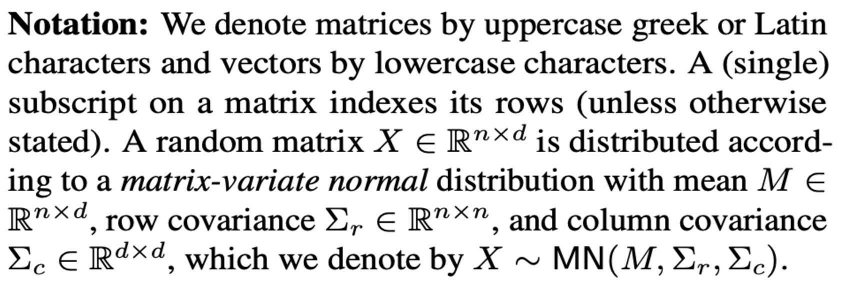
This article helped explain how to debias a dataset through a process called **logistic regression**. This process is something that I personally worked with the past couple of semesters.

Logistic Regression Algorithm

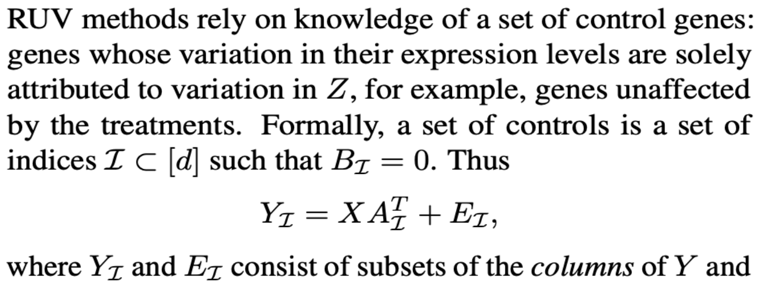
They propose a regression-based approach to re-moving implicit biases in representations. On tasks where the protected attribute is observed, the method is statistically more efficient than known approaches. Further, we show that this approach leads to debiased representations that satisfy a first order approximation of conditional parity.

- Preprocessing - 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
    - The bias code focuses on something specific which they put first and what the code focuses on second turns more bias… (e.g race - CAUC/AA)
- Method was first used for unwanted information used in genetics for gene expression data
- Removing Unwanted Variation (RUV) - basically explains itself, this is the data that is unwanted and removed

NOTATION



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 $X \sim \mathsf{MN}(M, \Sigma_r, \Sigma_c)$.

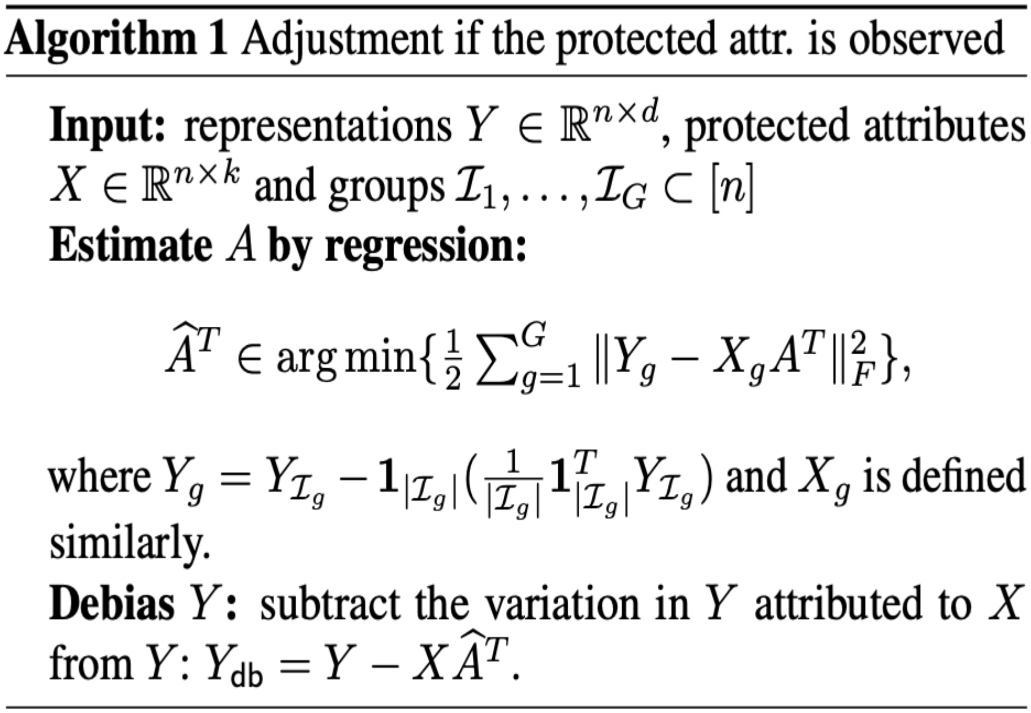


RUV methods rely on knowledge of a set of control genes: genes whose variation in their expression levels are solely attributed to variation in Z, for example, genes unaffected by the treatments. Formally, a set of controls is a set of indices $\mathcal{I} \subset [d]$ such that $B_{\mathcal{I}} = 0$.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

ALGORITHM



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 $\mathcal{I}_1, \ldots, \mathcal{I}_G \subset [n]$

Estimate A by regression:

$$\widehat{A}^T \in \arg\min\{\tfrac{1}{2}\textstyle\sum_{g=1}^{G} \|Y_g - X_g A^T\|_F^2\},$$

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

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

Looking at this algorithm you can tell that X is the protected attribute and you subtract that from Y. From this I had trouble reading the algorithm, not knowing what is typically inside each variable. If someone could help me out that would be great.

In the article there wasn't a "main" algorithm you could essentially follow which led me to have issues.