Statistics and Computing (2022) 32:77
https://doi.org/10.1007/s11222-022-10143-w

Achieving fairness with a simple ridge penalty

Marco Scutari[1] · **Francesca Panero**[2] · **Manuel Proissl**[3]

Received: 16 August 2021 / Accepted: 31 August 2022 / Published online: 18 September 2022

Abstract
In this paper, we present a general framework for estimating regression models subject to a user-defined level of fairness. We enforce fairness as a model selection step in which we choose the value of a ridge penalty to control the effect of sensitive attributes. We then estimate the parameters of the model conditional on the chosen penalty value. Our proposal is mathematically simple, with a solution that is partly in closed form and produces estimates of the regression coefficients that are intuitive to interpret as a function of the level of fairness. Furthermore, it is easily extended to generalised linear models, kernelised regression models and other penalties, and it can accommodate multiple definitions of fairness. We compare our approach with the regression model from Komiyama et al. (in: Proceedings of machine learning research. 35th international conference on machine learning (ICML), vol 80, pp 2737–2746, 2018), which implements a provably optimal linear regression model and with the fair models from Zafar et al. (J Mach Learn Res 20:1–42, 2019). We evaluate these approaches empirically on six different data sets, and we find that our proposal provides better goodness of fit and better predictive accuracy for the same level of fairness. In addition, we highlight a source of bias in the original experimental evaluation in Komiyama et al. (in: Proceedings of machine learning research. 35th international conference on machine learning (ICML), vol 80, pp 2737–2746, 2018).

Keywords Linear regression · Logistic regression · Generalised linear models · Fairness · Ridge regression

1 Introduction

Machine learning models are increasingly being used in applications where it is crucial to ensure the accountability and fairness of the decisions made on the basis of their outputs: some examples are criminal justice (Berk et al. 2021), credit risk modelling (Fuster et al. 2020) and screening job applications (Raghavan et al. 2020). In such cases, we are required to ensure that we are not discriminating individuals based on sensitive attributes such as gender and race, leading to disparate treatment of specific groups. At the same time, we would like to achieve the best possible predictive accuracy from other predictors.

The task of defining a non-discriminating treatment, though, does not come without challenges. The concept of fairness itself, in fact, has been characterised in different ways depending on the context. From an ethical and legal perspective, for example, it might depend on the type of distortion we wish to limit, which in turns varies with the type of application. Sometimes, we want to limit the adverse bias against a specific group, while in other instances we wish to protect single individuals. Alongside the legal and philosophical research debate, institutional regulations on the use of algorithms in society have been proposed in the last decade: for a comparison among the USA, EU and UK regulations, see Cath et al. (2018). The European Commission has recently released the first legal framework for the use of artificial intelligence (European Commission 2021), which is now under revision by the member states.

At the same time, there has been a growing interest towards fairness-preserving methods in the machine learning literature. From a statistical perspective, different charac-

✉ Marco Scutari
scutari@idsia.ch

Francesca Panero
f.panero@lse.ac.uk

Manuel Proissl
manuel.proissl@ubs.com

[1] Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Lugano, Switzerland

[2] London School of Economics; formerly Department of Statistics, University of Oxford, Oxford, UK

[3] UBS, Data and Analytics Center of Excellence, Zurich, Switzerland

terisations of fairness translate into different probabilistic models, which must then take into account the characteristics of the data they may be applied to (say, whether the variables are continuous or categorical, or whether a reliable ground truth is available or not). Despite this variety, all characterisations of fairness follow one of the two following approaches: *individual fairness* or *group fairness* (Del Barrio et al. 2020). The former requires that individuals that are similar receive similar predictions, while the latter aims at obtaining predictions that are similar across the groups identified by the sensitive variables.

