Achieving fairness with a simple ridge penalty

Francesca Panero
London School of Economics
and Political Science

2nd December 2022
OxCSML Seminar Series

LSE Department of Statistics

University of Oxford Department of Statistics

Achieving fairness with a simple ridge penalty

Statistics and Computing (2022) 32:77

Marco Scutari
Dalle Molle Institute
for Artificial Intelligence,
Lugano

Francesca Panero
London School of Economics
and Political Science,
London

Manuel Proissl
Accenture, Zurich
(previously at *UBS*)

How are we going to do this

1. Introduction about algorithmic fairness
   - Individual fairness
   - Group fairness

How are we going to do this

1. Introduction about algorithmic fairness
   - Individual fairness
   - Group fairness

2. Talk about the paper
   - Motivation of the paper
   - Technicalities
   - Experiments

Part 1

MIT Technology Review

Featured Topics Newsletters Events Podcasts

ARTIFICIAL INTELLIGENCE

Facebook's ad-serving algorithm discriminates by gender and race

Even if an advertiser is well-intentioned, the algorithm still prefers certain groups of people over others.

April 5, 2019

By Karen Hao

How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin

May 23, 2016

Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Forbes

AI

Achieving Instagram Growth In The Age Of AI And Algorithmic Bias

Annie Brown Former Contributor
Annie is the founder of Lips, an inclusive creative sharing platform.

Oct 18, 2021, 11:38pm EDT

Algorithms can be biased. How do we correct them?

MIT Technology Review

Featured Topics Newsletters Events Podcasts

ARTIFICIAL INTELLIGENCE

Facebook's ad-serving algorithm discriminates by gender and race

Even if an a
of people c

By Karen Ha

How We Analyzed the COMPAS Recidivism Algorithm

by Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin

May 23, 2016

Proceedings of Ma

Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification*

Joy Buolamwini JOYAB@MIT.EDU
MIT Media Lab 75 Amherst St. Cambridge, MA 02139

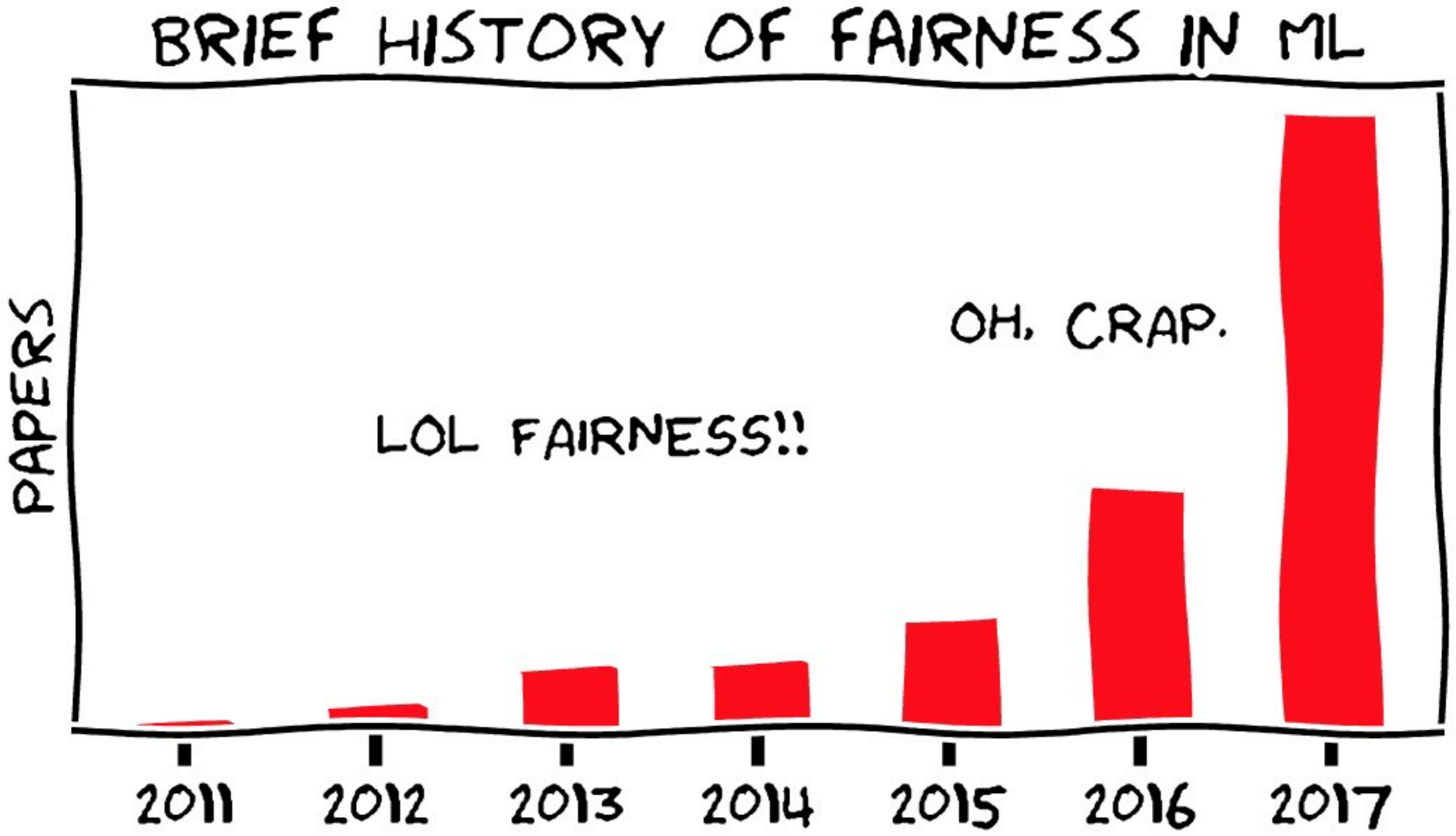
Timnit Gebru TIMNIT.GEBRU@MICROSOFT.COM
Microsoft Research 641 Avenue of the Americas, New York, NY 10011

