# A Novel Regularization Approach to Fair Machine Learning

Norman Matloff
University of California, Davis
Wenxi Zhang
Columbia University.

Bay Area R Users Group
GRAIL
July 18, 2023

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

**Plan of this talk:**

Here we will introduce a new method for fair machine learning.

Package code is available in

`https://github.com/matloff/EDFfair`

Norman Matloff
University of California, Davis
Wenxi Zhang
Columbia University.

# Overview

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Overview

Fairness in ML:

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Overview

Fairness in ML:

- The usual ML: predict Y from vector X.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Overview

Fairness in ML:

- The usual ML: predict Y from vector X.
- But X includes a sensitive variable S (race, gender, age etc.)

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Overview

Fairness in ML:

- The usual ML: predict Y from vector X.
- But X includes a sensitive variable S (race, gender, age etc.)
- Wish to exclude S or at least minimize its impact.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Overview

Fairness in ML:

- The usual ML: predict Y from vector X.
- But X includes a sensitive variable S (race, gender, age etc.)
- Wish to exclude S or at least minimize its impact.
- But there may be covariates C in X that are proxies for S,

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Overview

Fairness in ML:

- The usual ML: predict Y from vector X.
- But X includes a sensitive variable S (race, gender, age etc.)
- Wish to exclude S or at least minimize its impact.
- But there may be covariates C in X that are proxies for S, so that you end up "including" S anyway.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Overview

Fairness in ML:

- The usual ML: predict Y from vector X.
- But X includes a sensitive variable S (race, gender, age etc.)
- Wish to exclude S or at least minimize its impact.
- But there may be covariates C in X that are proxies for S, so that you end up "including" S anyway.
- Fairness-Utility Tradeoff: The greater the influence we allow for C, the greater our utility (pred. acc.), but the lesser our fairness.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example

The "Hello World" of Fair ML: COMPAS algorithm

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example

The "Hello World" of Fair ML: COMPAS algorithm

- Northpointe developed commercial product, COMPAS, to predict recidivism.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example

The "Hello World" of Fair ML: COMPAS algorithm

- Northpointe developed commercial product, COMPAS, to predict recidivism.
- Used by judges to aid in determining sentence for convicted criminals.

# Example

The "Hello World" of Fair ML: COMPAS algorithm

- Northpointe developed commercial product, COMPAS, to predict recidivism.
- Used by judges to aid in determining sentence for convicted criminals.
- *ProPublica* expose' claimed COMPAS biased against Black defendants.

# Example

The "Hello World" of Fair ML: COMPAS algorithm

- Northpointe developed commercial product, COMPAS, to predict recidivism.
- Used by judges to aid in determining sentence for convicted criminals.
- *ProPublica* expose' claimed COMPAS biased against Black defendants.
- S = race, C includes # of priors, educ. level etc.

# Criteria for "Fairness"

*Many* criteria have been proposed.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Criteria for "Fairness"

*Many* criteria have been proposed. (This is a research area, after all. :-) )

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Criteria for "Fairness"

*Many* criteria have been proposed. (This is a research area, after all. :-) )
Why does the criterion one uses matter?

- Criterion itself may be biased. (Northpointe claimed this about *ProPublica*.)
- Many ML unfairness remedies are based on exactly satisfying some chosen fairness citerion.

Norman Matloff
University of California, Davis
Wenxi Zhang
Columbia University.

# Example Criteria

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example Criteria

Say S is categorical (e.g. race, gender). Set $\widehat{Y}$ = predicted value or class.
Some common criteria:

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example Criteria

Say S is categorical (e.g. race, gender). Set $\widehat{Y}$ = predicted value or class.
Some common criteria:

- Demographic Parity

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example Criteria

Say S is categorical (e.g. race, gender). Set $\widehat{Y}$ = predicted value or class.
Some common criteria:

- Demographic Parity
  $\widehat{Y}$, $S$ independent

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example Criteria

Say S is categorical (e.g. race, gender). Set $\widehat{Y}$ = predicted value or class.
Some common criteria:

- Demographic Parity
  $\widehat{Y}$, $S$ independent
- Equalized Odds

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example Criteria

Say S is categorical (e.g. race, gender). Set $\widehat{Y}$ = predicted value or class.
Some common criteria:

- Demographic Parity
  $\widehat{Y}$, $S$ independent
- Equalized Odds
  $\widehat{Y}$ independent of S, given Y
  Retrospective, e.g. among those who end up *not* recidivating, $\widehat{Y}$ should not have been affected by S.