Group fairness has been explored the most in the literature. When defined as *statistical* or *demographic parity*, it requires that predictions and the sensitive variables are independent. If $\mathbf{X}$ is a matrix of predictors, $\mathbf{S}$ is a matrix of sensitive attributes, $\mathbf{y}$ is the response variable and $\widehat{\mathbf{y}}$ are the predictions provided by some model, statistical parity translates into $\widehat{\mathbf{y}}$ being independent of $\mathbf{S}$. Usually the requirement of complete independence is too strong for practical applications, and it is relaxed into a constraint that limits the strength of the dependence between $\mathbf{S}$ and $\widehat{\mathbf{y}}$. Statistical parity is a good definition when a reliable ground truth is not available: otherwise, a perfect classifier on a data set which is unbalanced in the outcome across groups would possibly not satisfy the definition. While this is usually seen as a weakness of this fairness definition, we must remember that historical data often display some sort of bias and allowing the perfect classifier to score optimally in the chosen metric would mean to preserve it in future decisions. In light of this, definitions that do not rely on ground truth are known as "bias-transforming", while notions that condition on the truth are known as "bias-preserving" (Wachter et al. 2021). A common bias-preserving definition of fairness is *equality of opportunity*, which requires the predictions $\widehat{\mathbf{y}}$ to be independent from the sensitive attributes $\mathbf{S}$ after conditioning on the ground truth $\mathbf{y}$. In the case of a binary classifier and a single categorical sensitive variable, it is commonly known as *equality of odds* and translates into having the same false positive and negative rates across different groups.

Learning fair black-box machine learning models such as deep neural networks provides many hard challenges (see, for instance, Chorać et al. 2020) that are currently being investigated. For this reason, a large part of the literature focuses on simpler models. In many settings, such models are preferable because there are not enough data to train a deep neural network, because of computational limitations, or because they are more interpretable. Most such research focuses on classification models. For instance, Woodworth et al. (2017) investigated equality of odds for a binary sensitive attribute; Zafar et al. (2019) investigated the unfairness of the decision boundary in logistic regression and support-vector machines under statistical parity; Russell et al. (2017) explored counterfactual fairness in graphical models; Agarwal et al. (2018) used ensembles of logistic regressions and gradient-boosted trees to reduce fair classification to a sequence of cost-sensitive classification problems. For a review on the vast field of fairness notions and methods, see Mehrabi et al. (2021), Del Barrio et al. (2020) and Pessach and Shmueli (2020). In our work, we will focus on models that satisfy statistical parity.

Fair regression models (with a continuous response variable) have not been investigated in the literature as thoroughly as classifiers. Fukuchi et al. (2013) considered a generative model that is neutral to a finite set of viewpoints. Calders et al. (2013) focused on discrete sensitive attributes that may be used to cluster observations. Pérez-Suay et al. (2017) used kernels as a regulariser to enforce fairness while allowing nonlinear associations and dimensionality reduction. Agarwal et al. (2019) chose to bound the regression error within an allowable limit for each group defined by the sensitive attributes; Berk et al. (2017) achieved a similar effect using individual and group penalty terms. Chzhen et al. (2020) used model recalibration on a discretised transform of the response to leverage fairness characterisations used in classification models with a single binary sensitive attribute. Mary et al. (2019) used the notion of Rényi correlation to propose two methods that achieve statistical parity and equality of odds. Steinberg et al. (2020) implemented a similar idea with mutual information.

Komiyama et al. (2018) proposed a quadratic optimisation approach for fair linear regression models that constrains least squares estimation by bounding the relative proportion of the variance explained by the sensitive attributes, falling into the statistical parity framework. In contrast with the approaches mentioned above, both predictors and sensitive attributes are allowed to be continuous as well as discrete; any number of predictors and sensitive attributes can be included in the model, and the level of fairness can be controlled directly by the user, without the need of model calibration to estimate it empirically. In the following, we call this approach NCLM (as in *non-convex linear model*). This approach comes with theoretical optimality guarantees. However, it produces regression coefficient estimates that are not in closed form and whose behaviour is not easy to interpret with respect to the level of fairness, and it is difficult to extend it beyond linear regression models.

These limitations motivated us to propose a simpler *fair ridge regression model* (FRRM) which is easier to estimate, to interpret and to extend. At the same time, we wanted to match the key strengths that distinguish Komiyama et al. (2018) from earlier work:

1. The ability to model any combination of discrete and continuous predictors as well as sensitive attributes;
2. The ability to control fairness directly via a tuning parameter with an intuitive, real-world interpretation.