Instagram Growth In The Age Of AI And Algorithmic Bias

Annie Brown Former Contributor

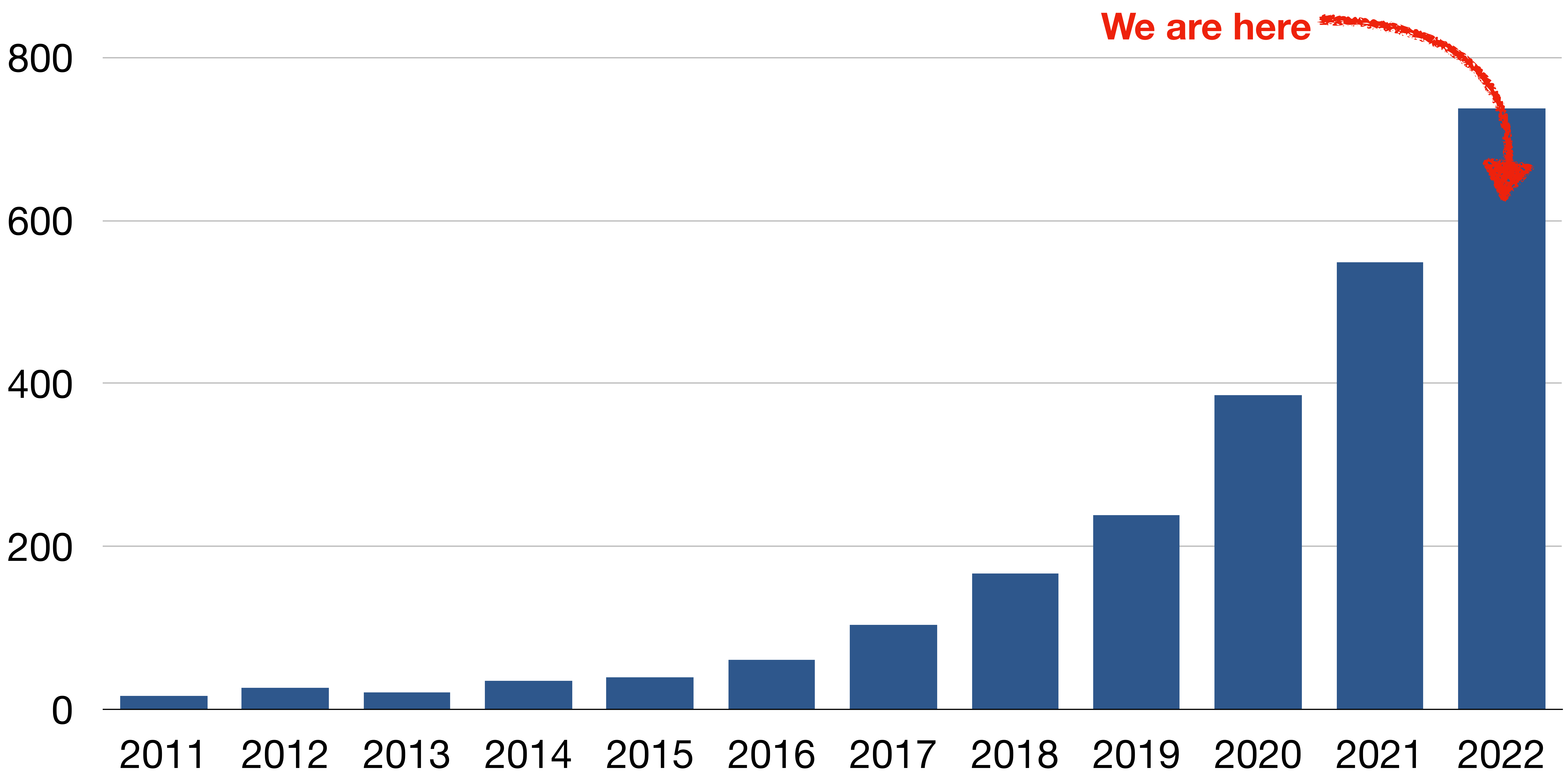
Annie is the founder of Lips, an inclusive creative sharing platform.

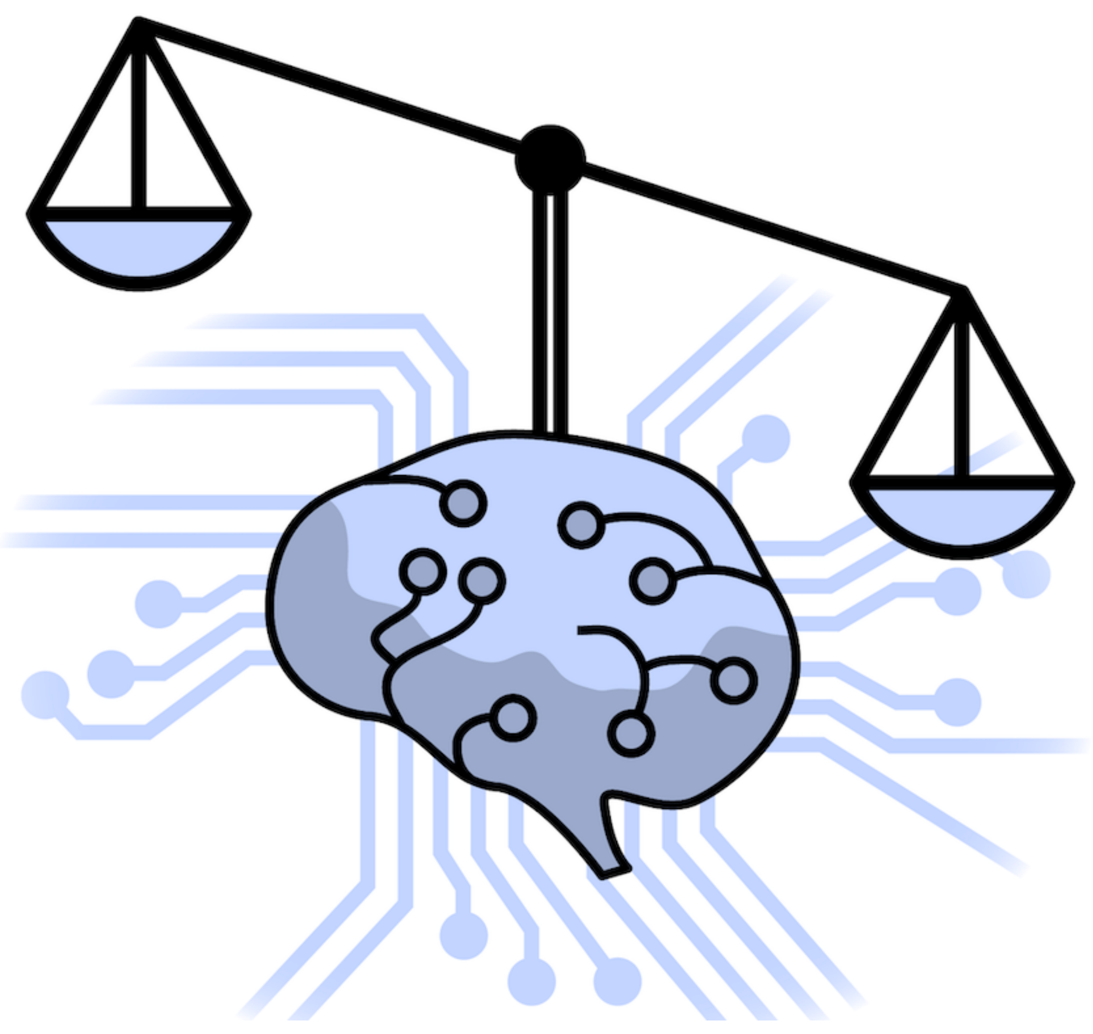
Oct 18, 2021, 11:38pm EDT


BRIEF HISTORY OF FAIRNESS IN ML
PAPERS
LOL FAIRNESS!!
OH, CRAP.
2011
2012
2013
2014
2015
2016
2017

Time evolution of the topic "Fairness"

Number of papers uploaded on arXiv (in CS, Maths and Stats)





Fairness in Machine Learning

Individual fairness

Group fairness

Individual fairness

Treat like cases alike

Aristotle, Nicomachean Ethics (IV century BC)

$$d(\text{decision}(\text{🚶}), \text{decision}(\text{🚶})) \leq d(\text{🚶}, \text{🚶})$$

“Fairness through awareness”, Dwork et al. Proceedings of the 3rd innovations in theoretical computer science conference (2012)
“What's Fair about Individual Fairness?” Fleisher. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2021)

Individual fairness

Treat like cases alike

Aristotle, Nicomachean Ethics (IV century BC)

d(decision(🚶), decision(🚶)) ≤ *d*(🚶, 🚶)

How do you define the distance d?

