Achieving Fairness with a Simple Ridge Penalty

Estimates and Uncertainty

Marco Scutari
scutari@bnlearn.com

Dalle Molle Institute for
Artificial Intelligence (IDSIA)

December 12, 2025

The Problem

- Machine learning (statistical?) models are being used in applications where it is crucial to ensure the accountability and fairness of the decisions made based on their outputs.
- Models are trained on historical data that contain various forms of bias, capture those biases and carry them over into current applications resulting in unfair discrimination of certain groups of people.
- The concept of fairness itself is difficult to define because it depends on the type of distortion we wish to limit and how we characterise it mathematically.
- How can we specify fair models that capture the non-discriminating information in the data and disregard the discriminating information?

Say that $\mathbf{y}$ is our response, $\hat{\mathbf{y}}$ are fitted values from the model, $\mathbf{S}$ are the sensitive attributes containing the discriminating information and $\mathbf{X}$ are the other predictors.

- Group fairness: predictions should be similar across the groups identified by the sensitive attributes.
    - Statistical or demographic parity ($\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{S}$).
    - Equality of opportunity ($\hat{\mathbf{y}} \perp\!\!\!\perp \mathbf{S} \mid \mathbf{y}$).
- Individual fairness: individuals that are similar receive similar predictions

$$f(\mathbf{y}, \mathbf{S}) = \sum\nolimits_{i,j} d_1(y_i, y_j) d_2(\mathbf{s}_i, \mathbf{s}_j).$$

Many, many mathematical characterisations in the literature [10, 4, 11].

We can enforce fairness at different stages of the model selection, estimation and validation process [2]:

- Pre-processing: transform the data to remove the underlying discrimination so that models are guaranteed to be fair.
- In-processing: modify model estimation to remove discrimination, either by changing its objective function (typically the log-likelihood) or by imposing constraints on its parameters.
- Post-processing: assess a previously-estimated model, treated as a black box, and alter its predictions to make them fair.

Komiyama *et al.* [7] did:

1. remove the association between $\mathbf{X}$ and $\mathbf{S}$ with $\mathbf{X} = \mathbf{B}^{\mathrm{T}}\mathbf{S} + \mathbf{U}$, estimating $\widehat{\mathbf{B}}_{\mathrm{OLS}} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{X}$;
2. take the decorrelated predictors $\widehat{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{B}}^{\mathrm{T}}_{\mathrm{OLS}}\mathbf{S}$ which contain the component of $\mathbf{X}$ that cannot be explained by $\mathbf{S}$ ($\widehat{\mathbf{U}} \perp\!\!\!\perp \mathbf{S}$);
3. formulate the regression model $\mathbf{y} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$;
4. formulate the fairness constraint
$$R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\mathrm{VAR}(\mathbf{S}\boldsymbol{\alpha})}{\mathrm{VAR}(\hat{\mathbf{y}})} = \frac{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\boldsymbol{\beta}};$$
5. solve the optimisation problem
$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \mathrm{E}\left[(\mathbf{y} - \hat{\mathbf{y}})^2\right] \quad \text{such that} \quad R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r, r \in [0, 1].$$