We achieve these aims by separating model selection and model estimation. Firstly, we choose the ridge penalty to achieve the desired level of fairness. Secondly, we estimate the model parameters conditional on the chosen penalty value. This is in contrast to other methods in the literature that do not have a separate model selection phase.

The paper is laid out as follows. In Sect. 2, we briefly review the NCLM approach from Komiyama et al. (2018), its formulation as an optimisation problem (Sect. 2.1), and we highlight a source of bias in its original experimental validation (Sect. 2.2). In Sect. 3 we discuss our proposal, FRRM, including its practical implementation (Sect. 3.1). We also discuss an analytical, closed-form estimate for ridge penalty and for the regression coefficients of the sensitive attributes in Sect. 3.2. We discuss several possible extensions of FRRM, including that to generalised linear models (FGRRM) and to different definitions of fairness in Sect. 4. In Sect. 5 we compare FRRM with NCLM and with the approach proposed by Zafar et al. (2019), investigating both linear (Sect. 5.1) and logistic (Sect. 5.2) regression models. We also consider the models from Steinberg et al. (2020) and Agarwal et al. (2018) in Sect. 5.3, insofar as they can be compared to FRRM given their limitations. Finally, we discuss the results in Sect. 6.

2 A nonconvex optimisation approach to fairness

Let $\mathbf{X}$ be a matrix of predictors, $\mathbf{S}$ be a matrix of sensitive attributes and $\mathbf{y}$ be a continuous response variable. Without loss of generality, we assume that all variables in $\mathbf{X}$, $\mathbf{S}$ and $\mathbf{y}$ are centred and that any categorical variables in $\mathbf{X}$ and $\mathbf{S}$ have been replaced with their one-hot encoding. Komiyama et al. (2018) start by removing the association between $\mathbf{X}$ and $\mathbf{S}$ using the auxiliary multivariate linear regression model

$$\mathbf{X} = \mathbf{B}^{\mathrm{T}}\mathbf{S} + \mathbf{U}.$$

They estimate the regression coefficients $\mathbf{B}$ by ordinary least squares (OLS) as $\widehat{\mathbf{B}}_{\mathrm{OLS}} = (\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{X}$, thus obtaining the residuals

$$\widehat{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{B}}_{\mathrm{OLS}}^{\mathrm{T}}\mathbf{S}. \tag{1}$$

Due to the properties of OLS, $\mathbf{S}$ and $\widehat{\mathbf{U}}$ are orthogonal and $\mathrm{COV}(\mathbf{S}, \widehat{\mathbf{U}}) = \mathbf{0}$, where $\mathbf{0}$ is a matrix of zeroes. $\widehat{\mathbf{B}}_{\mathrm{OLS}}^{\mathrm{T}}\mathbf{X}$ can then be interpreted as the component of $\mathbf{X}$ that is explained by $\mathbf{S}$, and $\widehat{\mathbf{U}}$ as the component of $\mathbf{X}$ that cannot be explained by $\mathbf{S}$ (the de-correlated predictors).

Komiyama et al. (2018) then define their main predictive model as the OLS regression

$$\mathbf{y} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\alpha}$ is the vector of the coefficients associated with the sensitive attributes $\mathbf{S}$, and $\boldsymbol{\beta}$ is associated with the de-correlated predictors $\widehat{\mathbf{U}}$. In keeping with classical statistics, they measure the goodness of fit of the model with the coefficient of determination $R^2(\boldsymbol{\alpha}, \boldsymbol{\beta})$, which can be interpreted as the proportion of variance explained by the model. Furthermore, $R^2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is also proportional to the cross-entropy of the model, which is a function of $1 - R^2(\boldsymbol{\alpha}, \boldsymbol{\beta})$. Since $\mathbf{S}$ and $\widehat{\mathbf{U}}$ are orthogonal, and since (2) is fitted with OLS, $R^2(\boldsymbol{\alpha}, \boldsymbol{\beta})$ decomposes as

$$\begin{aligned} R^2(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{\mathrm{VAR}(\widehat{\mathbf{y}})}{\mathrm{VAR}(\mathbf{y})} = \frac{\mathrm{VAR}(\mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta})}{\mathrm{VAR}(\mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta} + \boldsymbol{\varepsilon})} \\ &= \frac{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\boldsymbol{\beta}}{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\boldsymbol{\beta} + \mathrm{VAR}(\boldsymbol{\varepsilon})}, \end{aligned} \tag{3}$$