“Fairness through awareness”, Dwork et al. Proceedings of the 3rd innovations in theoretical computer science conference (2012)
“What's Fair about Individual Fairness?” Fleisher. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2021)

Group fairness

- $\mathbf{y}$ response
- $\mathbf{X}$ non-sensitive covariates
- $\mathbf{S}$ sensitive covariates

Let's focus on groups (identified by their sensitive attributes $\mathbf{S}$)

We'd like a <u>fair</u> group decision

Group fairness

- $\mathbf{y}$ response
- $\mathbf{X}$ non-sensitive covariates
- $\mathbf{S}$ sensitive covariates

Let's focus on groups (identified by their sensitive attributes $\mathbf{S}$)

We'd like a <u>fair</u> group decision

- **Statistical parity:** $\hat{\mathbf{y}}$ independent of $\mathbf{S}$

$$\mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 1) = \mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 0)$$

Group fairness

- $\mathbf{y}$ response
- $\mathbf{X}$ non-sensitive covariates
- $\mathbf{S}$ sensitive covariates

Let's focus on groups (identified by their sensitive attributes $\mathbf{S}$)

We'd like a fair group decision

- **Statistical parity:** $\hat{\mathbf{y}}$ independent of $\mathbf{S}$

$$\mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 1) = \mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 0)$$

- **Equality of odds:** $\hat{\mathbf{y}}$ independent of $\mathbf{S}$, conditional of $\mathbf{y}$

$$\mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 1, \mathbf{y} = \tilde{\mathbf{y}}) = \mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 0, \mathbf{y} = \tilde{\mathbf{y}})$$

Same accuracy and misclassification error among groups

Group fairness

- $\mathbf{y}$ response
- $\mathbf{X}$ non-sensitive covariates
- $\mathbf{S}$ sensitive covariates

Let's focus on groups (identified by their sensitive attributes $\mathbf{S}$)

We'd like a fair group decision

- **Statistical parity:** $\hat{\mathbf{y}}$ independent of $\mathbf{S}$

$$\mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 1) = \mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 0)$$

- **Equality of odds:** $\hat{\mathbf{y}}$ independent of $\mathbf{S}$, conditional of $\mathbf{y}$

$$\mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 1, \mathbf{y} = \tilde{\mathbf{y}}) = \mathbb{P}(\hat{\mathbf{y}} = \mathbf{y}^* \,|\, \mathbf{S} = 0, \mathbf{y} = \tilde{\mathbf{y}})$$

Same accuracy and misclassification error among groups

- ...

Group fairness

- $\mathbf{y}$ response
- $\mathbf{X}$ non-sensitive covariates
- $\mathbf{S}$ sensitive covariates

Let's focus on groups (identified by their sensitive attributes $\mathbf{S}$)

We'd like a fair group decision

- **Statistical parity:** $\hat{\mathbf{y}}$ independent of $\mathbf{S}$

$$|\,\mathbb{P}(\hat{y} = y^* \,|\, S = 1) - \mathbb{P}(\hat{y} = y^* \,|\, S = 0)\,| \leq r$$

User-defined unfairness level

- **Equality of odds:** $\hat{\mathbf{y}}$ independent of $\mathbf{S}$, conditional of $\mathbf{y}$

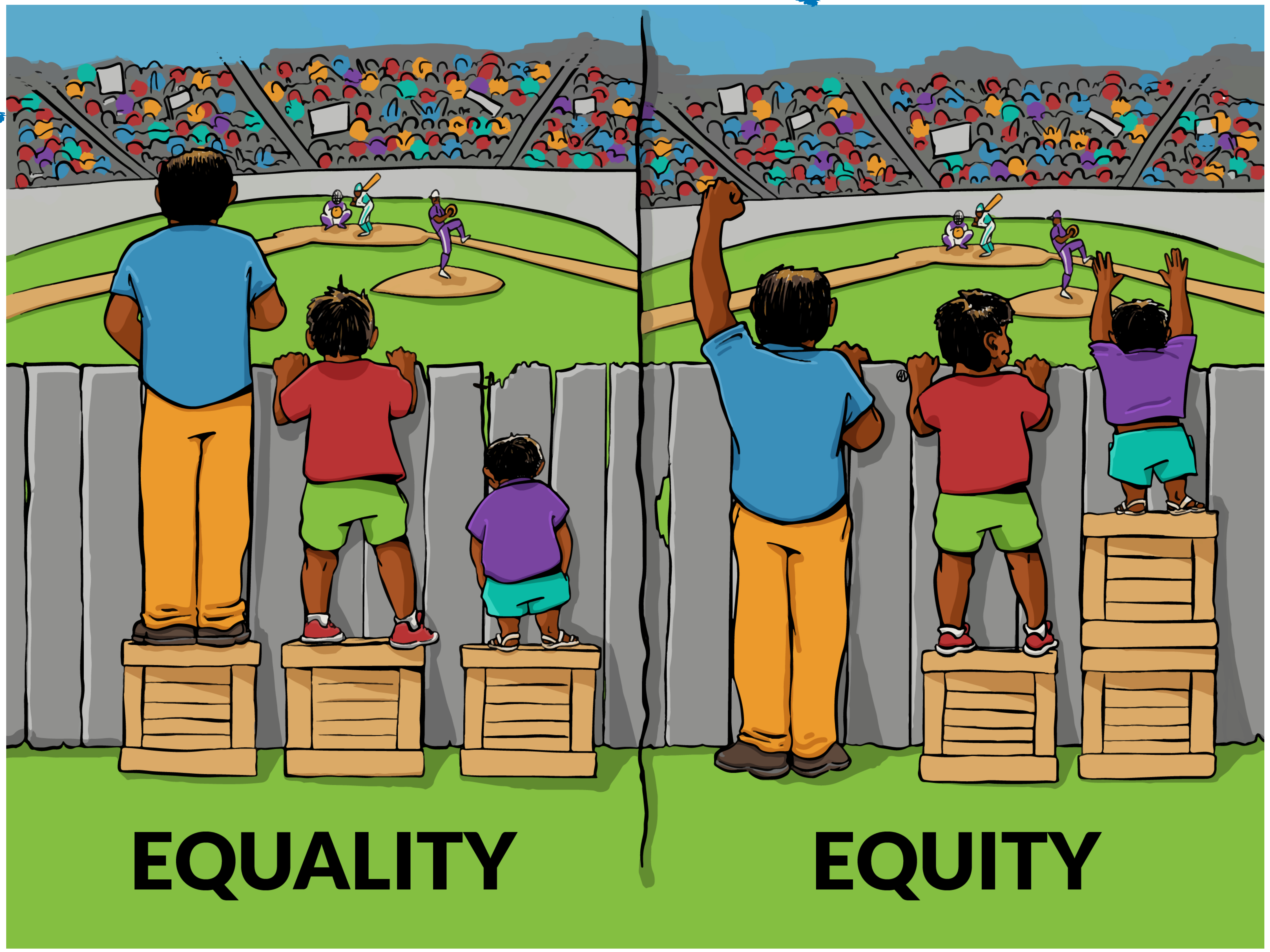
$$|\,\mathbb{P}(\hat{y} = y^* \,|\, S = 1, Y = y) - \mathbb{P}(\hat{y} = y^* \,|\, S = 0, Y = y)\,| \leq r$$

- ...

Equality vs equity

Bias transforming
(statistical parity)

Bias preserving
(equality of odds)



"Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law" Wachter et al W. Va. Law Rev. **123**, 735 (2021)

How do we solve this?

3 main approaches

- Pre-processing of the data

How do we solve this?

3 main approaches

- Pre-processing of the data
- Post-processing of the outcome of the model

How do we solve this?

3 main approaches

- Pre-processing of the data
- Post-processing of the outcome of the model
- Change the model: minimise the loss subject to a fairness criterion

How do we solve this?