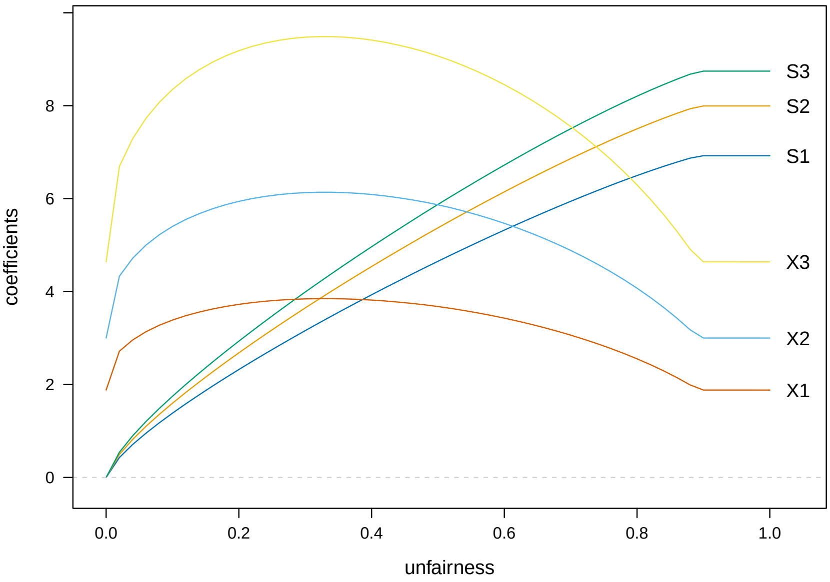
PROS:

- The formulation is simple.
- Discriminating and non-discriminating information are separated.
- The optimisation problem is QCQP, for which there are solvers.
- The fairness constraint is defined in terms of explained variance, the natural measure of information in a linear model.
- The bound is interpretable: 0 = complete fairness, 1 = no constraint.

CONS:

- No distributional assumptions.
- Cannot be extended without losing the ability to use QCQP solvers.
- The behaviour of the estimated coefficients is weird.

Coefficient Profile Plots in Komiyama *et al.*



Take two vintage pieces of statistics from the 1970s-1980s:

1. ridge regression (RR) [6];
2. generalised linear models (GLMs) [9].

With them, we build a Fair (Generalised) Ridge Regression Model (F(G)RRM) that fixes the CONS above and keep all the PROS:

- Modular: swappable characterisation of fairness.
- Versatile: supports all generalised linear models.
- Interpretable: both the model and the fairness constraints are interpretable, and all the best practices from the literature apply.
- Statistical: model selection, model validation, hypothesis testing, confidence intervals, etc. are already available in the literature.

Fair Ridge Regression Model (FRRM)

Let's start again from $\mathbf{y} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. We want to re-create the shrinkage effects on the coefficients $\boldsymbol{\alpha}$ associated with $\mathbf{S}$ that we see in Komiyama *et al.*: we can do that with a ridge penalty,

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = \operatorname*{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{S}\boldsymbol{\alpha} - \widehat{\mathbf{U}}\boldsymbol{\beta}\|_2^2 + \lambda(r)\|\boldsymbol{\alpha}\|_2^2,$$

which we only apply to $\boldsymbol{\alpha}$ because by construction there is no discriminating information in $\widehat{\mathbf{U}}$. The parameter estimates are in closed form:

$$\begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} \end{bmatrix} = \begin{bmatrix} \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda(r)\mathbf{I}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y} \\ (\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}})^{-1}\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{y} \end{bmatrix}.$$

But how do we control the fairness of the model?

Fair Ridge Regression Model (FRRM)

For a given level of fairness $r \in [0, 1]$:

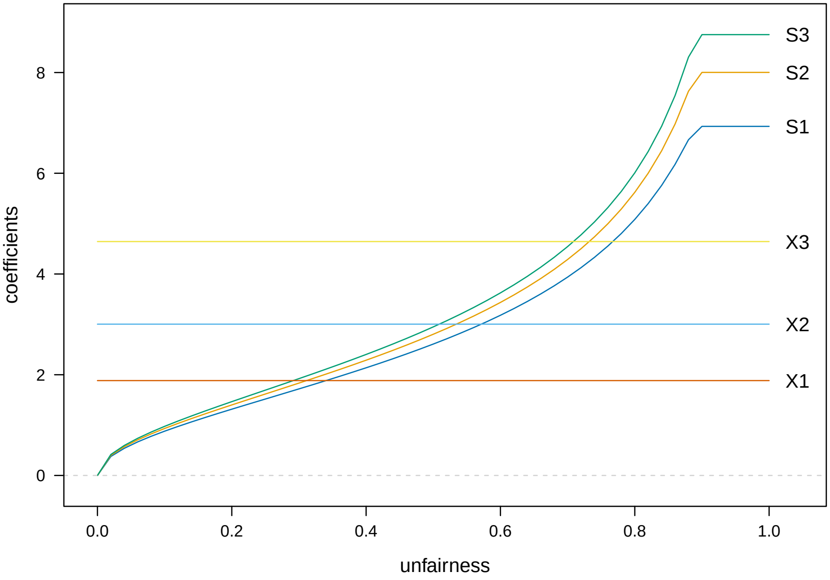
1. Compute $\widehat{\mathrm{U}}$ from $\mathbf{X}, \mathbf{S}$.
2. Estimate $\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} = (\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}})^{-1}\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{y}$.
3. Estimate $\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y}$. Then:
   - 3.1 If $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) \leqslant r$, set $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} = \widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}$.
   - 3.2 Otherwise, find the value of $\lambda(r)$ that satisfies

     $$\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} = \frac{r}{1-r}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$$

     and estimate the associated $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ in the process.

A single solution, requiring a simple univariate root finding algorithm regardless of the number of variables involved.

Coefficients in FRRM



The Flexibility of FRRM

We can easily replace $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ with other constraints.

1. $R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi) = \frac{\mathrm{VAR}(\mathbf{S}\boldsymbol{\phi})}{\mathrm{VAR}(\mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi})}$, from $\hat{\mathbf{y}} = \mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi} + \boldsymbol{\varepsilon}^*$.
2. $f(\boldsymbol{\alpha}, \mathbf{y}, \mathbf{S}) = \sum_{i,j} d(y_i, y_j)(\mathbf{s}_i\boldsymbol{\alpha} - \mathbf{s}_j\boldsymbol{\alpha})^2$ and $D_{\mathrm{IF}} = f(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \mathbf{y}, \mathbf{S}) / f(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \mathbf{y}, \mathbf{S})$,
3. Any convex combination of $R^2_{\mathbf{S}}(\cdot)$, $R^2_{\mathrm{EO}}(\cdot)$, $D_{\mathrm{IF}}(\cdot)$ and others.

We can draw on [5, 12, 13] to estimate the FGRRM

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = \operatorname*{argmin}_{\boldsymbol{\alpha}, \boldsymbol{\beta}} D(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda(r)\|\boldsymbol{\alpha}\|_2^2.$$

where $D(\cdot)$ is the deviance of a GLM + RR, choosing $\lambda(r)$ to give

$$\frac{D(\boldsymbol{\alpha}, \boldsymbol{\beta}) - D(\mathbf{0}, \boldsymbol{\beta})}{D(\boldsymbol{\alpha}, \boldsymbol{\beta}) - D(\mathbf{0}, \mathbf{0})} \leqslant r.$$

A Bayesian Construction Similar to FGRRM?

Take a logistic FGRRM with linear component $\boldsymbol{\eta} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\beta$:

$$\mathbf{y} \mid \widehat{\mathbf{U}}, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\beta} \sim \mathrm{Ber}(\boldsymbol{\pi}), \qquad \boldsymbol{\pi} = \mathrm{logit}^{-1}(\boldsymbol{\eta}) = \frac{\exp(\boldsymbol{\eta})}{1 + \exp(\boldsymbol{\eta})}.$$

We can implement it with Bayesian logistic regression where

$$\boldsymbol{\alpha} \sim N\left(0, (\sigma^2_{\boldsymbol{\alpha}}/\lambda)\mathbf{I}_q\right), \qquad \boldsymbol{\beta} \sim N\left(0, \sigma^2_{\boldsymbol{\beta}}\mathbf{I}_p\right),$$