where $\widehat{\mathbf{y}}$ are the fitted values produced by OLS, and $\mathrm{VAR}(\widehat{\mathbf{U}})$, $\mathrm{VAR}(\mathbf{S})$ are the covariance matrices of $\mathbf{X}$ and $\mathbf{S}$, respectively. Both matrices are assumed to be full rank. The proportion of the overall explained variance that is attributable to the sensitive attributes then is

$$R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{\mathrm{VAR}(\mathbf{S}\boldsymbol{\alpha})}{\mathrm{VAR}(\widehat{\mathbf{y}})} = \frac{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\boldsymbol{\beta}}. \tag{4}$$

Komiyama et al. (2018) choose to bound $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ to a value $r \in [0, 1]$ that determines how fair the model is. Setting $r = 0$ corresponds to a completely fair model (that is, statistical parity: $\widehat{\mathbf{y}}$ is independent from $\mathbf{S}$) because it implies $\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} = 0$, which can only be true if all regression coefficients $\boldsymbol{\alpha}$ are equal to zero. On the other hand, setting $r = 1$ means the constraint is always satisfied because $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant 1$ by construction.

2.1 The optimisation problem

Fitting (2) subject to $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r$ by OLS can be formally be written as

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \mathrm{E}\left((\mathbf{y} - \widehat{\mathbf{y}})^2\right) \\ \text{s.t.} \quad & R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r \end{aligned} \tag{5}$$

which in light of (3) and (4) takes the form

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} + \boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\boldsymbol{\beta} \\ & - 2\left(\mathrm{E}(\mathbf{y}\mathbf{S}^{\mathrm{T}}\boldsymbol{\alpha}) + \mathrm{E}(\mathbf{y}\widehat{\mathbf{U}}^{\mathrm{T}}\boldsymbol{\beta})\right) \\ \text{s.t.} \quad & (1 - r)\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} - r\boldsymbol{\beta}^{\mathrm{T}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\boldsymbol{\beta} \leqslant 0. \end{aligned} \tag{6}$$

The optimisation in (6) is a quadratic programming problem subject to quadratic constraints, and it is not convex in $(\boldsymbol{\alpha}, \boldsymbol{\beta})$.

Instead of solving (6), Komiyama et al. (2018) convert it into the convex quadratic problem

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta},\gamma} \quad & \gamma \\ \text{s.t.} \quad & \begin{bmatrix}\boldsymbol{\alpha}^{\mathrm{T}} & \boldsymbol{\beta}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathrm{VAR}(\mathbf{S}) & \mathbf{0} \\ \mathbf{0} & \mathrm{VAR}(\widehat{\mathbf{U}})\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \\ & -2\begin{bmatrix}\mathrm{E}(\mathbf{yS}) & \mathrm{E}(\mathbf{y}\widehat{\mathbf{U}})\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - \gamma \leqslant 0, \\ & \begin{bmatrix}\boldsymbol{\alpha}^{\mathrm{T}} & \boldsymbol{\beta}^{\mathrm{T}}\end{bmatrix}\begin{bmatrix}\mathrm{VAR}(\mathbf{S})/r & \mathbf{0} \\ \mathbf{0} & \mathbf{0}\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} \\ & -2\begin{bmatrix}\mathrm{E}(\mathbf{yS}) & \mathrm{E}(\mathbf{y}\widehat{\mathbf{U}})\end{bmatrix}\begin{bmatrix}\boldsymbol{\alpha} \\ \boldsymbol{\beta}\end{bmatrix} - \gamma \leqslant 0, \end{aligned} \tag{7}$$

using previous results from Yamada and Takeda (2018). (γ is an auxiliary parameter without a real-world interpretation.) The relaxed problem in (7) yields an optimal solution $(\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}})$ for (6), under the assumptions discussed earlier as well as those in Yamada and Takeda (2018). It also has two additional favourable properties: it can be solved by off-the-shelf optimisers (the authors used Gurobi) and it can be extended by replacing $\mathrm{VAR}(\mathbf{S})$ and $\mathrm{VAR}(\widehat{\mathbf{U}})$ with more complex estimators than the respective empirical covariance matrices. Two examples covered in Komiyama et al. (2018) are the use of kernel transforms to capture nonlinear relationship and regularisation adding a ridge penalty term.