There are various others that are popular in the research realm. These can also be phrased in terms of FPR, TPR etc.

Norman Matloff
University of California, Davis
Wenxi Zhang
Columbia University.

# Relaxing a Criterion

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Relaxing a Criterion

- Allow the criterion to be only approximately met.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Relaxing a Criterion

- Allow the criterion to be only approximately met.
- Set a "slider" with which the user can a select a point in the Fairness-Utility spectrum.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Relaxing a Criterion

- Allow the criterion to be only approximately met.
- Set a "slider" with which the user can a select a point in the Fairness-Utility spectrum.
- For continuous Y, use correlation.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Relaxing a Criterion

- Allow the criterion to be only approximately met.
- Set a "slider" with which the user can a select a point in the Fairness-Utility spectrum.
- For continuous Y, use correlation.
- For binary case, we use $R(T, W)$. T and W are $\widehat{Y}$ and S if Y is continuous, $\widehat{P}(Y = 1|X)$ if Y is binary, similarly if S is binary.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Work by Komiyama *et al* and Scutari *et al*

- (Many others, just two examples.)
- Regress Y on S, then regress the residuals U on $\tilde{X} = X$ without the X component.
- User sets an upper bound on $R^2(\widehat{Y}, S)$ to set the level of Fairness-Utility Tradeoff.
- Komiyama use quadratic programming optimization, thus iterative.
- Scutari approach the problem via ridge regression (with $\lambda$ for the regression on U).
- Scutari is implemented in their **fairml** package on CRAN.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Scutari Approach

- Scutari *et al* (2022), "Achieving Fairness with a Simple Ridge Penalty"

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Scutari Approach

- Scutari *et al* (2022), "Achieving Fairness with a Simple Ridge Penalty"
- Lots of technical detail. Summary: first linear regress Y on S, then linear ridge-regress residuals on (non-S part of) X.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Our Approach (Linear Case)

- Again use ridge, but differently.

$$\text{argmin}_b \; ||\mathbb{Y} - \mathbb{X}b||^2 + ||Db||^2 \tag{1}$$

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Our Approach (Linear Case)

- Again use ridge, but differently.

$$\text{argmin}_b \; ||\mathbb{Y} - \mathbb{X}b||^2 + ||Db||^2 \tag{1}$$

- D = diag($d_1, ..., d_p$) is a hyperparameter, set to desirable point in Fairness-Utility Tradeoff.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Our Approach (Linear Case)

- Again use ridge, but differently.

$$\text{argmin}_b \; ||\mathbb{Y} - \mathbb{X}b||^2 + ||Db||^2 \tag{1}$$

- D = diag($d_1, ..., d_p$) is a hyperparameter, set to desirable point in Fairness-Utility Tradeoff.
- Presumably $d_{i_S} = 0$.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Our Approach (Linear Case)

- Again use ridge, but differently.

$$\text{argmin}_b \; ||\mathbb{Y} - \mathbb{X}b||^2 + ||Db||^2 \tag{1}$$

- D = diag($d_1, ..., d_p$) is a hyperparameter, set to desirable point in Fairness-Utility Tradeoff.
- Presumably $d_{i_S} = 0$.
- The positive $d_j$ are for the proxies, i.e. C.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Our Approach (Linear Case)

- Again use ridge, but differently.

$$\text{argmin}_b \; ||\mathbb{Y} - \mathbb{X}b||^2 + ||Db||^2 \tag{1}$$

- D = diag($d_1, ..., d_p$) is a hyperparameter, set to desirable point in Fairness-Utility Tradeoff.
- Presumably $d_{i_S} = 0$.
- The positive $d_j$ are for the proxies, i.e. C.
- Advtantage over Scutari *et al*: Not "One size fits all," different $d_i$ for different $S_i$.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Our Approach (Linear Case)

- Again use ridge, but differently.

$$\text{argmin}_b \; ||\mathbb{Y} - \mathbb{X}b||^2 + ||Db||^2 \tag{1}$$

- D = diag($d_1, ..., d_p$) is a hyperparameter, set to desirable point in Fairness-Utility Tradeoff.
- Presumably $d_{i_S} = 0$.
- The positive $d_j$ are for the proxies, i.e. C.
- Advtantage over Scutari *et al*: Not "One size fits all," different $d_i$ for different $S_i$.
- Closed-form solution for b:

$$b = [\mathbb{X}'\mathbb{X} + D^2]^{-1}\mathbb{X}'\mathbb{Y}$$

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Computational Trick

A well-known trick for ridge-regression generalizes to our setting.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Computational Trick

A well-known trick for ridge-regression generalizes to our setting.