3 main approaches

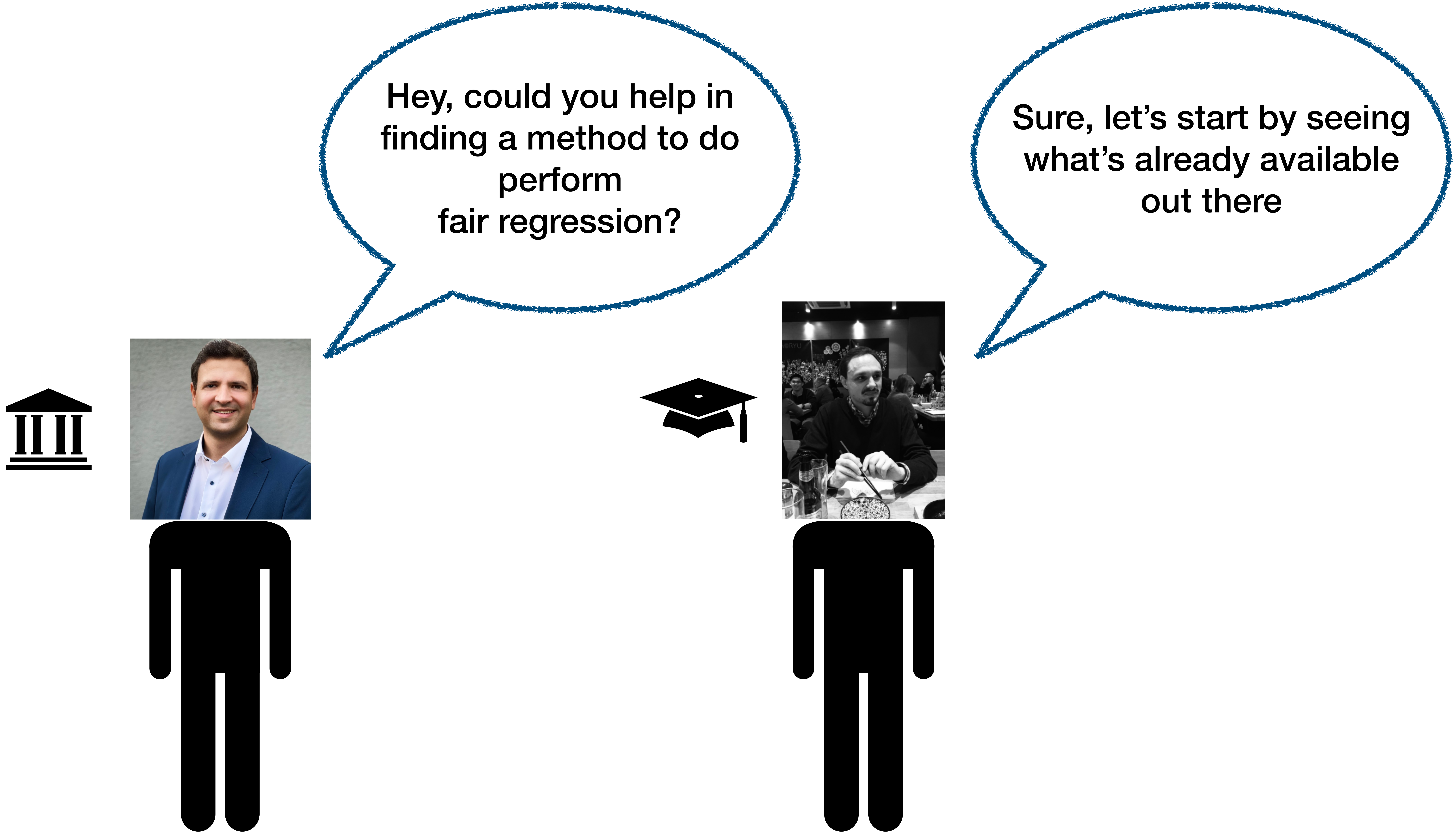
- Pre-processing of the data
- Post-processing of the outcome of the model
- Change the model: minimise the loss subject to a fairness criterion

Part 2

So…what did we do?

Achieving fairness with a simple ridge penalty

Propose a regression model that achieves statistical parity (and other definitions of fairness) using a ridge penalty


Hey, could you help in finding a method to do perform fair regression?
Sure, let's start by seeing what's already available out there

"Nonconvex optimization for regression with fairness constraints"

Komiyama et al. Proceedings of ICML (2018)

Statistical Parity:

$\hat{\mathbf{y}}$ independent of $\mathbf{S}$

- Let's disentangle the contribution of $\mathbf{S}$ from $\mathbf{X}$

$$\mathbf{X} = \mathbf{B}^T\mathbf{S} + \mathbf{U}$$
$$\hat{\mathbf{U}} = \mathbf{X} - \hat{\mathbf{B}}_{OLS}^T\mathbf{S}$$

“Nonconvex optimization for regression with fairness constraints”

Komiyama et al. Proceedings of ICML (2018)

Statistical Parity:

$\hat{\mathbf{y}}$ independent of $\mathbf{S}$

- Let’s disentangle the contribution of $\mathbf{S}$ from $\mathbf{X}$

$$\mathbf{X} = \mathbf{B}^T\mathbf{S} + \mathbf{U}$$
$$\hat{\mathbf{U}} = \mathbf{X} - \hat{\mathbf{B}}^T_{OLS}\mathbf{S}$$

$\hat{\mathbf{U}}$ independent of $\mathbf{S}$

“Nonconvex optimization for regression with fairness constraints”

Komiyama et al. Proceedings of ICML (2018)

Statistical Parity:

$\hat{\mathbf{y}}$ independent of $\mathbf{S}$

- Let’s disentangle the contribution of $\mathbf{S}$ from $\mathbf{X}$

$$\mathbf{X} = \mathbf{B}^T\mathbf{S} + \mathbf{U}$$
$$\hat{\mathbf{U}} = \mathbf{X} - \hat{\mathbf{B}}_{OLS}^T\mathbf{S}$$

$\hat{\mathbf{U}}$ independent of $\mathbf{S}$

$$\mathbf{y} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

“Nonconvex optimization for regression with fairness constraints”

Komiyama et al. Proceedings of ICML (2018)

Statistical Parity:

$\hat{\mathbf{y}}$ independent of $\mathbf{S}$

- Let’s disentangle the contribution of $\mathbf{S}$ from $\mathbf{X}$

$$\mathbf{X} = \mathbf{B}^T\mathbf{S} + \mathbf{U}$$
$$\hat{\mathbf{U}} = \mathbf{X} - \hat{\mathbf{B}}^T_{OLS}\mathbf{S}$$

$\hat{\mathbf{U}}$ independent of $\mathbf{S}$

$$\mathbf{y} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

- Let’s enforce statistical parity through limiting the variance of $\hat{\mathbf{y}}$ explained by $\mathbf{S}$

$\min_{\alpha,\beta} \mathbb{E}[(\mathbf{y} - \hat{\mathbf{y}})^2]$ such that $R^2_S(\alpha, \beta) \leq r$

$$R^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\hat{\mathbf{y}})} = \frac{\alpha^T Var(\mathbf{S})\alpha}{\alpha^T Var(\mathbf{S})\alpha + \beta^T Var(\hat{\mathbf{U}})\beta}$$

“Nonconvex optimization for regression with fairness constraints”

Komiyama et al. Proceedings of ICML (2018)

Statistical Parity:

$\hat{\mathbf{y}}$ independent of $\mathbf{S}$

$r = 0$ Full fairness
$r = 1$ OLS

- Let’s disentangle the contribution of $\mathbf{S}$ from $\mathbf{X}$

$$\mathbf{X} = \mathbf{B}^T\mathbf{S} + \mathbf{U}$$
$$\hat{\mathbf{U}} = \mathbf{X} - \hat{\mathbf{B}}^T_{OLS}\mathbf{S}$$

$\hat{\mathbf{U}}$ independent of $\mathbf{S}$

$$\mathbf{y} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

- Let’s enforce statistical parity through limiting the variance of $\hat{\mathbf{y}}$ explained by $\mathbf{S}$

$\min_{\alpha,\beta} \mathbb{E}[(\mathbf{y} - \hat{\mathbf{y}})^2]$ such that $R^2_S(\alpha, \beta) \leq r$

$$R^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\hat{\mathbf{y}})} = \frac{\alpha^T Var(\mathbf{S})\alpha}{\alpha^T Var(\mathbf{S})\alpha + \beta^T Var(\hat{\mathbf{U}})\beta}$$

Komiyama's approach

To address collinearity in $\mathbf{S}$, they construct $\hat{\mathbf{U}}$ with regularised regression (with penalty λ) which makes $\hat{\mathbf{U}}$ correlated with $\mathbf{S}$.

Let's call this version $\tilde{\mathbf{U}}$.

Komiyama's approach

To address collinearity in $\mathbf{S}$, they construct $\hat{\mathbf{U}}$ with regularised regression (with penalty λ) which makes $\hat{\mathbf{U}}$ correlated with $\mathbf{S}$.

Let's call this version $\tilde{\mathbf{U}}$.

$$R^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha + \hat{\mathbf{U}}\beta)} = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\hat{\mathbf{U}}\beta)}$$

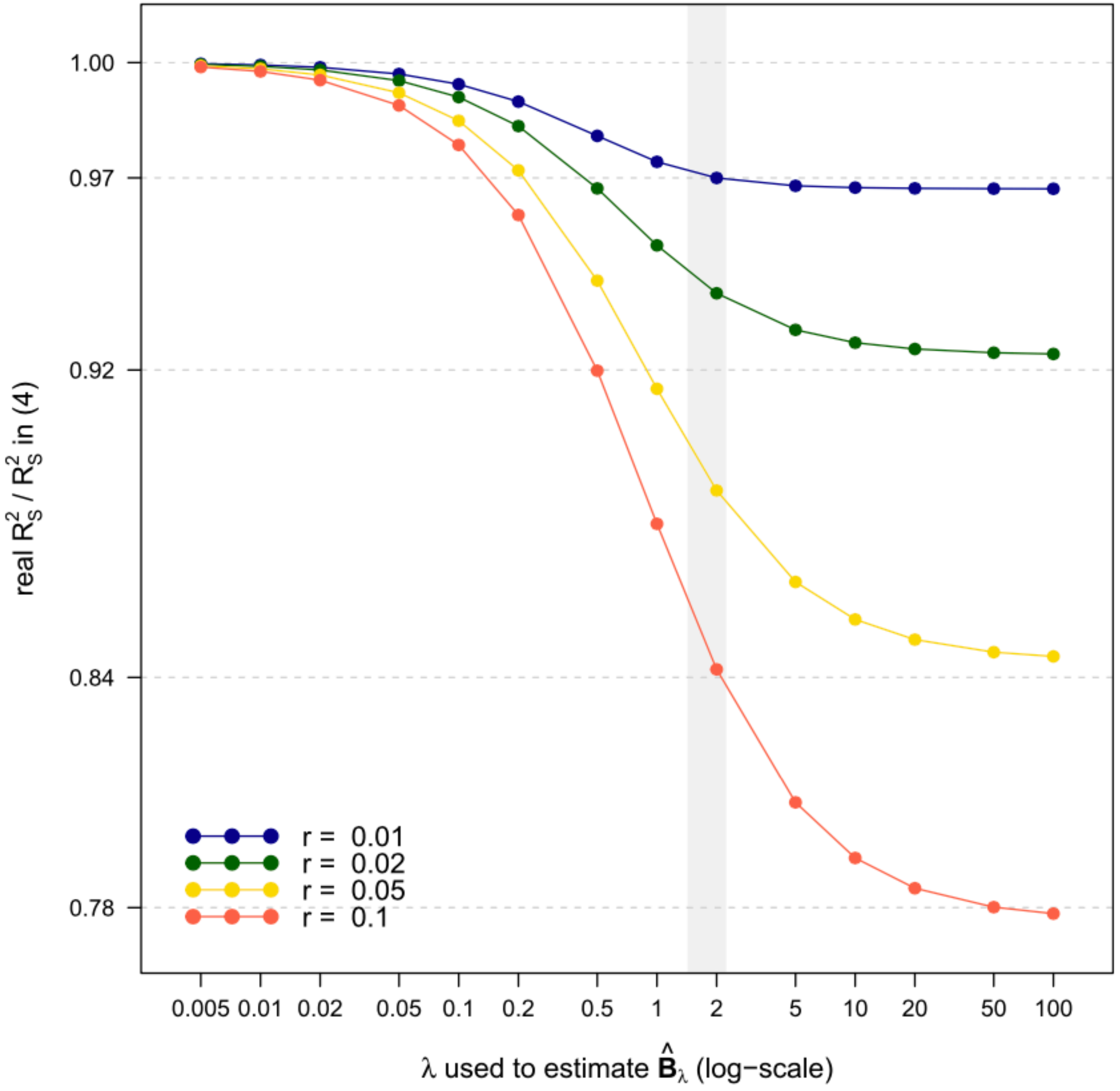
$$\tilde{R}^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha + \tilde{\mathbf{U}}\beta)} = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\tilde{\mathbf{U}}\beta) - 2Cov(\mathbf{S}\alpha, \tilde{\mathbf{U}}\beta)}$$

Komiyama's approach

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ S_1 \\ S_2 \\ S_3 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.3 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 1 & 0.3 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 1 & 0.3 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 1 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 1 & 0.3 \\ 0.3 & 0.3 & 0.3 & 0.3 & 0.3 & 1 \end{bmatrix} \right)$$

$$\mathbf{y} = 2X_1 + 3X_2 + 4x_3 + 5S_1 + 6S_2 + 7S_3 + \boldsymbol{\varepsilon}$$

$$\frac{\tilde{R}^2_{\mathbf{S}}(\alpha, \beta)}{R^2_{\mathbf{S}}(\alpha, \beta)}$$



$\tilde{U}(\lambda)$

$$R^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha + \hat{\mathbf{U}}\beta)} = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\hat{\mathbf{U}}\beta)}$$