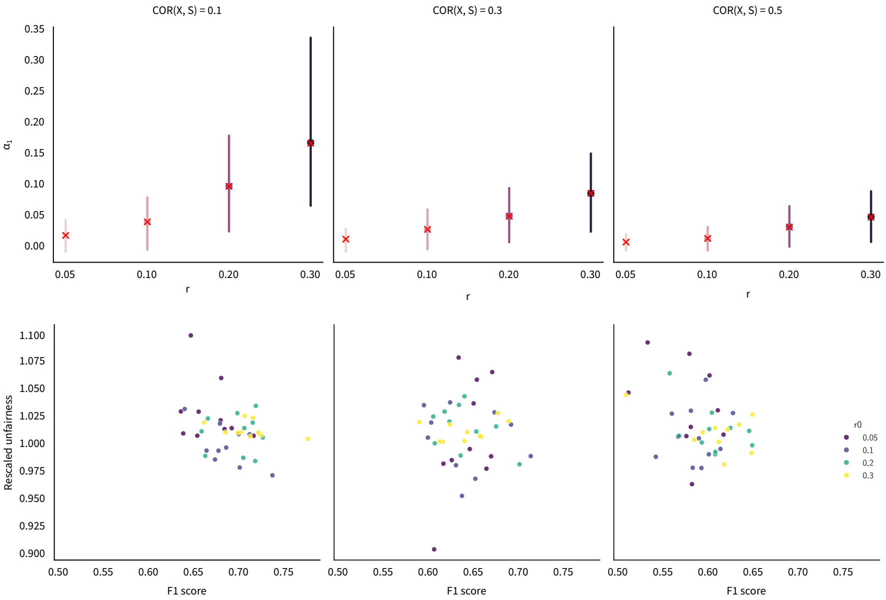
and tie everything together with a prior on $\lambda(r)$

$$\mathrm{P}\left(\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda(r) \mid \mathbf{y}, \widehat{\mathbf{U}}, \mathbf{S}\right) \propto \mathrm{P}\left(\mathbf{y} \mid \widehat{\mathbf{U}}, \mathbf{S}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) \mathrm{P}\left(\boldsymbol{\beta}\right) \mathrm{P}\left(\boldsymbol{\alpha} \mid \lambda(r)\right) \mathrm{P}\left(\lambda(r)\right).$$

$\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\lambda(r)$ are stochastic, and induce a posterior distribution on r through the fairness constraint expression.

$\mathrm{P}(r \mid \widehat{\mathbf{U}}, \mathbf{S} < r_0)$ then controls the fairness of the model.

Marginal Plots from the Posterior Distribution



PROs and CONs

PROs:

- Putting Zellner's g-priors on $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ gives the MLE for $g \to \infty$, which is useful for comparisons.
- Posterior inference (jointly) on $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and r via MCMC.
- Posterior sensitivity analysis on the choice of r.

CONs:

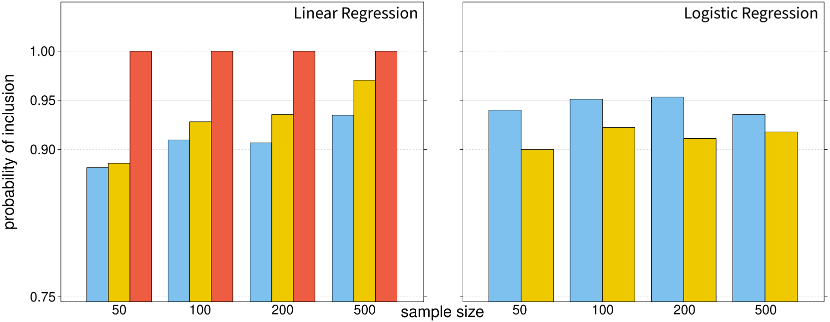
- It appears to be less numerically stable when $\mathbf{X}$ and $\mathbf{S}$ are collinear.
- MCMC convergence to its stationary distribution can be problematic in this and other less-than-smooth settings.