- Set

$$A = \begin{pmatrix} \mathbb{X} \\ D \end{pmatrix} \tag{2}$$

$$B = \begin{pmatrix} \mathbb{Y} \\ 0 \end{pmatrix} \tag{3}$$

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Computational Trick

A well-known trick for ridge-regression generalizes to our setting.

- Set

$$A = \begin{pmatrix} \mathbb{X} \\ D \end{pmatrix} \tag{2}$$

$$B = \begin{pmatrix} \mathbb{Y} \\ 0 \end{pmatrix} \tag{3}$$

- Then run **lm()** as usual, using A and B as the design matrix and response variable data, instead of $\mathbb{X}$ and $\mathbb{Y}$.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Computational Trick

A well-known trick for ridge-regression generalizes to our setting.

- Set

$$A = \begin{pmatrix} \mathbb{X} \\ D \end{pmatrix} \tag{2}$$

$$B = \begin{pmatrix} \mathbb{Y} \\ 0 \end{pmatrix} \tag{3}$$

- Then run **lm()** as usual, using A and B as the design matrix and response variable data, instead of $\mathbb{X}$ and $\mathbb{Y}$.
- This gives us our desired ridge estimator.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Example

Package in `https://github.com/matloff`, small issues.

```
> data(compas)
> z <- qeFairRidgeLog(compas,'two_year_recid',
>    list(decile_score=0.8,gender=0.8,
>       priors_count=0.8,age=0.8),
>    'race',yesYVal='Yes',holdout=NULL)
# try a prediction, like row 1 but age 33 not 6
> newx <- compas[1,-9]
> newx['age'] <- 33
> predict(z,newx)
         1
0.2854387  # 28.5% chance to recidivate
```

Norman Matloff
University of California, Davis
Wenxi Zhang
Columbia University.

# Extension to Other ML Algorithms

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Extension to Other ML Algorithms

- Random forests: Set node-split probability lower for features in C than in the rest of X. (The **ranger** package could be used.)

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Extension to Other ML Algorithms

- Random forests: Set node-split probability lower for features in C than in the rest of X. (The **ranger** package could be used.)
- k-nearest neighbors (k-NN): In defining the distance metric, place smaller weight on the coordinates corresponding to $C$. (Could use **qeKNN** in my forthcoming **qeML** package.)

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# Extension to Other ML Algorithms

- Random forests: Set node-split probability lower for features in C than in the rest of X. (The **ranger** package could be used.)
- k-nearest neighbors (k-NN): In defining the distance metric, place smaller weight on the coordinates corresponding to $C$. (Could use **qeKNN** in my forthcoming **qeML** package.)
- Support vector machines: Apply an $\ell_2$ constraint on the portion of the vector $w$ of hyperplane coefficients corresponding to $C$.

# A Related Package

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# A Related Package

**Coming Soon!**

- **dsld**, Data Science Looks at Discrimination
- Race, gender, age etc.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# A Related Package

**Coming Soon!**

- **dsld**, Data Science Looks at Discrimination
- Race, gender, age etc.
- "Statistical discrimination analysis in a box"

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# A Related Package

**Coming Soon!**

- **dsld**, Data Science Looks at Discrimination
- Race, gender, age etc.
- "Statistical discrimination analysis in a box"
- Will include paired Quarto book teaching the stat concepts needed for investigation of discrimination.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# A Related Package

**Coming Soon!**

- **dsld**, Data Science Looks at Discrimination
- Race, gender, age etc.
- "Statistical discrimination analysis in a box"
- Will include paired Quarto book teaching the stat concepts needed for investigation of discrimination.
- Part I: General discrimination analysis—effect of S. Part II: Fair ML—predict while avoiding use of S.

Norman Matloff University of California, Davis
Wenxi Zhang Columbia University.

# A Related Package

**Coming Soon!**

- **dsld**, Data Science Looks at Discrimination
- Race, gender, age etc.
- "Statistical discrimination analysis in a box"
- Will include paired Quarto book teaching the stat concepts needed for investigation of discrimination.
- Part I: General discrimination analysis—effect of S. Part II: Fair ML—predict while avoiding use of S.
- Anticipated usage includes:
  - Teaching and research in the social sciences/economics.
  - Litigation support.
  - Government agencies.
  - Corporate HR analysis.
  - Consulting.