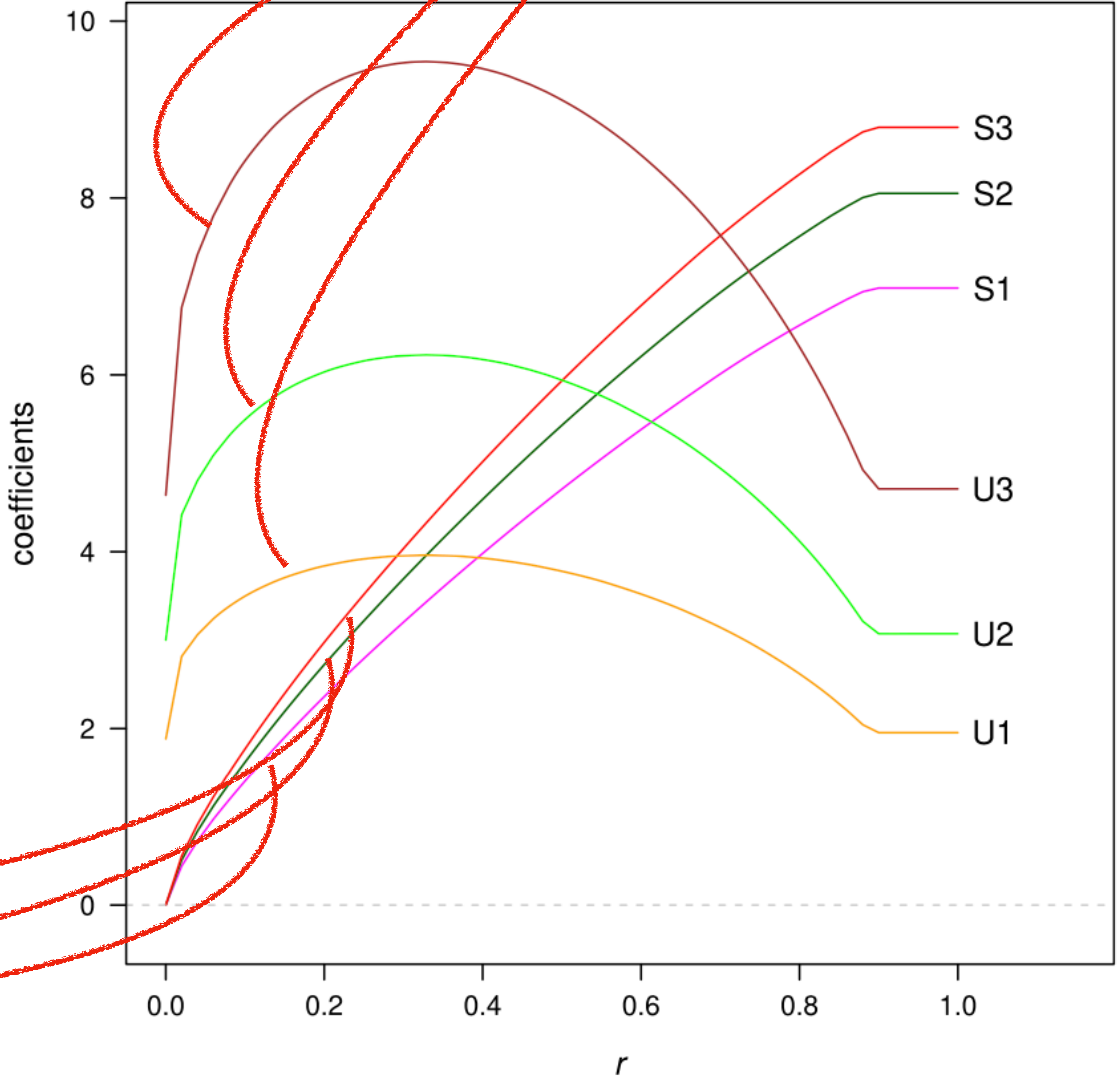
$$\tilde{R}^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha + \tilde{\mathbf{U}}\beta)} = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\tilde{\mathbf{U}}\beta) - 2Cov(\mathbf{S}\alpha, \tilde{\mathbf{U}}\beta)}$$

Komiyama's approach

$\beta_1, \beta_2, \beta_3$

$$\mathbf{y} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

$$\mathbf{y} = \mathbf{S}\alpha + \tilde{\mathbf{U}}\beta + \epsilon$$



$\alpha_1, \alpha_2, \alpha_3$

Our proposal: use ridge regression

$$(\hat{\alpha}_{FRRM}, \hat{\beta}_{FRRM}) = \text{argmin}_{\alpha,\beta} \|\mathbf{y} - \mathbf{S}\alpha - \hat{\mathbf{U}}\beta\|^2 + \lambda(r)\|\alpha\|_2^2,$$

$$\text{with } \lambda(r) \text{ s.t. } R^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\hat{\mathbf{U}}\beta)} \leq r$$

Our proposal: use ridge regression

$$(\hat{\alpha}_{FRRM}, \hat{\beta}_{FRRM}) = \text{argmin}_{\alpha,\beta} \|\mathbf{y} - \mathbf{S}\alpha - \hat{\mathbf{U}}\beta\|^2 + \lambda(r)\|\alpha\|_2^2,$$

$$\text{with } \lambda(r) \text{ s.t. } R^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\hat{\mathbf{U}}\beta)} \leq r$$

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

Our proposal: use ridge regression

$$(\hat{\alpha}_{FRRM}, \hat{\beta}_{FRRM}) = \text{argmin}_{\alpha,\beta} \|\mathbf{y} - \mathbf{S}\alpha - \hat{\mathbf{U}}\beta\|^2 + \lambda(r)\|\alpha\|_2^2,$$

$$\text{with } \lambda(r) \text{ s.t. } R^2_{\mathbf{S}}(\alpha, \beta) = \frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\hat{\mathbf{U}}\beta)} \leq r$$

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

1. Compute $\hat{\mathbf{U}}$
2. $\hat{\beta}_{FRRM} = (\hat{\mathbf{U}}^T\hat{\mathbf{U}})^{-1}\hat{\mathbf{U}}^T\mathbf{y}$
3. Compute $\hat{\alpha}_{OLS} = (\mathbf{S}^T\mathbf{S})^{-1}\mathbf{S}^T\mathbf{y}$.
   If $R^2_{\mathbf{S}}(\hat{\alpha}_{OLS}, \hat{\beta}_{FRRM}) \leq r$: $\hat{\alpha}_{FRRM} = \hat{\alpha}_{OLS}$
   Else: find $\lambda(r)$ s.t. $R^2_{\mathbf{S}}(\hat{\alpha}_{FRRM}, \hat{\beta}_{FRRM}) = r$ and the corresponding $\hat{\alpha}_{FRRM}$

Properties

$$\hat{\mathbf{y}} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

- The problem is guaranteed to have a single solution

$$\frac{Var(\mathbf{S}\alpha)}{Var(\mathbf{S}\alpha) + Var(\hat{\mathbf{U}}\beta)} = r$$

Properties

$$\hat{\mathbf{y}} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

- The problem is guaranteed to have a single solution

$$\frac{Var(\mathbf{S}(\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^T\mathbf{y})}{Var(\mathbf{S}(\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^T\mathbf{y}) + Var(\hat{\mathbf{U}}\beta)} = r$$