2.2 Avoiding bias in the auxiliary model

A key assumption that underlies the results in the previous section is the use of OLS in creating the de-correlated predictors in (1): it ensures that $\widehat{\mathbf{U}}$ is orthogonal to $\mathbf{S}$ and therefore that is does not contain any information from the sensitive attributes. However, Komiyama et al. (2018) in their experimental section state that "The features $\widehat{\mathbf{U}}$ were built from $\mathbf{X}$ by de-correlating it from $\mathbf{S}$ by using regularised least squares regression",[1] where the "regularised least squares regression" is a ridge regression model. This divergence from the theoretical construction leading to (7) introduces bias in the model by making $\widehat{\mathbf{U}}$ correlated to $\mathbf{S}$ in proportion to the amount of regularisation.

As noted in van Wieringen (2018), the residuals in a ridge regression are not orthogonal to the fitted values for any penalisation coefficient $\lambda > 0$. Let $\widetilde{\mathbf{U}}$ be the ridge estimate of $\mathbf{U}$, that is, $\widetilde{\mathbf{U}} = \mathbf{X} - \widehat{\mathbf{B}}_\lambda \mathbf{X}$. Let X_i be the ith column of $\mathbf{X}$ (that is, one of the predictors) and $\widetilde{U}_i$ be the corresponding column of $\widetilde{\mathbf{U}}$. Then

$$\widetilde{U}_i = X_i - \mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^{\mathrm{T}}X_i$$

while the corresponding OLS estimate from $\widehat{\mathbf{U}}$ is

$$\widehat{U}_i = X_i - \mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}X_i.$$

Their difference is

$$\begin{aligned} \widetilde{U}_i - \widehat{U}_i &= \mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1}\mathbf{S}^{\mathrm{T}}X_i - \mathbf{S}(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{I})^{-1}\mathbf{S}^{\mathrm{T}}X_i \\ &= \mathbf{S}\left[(\mathbf{S}^{\mathrm{T}}\mathbf{S})^{-1} - (\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{I})^{-1}\right]\mathbf{S}^{\mathrm{T}}X_i. \end{aligned} \tag{8}$$

Given the spectral decomposition $\mathbf{S}^{\mathrm{T}}\mathbf{S} = \mathbf{A}\mathbf{3}\mathbf{A}^{\mathrm{T}}$, where $\mathbf{3} = \mathrm{diag}(l_j)$, (8) can be rewritten as

$$\begin{aligned} \widetilde{U}_i - \widehat{U}_i &= \mathbf{S}\left[\mathbf{A}\mathbf{3}^{-1}\mathbf{A}^{\mathrm{T}} - (\mathbf{A}\mathbf{3}\mathbf{A}^{\mathrm{T}} + \lambda\mathbf{I})^{-1}\right]\mathbf{S}^{\mathrm{T}}X_i \\ &= \mathbf{S}\left[\mathbf{A}\,\mathrm{diag}\left(\frac{1}{l_j} - \frac{1}{l_j+\lambda}\right)\mathbf{A}^{\mathrm{T}}\right]\mathbf{S}^{\mathrm{T}}X_i, \end{aligned}$$

thus giving

$$\begin{aligned} \mathbf{S}^{\mathrm{T}}(\widetilde{U}_i - \widehat{U}_i) &= \mathbf{S}^{\mathrm{T}}\mathbf{S}\left[\mathbf{A}\,\mathrm{diag}\left(\frac{1}{l_j} - \frac{1}{l_j+\lambda}\right)\mathbf{A}^{\mathrm{T}}\right]\mathbf{S}^{\mathrm{T}}X_i \\ &= \mathbf{A}\mathbf{3}\mathbf{A}^{\mathrm{T}}\mathbf{A}\,\mathrm{diag}\left(\frac{1}{l_j} - \frac{1}{l_j+\lambda}\right)\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}X_i \\ &= \mathbf{A}\,\mathrm{diag}\left(1 - \frac{l_j}{l_j+\lambda}\right)\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}X_i. \end{aligned}$$