How can we approximate the MCMC inference in the original FGRRM?

Much has been written on confidence intervals for RR coefficients.

For FGRRM:

- Nonparametric boostrap [3] works best.
- Double bootstrap [14] is costly, and works only for FRRM.
- Residual bootstrap [8] does not work.

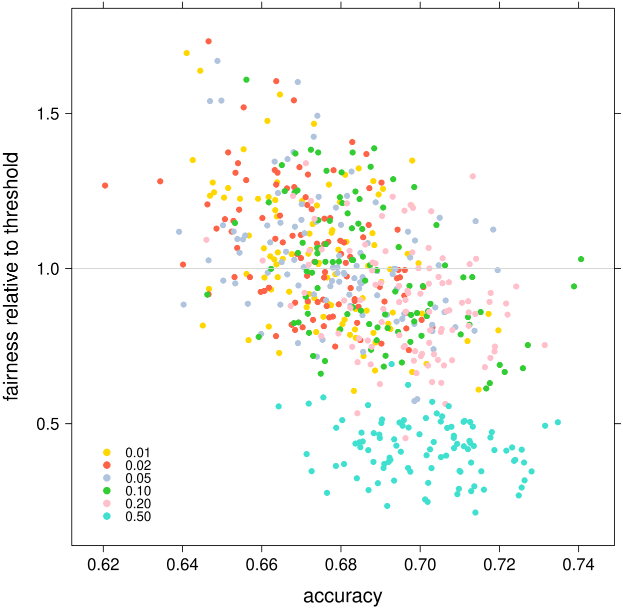


Risk Estimation: Fairness vs Accuracy

At some computational cost, nested bootstrap allows us to marginalise $\boldsymbol{\alpha}, \boldsymbol{\beta}, \lambda(r)$ and training-validation split variability.

1. For $b = 1, \dots, B_1$:
   1.1 Draw a bootstrap sample $\mathbf{y}^{(b)}, \mathbf{X}^{(b)}$ and $\mathbf{S}^{(b)}$ from $\mathbf{y}, \mathbf{X}, \mathbf{S}$.
   1.2 Split training $\left(\mathbf{y}^{(b)}_{\mathrm{TR}}, \mathbf{X}^{(b)}_{\mathrm{TR}}, \mathbf{S}^{(b)}_{\mathrm{TR}}\right)$ and validation $\left(\mathbf{y}^{(b)}_{\mathrm{VA}}, \mathbf{X}^{(b)}_{\mathrm{VA}}, \mathbf{S}^{(b)}_{\mathrm{VA}}\right)$.
   1.3 For $b' = 1, \dots, B_2$:
      1.3.1 Draw a bootstrap sample $\mathbf{y}^{(b')}_{\mathrm{TR}}, \mathbf{X}^{(b')}_{\mathrm{TR}}, \mathbf{S}^{(b')}_{\mathrm{TR}}$ from $\mathbf{y}^{(b)}_{\mathrm{TR}}, \mathbf{X}^{(b)}_{\mathrm{TR}}, S^{(b)}_{\mathrm{TR}}$.
      1.3.2 Estimate a model $\mathcal{M}^{(b')}$ with fairness r from $\mathbf{y}^{(b')}_{\mathrm{TR}}, \mathbf{X}^{(b')}_{\mathrm{TR}}, \mathbf{S}^{(b')}_{\mathrm{TR}}$.
      1.3.3 Estimate the fairness loss for $\mathcal{M}^{(b')}$ on the validation set $\left(\mathbf{y}^{(b)}_{\mathrm{VA}}, \mathbf{X}^{(b)}_{\mathrm{VA}}, \mathbf{S}^{(b)}_{\mathrm{VA}}\right)$.
      1.3.4 Predict $\mathbf{y}^{(b)}_{\mathrm{VA}}$ from $\mathbf{X}^{(b)}_{\mathrm{VA}}, \mathbf{S}^{(b)}_{\mathrm{VA}}$ using $\mathcal{M}^{(b')}$ to estimate predictive accuracy.
2. Assess the joint distribution of empirical fairness and predictive accuracy either by visual inspection or using a pinball loss. [1]