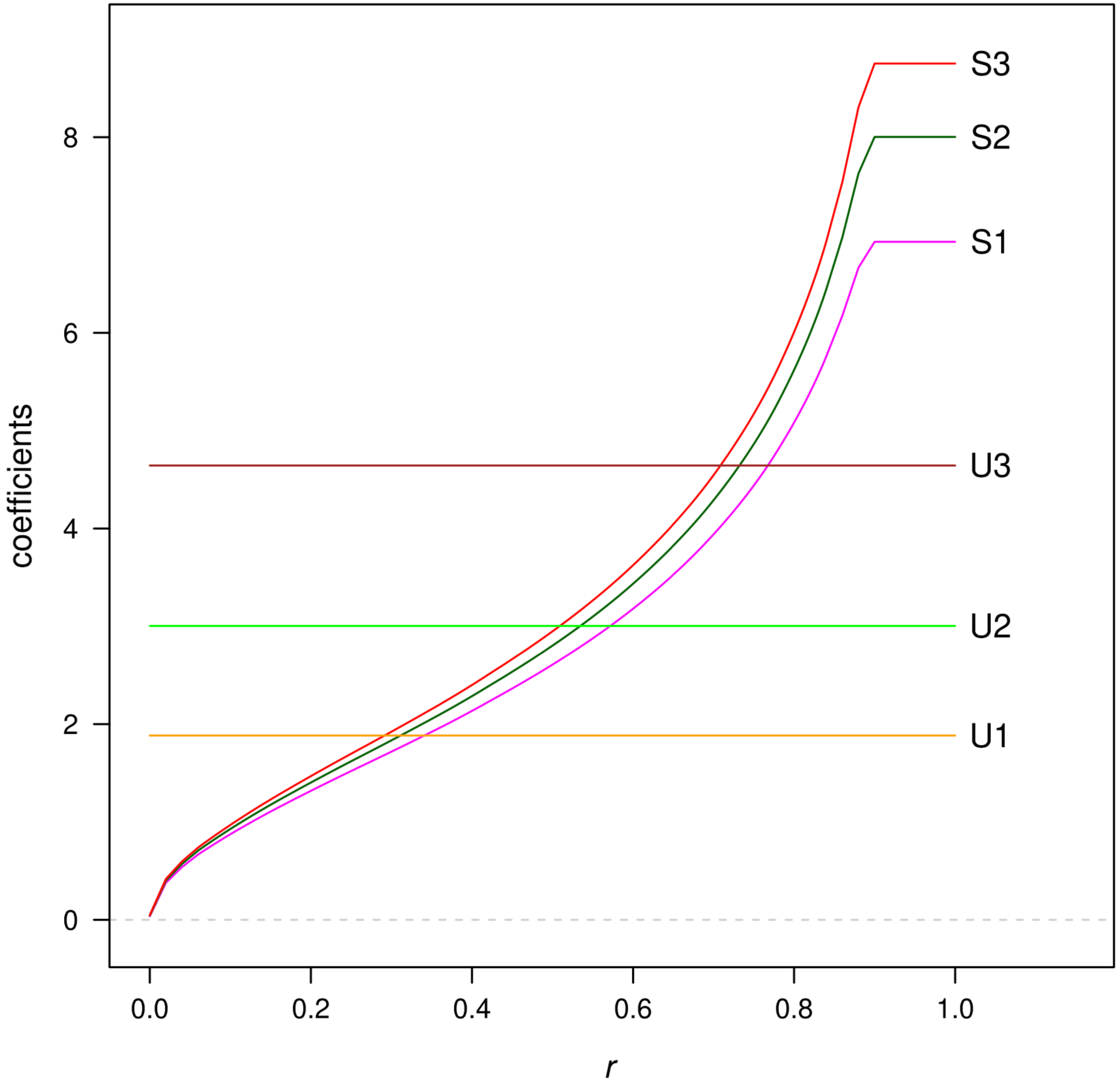
Properties

$$\hat{\mathbf{y}} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

- The problem is guaranteed to have a single solution

$$\frac{Var(\mathbf{S}(\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^T\mathbf{y})}{Var(\mathbf{S}(\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^T\mathbf{y}) + Var(\hat{\mathbf{U}}\beta)} = r$$

- Coefficients behave monotonically in r



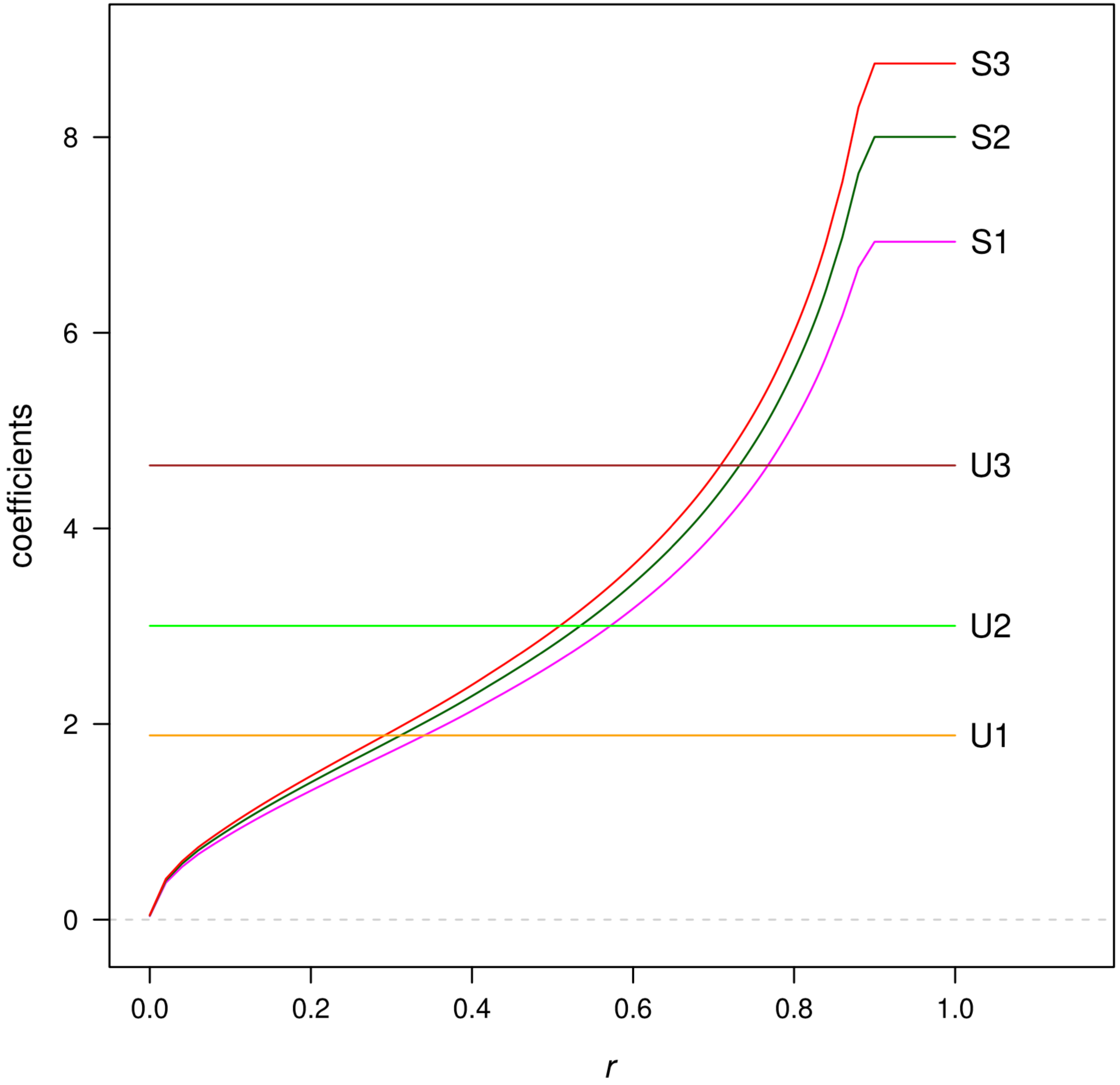
Properties

$$\hat{\mathbf{y}} = \mathbf{S}\alpha + \hat{\mathbf{U}}\beta + \epsilon$$

- The problem is guaranteed to have a single solution

$$\frac{Var(\mathbf{S}(\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^T\mathbf{y})}{Var(\mathbf{S}(\mathbf{S}^T\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^T\mathbf{y}) + Var(\hat{\mathbf{U}}\beta)} = r$$

- Coefficients behave monotonically in r
- Easier to optimise (than Komiyama)