Since $\mathbf{S}^{\mathrm{T}}\widetilde{U}_i = \mathbf{S}^{\mathrm{T}}(\widetilde{U}_i - \widehat{U}_i)$ due to $\mathbf{S}^{\mathrm{T}}\widehat{U}_i = \mathbf{0}$, and $\mathrm{COV}(\mathbf{S}, \widetilde{U}_i) \propto \mathbf{S}^{\mathrm{T}}\widehat{U}_i$, we have that $\mathrm{COV}(\mathbf{S}, \widetilde{U}_i)$ vanishes as $\lambda \to 0$ and $\widehat{\mathbf{B}}_\lambda \to \widehat{\mathbf{B}}_{\mathrm{OLS}}$. On the other hand, $|\mathrm{COV}(\mathbf{S}, \widetilde{U}_i)|$ becomes increasingly large as $\lambda \to \infty$, eventually reaching $\mathrm{COV}(\mathbf{S}, X_i)$. If we replace $\widehat{\mathbf{U}}$ with $\widetilde{\mathbf{U}}$, the denominator of $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ then becomes

$$\begin{aligned} \mathrm{VAR}(\widetilde{\mathbf{y}}) &= \mathrm{VAR}(\mathbf{S}\boldsymbol{\alpha} + \widetilde{\mathbf{U}}\boldsymbol{\beta}) \\ &= \mathrm{VAR}(\mathbf{S}\boldsymbol{\alpha}) + \mathrm{VAR}(\widetilde{\mathbf{U}}\boldsymbol{\beta}) - 2\,\mathrm{COV}(\mathbf{S}\boldsymbol{\alpha}, \widetilde{\mathbf{U}}\boldsymbol{\beta}) \end{aligned}$$

which can be either larger or smaller than that of the $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in (4).

Example 1 Consider three predictors $\mathbf{X} = \{X_1, X_2, X_3\}$ and three sensitive attributes $\mathbf{S} = \{S_1, S_2, S_3\}$ distributed as

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

[1] We substituted their notation with ours for clarity.

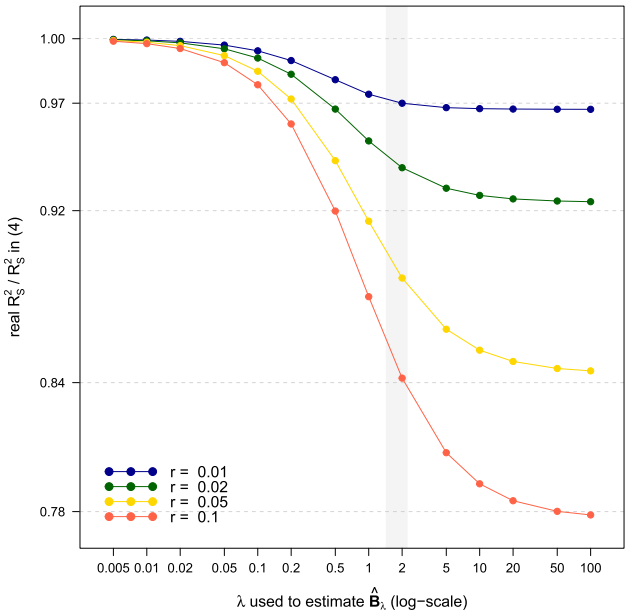

Fig. 1 Bias introduced by penalised regression in $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$. The shaded area corresponds to the range of optimal values for λ used in computing $\widehat{\mathbf{B}}_\lambda$

giving the response variable $\mathbf{y}$ by way of the linear model

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

with with independent and identically distributed errors $\varepsilon_i \sim N(0, 100)$. Hence $\boldsymbol{\beta} = [2, 3, 4]^{\mathrm{T}}$ and $\boldsymbol{\alpha} = [5, 6, 7]^{\mathrm{T}}$ using the notation established in (2).