Fairness-Accuracy Trade-Off Plot



- ✓ Introduction
- ✓ Fair Linear Models
- ✓ Fair Ridge Regression
- ✓ Bayesian Fair Regression
- ✓ Uncertainty Quantification
- → Conclusions & Acknowledgements

Conclusions

- Fairness is increasingly a concern as machine learning models become an integral part of automated decision support systems.
- Explainable AI investigates black-box models such as neural networks, but simpler models are also in common use and should be made fair.
- The literature studies fairness as an optimisation problem, producing models whose statistical properties and best practices are unknown.
- Classical statistics provides all the tools to formulate versatile fair models that we know how to interpret and use.
- Uncertainty quantification of fair models remains key, and is challenging because of the nature of the fairness constraint.

Acknowledgements

Ernst Wit

Faculty of Informatics

Università della Svizzera Italiana

Francesca Panero

Department of Methods and Models for Economics, Territory and Finance

Sapienza Università di Roma

Manuel Proissl

Quantum Industry Applications Lead

IBM

Software: https://cran.r-project.org/web/packages/fairml/

That's all!

Happy to discuss in more detail.

scutari@bnlearn.com

[1] Y. Chung, W. Neiswanger, I. Char, and J. Schneider.
Beyond Pinball Loss: Quantile Methods for Calibrated Uncertainty Quantification.
In *Advances in Neural Information Processing Systems*, pages 10971–10984, 2021.

[2] B. D'Alessandro, C. O'Neil, and T. LaGatta.
Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification.
Big Data, 5(2):120–134, 2017.

[3] A. C. Davison and D. V. Hinkley.
Bootstrap Methods and Their Application.
Cambridge University Press, 1997.

[4] E. Del Barrio, P. Gordaliza, and J. M. Loubes.
Review of Mathematical Frameworks for Fairness in Machine Learning, 2020.

[5] J. Friedman, T. Hastie, and R. Tibshirani.
Regularization Paths for Generalized Linear Models via Coordinate Descent.
Journal of Statistical Software, 33(1):1–22, 2010.

[6] A. E. Hoerl and R. W. Kennard.
Ridge Regression: Biased Estimation for Nonorthogonal Problems.
Technometrics, 12(1):55–67, 1970.

[7] J. Komiyama, A. Takeda, J. Honda, and H. Shimao.
Nonconvex Optimization for Regression with Fairness Constraints.
Proceedings of Machine Learning Research, 80:2737–2746, 2018.
35th International Conference on Machine Learning.

[8] P. Legendre.
Comparison of Permutation Methods for the Partial Correlation and Partial Mantel Tests.
Journal of Statistical Computation and Simulation, 67(1):37–73, 2000.

[9] P. McCullagh and J. A. Nelder.
Generalized Linear Models.
CRC press, 2nd edition, 1989.

[10] N. Mehrabi, F. Morstatter, N. Saxena, et al.
A Survey on Bias and Fairness in Machine Learning.
ACM Computing Surveys, 54(6):115, 2021.

[11] D. Pessach and E. Shmueli.
A Review on Fairness in Machine Learning.
ACM Computing Surveys, 55(3):51, 2022.

[12] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani.
Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent.
Journal of Statistical Software, 39(5):1–13, 2011.

[13] J. K. Tay, B. Narasimhan, and T. Hastie.
Elastic Net Regularization Paths for All Generalized Linear Models.
Journal of Statistical Software, 106(1):1–31, 2023.

[14] H. Vinod.
Double Bootstrap for Shrinkage Estimators.
Journal of Econometrics, 68(2):287–302, 1995.