Possible extensions

Different penalties

$$(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \widehat{\boldsymbol{\beta}}_{\text{FRRM}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{S}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1(r)\|\boldsymbol{\alpha}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2$$

- Improve accuracy and address collinearity
- Variable selection: LASSO or elastic net penalties

Possible extensions

Different penalties

- Improve accuracy and address collinearity
- Variable selection: LASSO or elastic net penalties

$$(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{S}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1(r)\|\boldsymbol{\alpha}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2$$

Different definitions of fairness

- Equality of odds: $\hat{\mathbf{y}}$ independent of $\mathbf{S}$, conditional of $\mathbf{y}$
- Individual fairness

$$R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi) = \frac{\mathrm{VAR}(\mathbf{S}\boldsymbol{\phi})}{\mathrm{VAR}(\mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi})}$$

$$\widehat{\mathbf{y}} = \mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi} + \varepsilon^*.$$

Possible extensions

Different penalties

- Improve accuracy and address collinearity
- Variable selection: LASSO or elastic net penalties

$$(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \widehat{\boldsymbol{\beta}}_{\text{FRRM}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{S}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1(r)\|\boldsymbol{\alpha}\|_2^2 + \lambda_2\|\boldsymbol{\beta}\|_2^2$$

Different definitions of fairness

- Equality of odds: $\hat{\mathbf{y}}$ independent of $\mathbf{S}$, conditional of $\mathbf{y}$
- Individual fairness

$$R^2_{\text{EO}}(\boldsymbol{\phi}, \psi) = \frac{\text{VAR}(\mathbf{S}\boldsymbol{\phi})}{\text{VAR}(\mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi})}$$

$$\widehat{\mathbf{y}} = \mathbf{y}\psi + \mathbf{S}\boldsymbol{\phi} + \varepsilon^*.$$

Different models

- Generalised linear models (GLM)
- Cox proportional hazard model
- Kernel ridge regression model

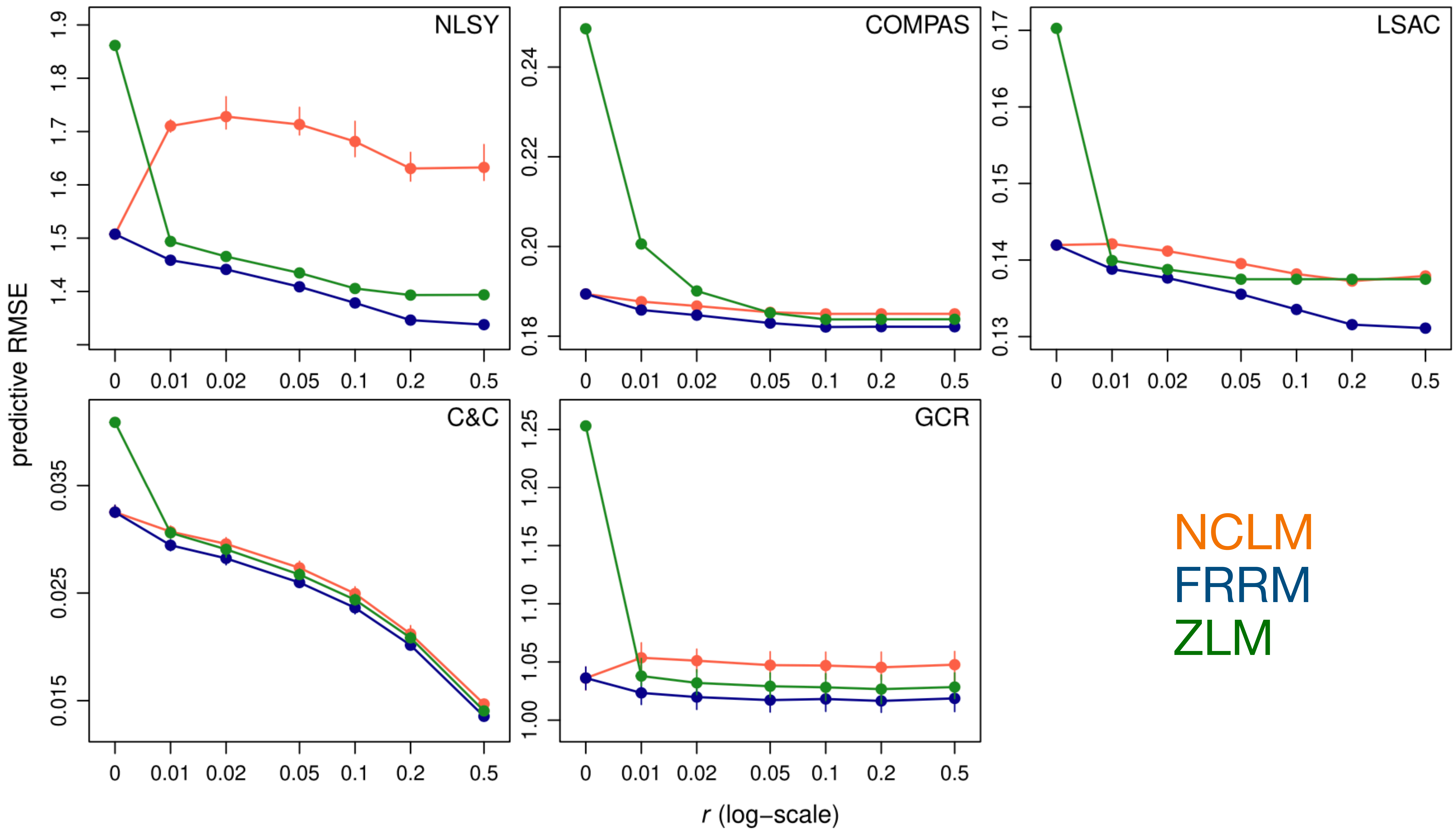
$$(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \widehat{\boldsymbol{\beta}}_{\text{FRRM}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}}\, D(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda(r)\|\boldsymbol{\alpha}\|_2^2$$

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

Real data experiments and comparisons

- Communities and Crime (810 observations, 101 socio-economics predictors)
  $\mathbf{y}$: normalised crime rate
  $\mathbf{S}$: proportions of African-American people and foreign born people
- COMPAS (5855 observations, 13 predictors)
  $\mathbf{y}$: % recidivating within 2 years
  $\mathbf{S}$: offender's gender and race
- National Longitudinal Survey of Youth (4908 observations, 13 labour market predictors)
  $\mathbf{y}$: income in 1990
  $\mathbf{S}$: gender and age
- Law School Admissions Council
  $\mathbf{y}$: GPA
  $\mathbf{S}$: race and age
- German Credit (1000 observations, 42 predictors)
  $\mathbf{y}$: % of good and bad loans
  $\mathbf{S}$: age, gender and foreign-born status

Real data experiments and comparisons



Zafar, Valera, Gomez-Rodriguez, Gummadi: Fairness constraints: a flexible approach for fair classification. J. Mach. Learn. Res. **20** (2019)

Summary

PROS

- Easy to understand
- Easy and fast to run (`fairml` R package)
- Works with different types of response variables
- Works with multivariate sensitive variables, of different type
- Works with different definitions of fairness

Summary

PROS

- Easy to understand
- Easy and fast to run (`fairml` R package)
- Works with different types of response variables
- Works with multivariate sensitive variables, of different type
- Works with different definitions of fairness

CONS

- Criticism agains use of $R^2_{\mathbf{S}}$
- You need to specify $\mathbf{S}$

Thank you very much!

Questions?