Figure 1 shows the ratio between

$$\frac{\mathrm{VAR}(\mathbf{S}\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}})}{\mathrm{VAR}(\mathbf{S}\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}) + \mathrm{VAR}(\widetilde{\mathbf{U}}\widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}}) - 2\,\mathrm{COV}(\mathbf{S}\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}, \widetilde{\mathbf{U}}\widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}})}$$

and

$$\frac{\mathrm{VAR}(\mathbf{S}\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}})}{\mathrm{VAR}(\mathbf{S}\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}) + \mathrm{VAR}(\widetilde{\mathbf{U}}\widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}})}$$

for various values of the penalisation coefficient λ and $r = 0.01, 0.02, 0.05, 0.10$. The shaded area represents the range of the optimal λs for the various $\widetilde{U}_i$, chosen as those within 1 standard error from the minimum in 10-fold cross-validation as suggested in Hastie et al. (2009). The relative difference between the two is between 3 and 5% for small values like $r = 0.01, 0.02$, and it can grow as large as 16% for $r = 0.10$. Note that variables are only weakly correlated in this example; higher degrees of collinearity will result in even stronger bias.

Using ridge regression to estimate (1) can be motivated by the need to address collinearity in $\mathbf{S}$, which would make $\widehat{\mathbf{B}}_{\mathrm{OLS}}$ numerically unstable or impossible to estimate. As an alternative, we can replace $\mathbf{S}$ with a lower-dimensional, full-rank approximation based on a reduced number of principal components in both (1) and (7). This satisfies the assumption that $\mathrm{VAR}(\mathbf{S})$ is full rank in the process of deriving (7).

3 An alternative penalised regression approach

The approach proposed by Komiyama et al. (2018) has four limitations that motivated our work.

1. The dimension of the optimisation problem that is solved numerically in (7) scales linearly in the number of variables.
2. The formulation of (7) allows us to use numeric solvers that can handle quadratic programming with quadratic constraints, but cannot be translated to regression models other than a linear regression.
3. The second constraint in (7) is undefined in the limit case $r = 0$ and can potentially make $(\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}})$ numerically unstable as $r \to 0$.
4. The behaviour of the estimated regression coefficients is not intuitive to interpret. The constraints in (7) are functions of both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$: as a result, $\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}$ and $\widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}}$ are not independent as they would be in an unconstrained OLS regression (because $\mathbf{S}$ and $\widehat{\mathbf{U}}$ are orthogonal). Changing the value of the bound r then affects the coefficients $\widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}}$, shrinking or inflating them, as well as the $\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}$.

Example 1 (continued) Consider again the example from Sect. 2.2. The estimates of the regression coefficients given by NCLM over $r \in [0, 1]$ are shown in the profile plot in Fig. 2. As expected, we can see that we have all $\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}$ converge to zero as $r \to 0$ because $\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}_{\mathrm{NCLM}}\,\mathrm{VAR}(\mathbf{S})\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}} \to 0$. For $r = 0$, we can say that $\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}} = \mathbf{0}$ for continuity. As r increases, all $\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}$ gradually increase in magnitude. The constraint becomes inactive when $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) < r$, hence NCLM reverts back to a standard OLS regression model for large values of r. As a result, all coefficients stop changing once $r \geqslant 0.85$.

The behaviour of the $\widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}}$ is, however, difficult to explain on an intuitive level. They do not change monotonically as r increases, unlike what happens, for instance, in ridge regression (Hoerl and Kennard 1970) or the LASSO (Tibshirani 1996). They first increase above $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$, which helps in reducing $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}})$ by increasing its denominator. They then plateau around $r \approx 0.3$, and start decreasing as $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{NCLM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{NCLM}})$ is allowed to grow.

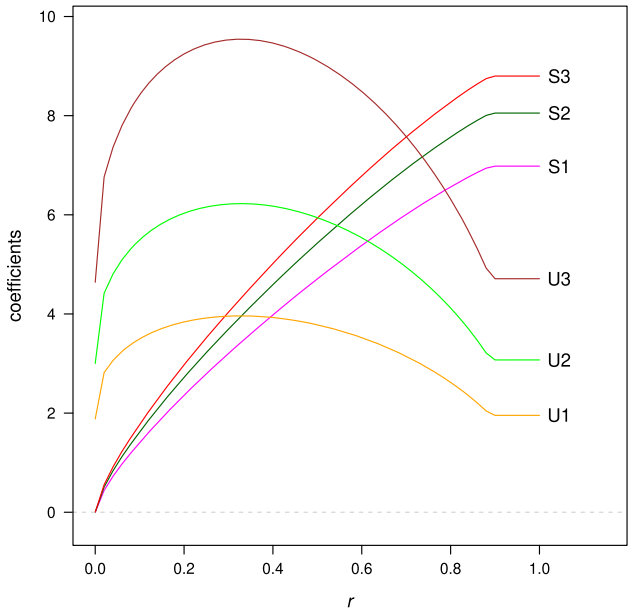

Fig. 2 Profile plot for the coefficients estimated by NLCM as a function of the bound r for the example in Sect. 3

3.1 Fairness by ridge penalty

In order to overcome the issues of NCLM we just discussed, we propose an alternative constrained optimisation framework we call the *fair ridge regression model* (FRRM). The key idea behind our proposal is to solve the constrained optimisation problem stated in (5) by imposing a ridge penalty on $\boldsymbol{\alpha}$ while leaving $\boldsymbol{\beta}$ unconstrained. Formally,

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \quad & \mathrm{E}\left((\mathbf{y}-\widehat{\mathbf{y}})^2\right) \\ \text{s.t.} \quad & \|\boldsymbol{\alpha}\|_2^2 \leqslant t(r) \end{aligned} \tag{9}$$

where $t(r) \geqslant 0$ is such that $R^2_{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leqslant r$ by bounding $\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha}$ through $\|\boldsymbol{\alpha}\|_2^2$. Equivalently, we can write (9) as

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

where $\lambda(r) \geqslant 0$ is the value of the ridge penalty that makes $R^2_{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta}) \leqslant r$. There is a one-to-one relationship between the values of $t(r)$ and $\lambda(r)$, so we choose to focus on the latter. As $r \to 0$, $\lambda(r)$ should diverge so that all $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ converge to zero asymptotically and $\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\mathbf{S})\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} \to 0$ as in NCLM. Note that zero is a valid value for r in (9) while it is not for NCLM in (7). Furthermore, note that (9) is not specifically tied to $R^2_{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta})$ (we will show how to replace it with different fairness constraints in Sect. 4.3) and that it can be easily reformulated with other penalties (which we will discuss in Sect. 4.2).

The $\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$ are now independent from the $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ because the ridge penalty does not involve the former. Starting from the classical estimator for the coefficients of a ridge regression (as it can be found in van Wieringen (2018) among others), and taking into account that $\mathbf{S}$ and $\widehat{\mathbf{U}}$ are orthogonal, it is easy to show that

$$\begin{aligned} \begin{bmatrix} \widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} \\ \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} \end{bmatrix} &= \left(\begin{bmatrix} \mathbf{S}^{\mathrm{T}} \\ \widehat{\mathbf{U}}^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \mathbf{S} & \widehat{\mathbf{U}} \end{bmatrix} + \begin{bmatrix} \lambda(r)\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right)^{-1} \begin{bmatrix} \mathbf{S}^{\mathrm{T}} \\ \widehat{\mathbf{U}}^{\mathrm{T}} \end{bmatrix} \mathbf{y} \\ &= \begin{bmatrix} \mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda(r)\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{S}^{\mathrm{T}} \\ \widehat{\mathbf{U}}^{\mathrm{T}} \end{bmatrix} \mathbf{y} \\ &= \begin{bmatrix} \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda(r)\mathbf{I}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{y} \\ (\widehat{\mathbf{U}}^{\mathrm{T}}\widehat{\mathbf{U}})^{-1}\widehat{\mathbf{U}}^{\mathrm{T}}\mathbf{y} \end{bmatrix}. \end{aligned} \tag{11}$$

The $\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$ can be estimated in closed form, only depend on $\widehat{\mathbf{U}}$, and do not change as r varies. The $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ depend on $\mathbf{S}$ and also on r through $\lambda(r)$, and they must be estimated numerically. However, the form of $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ in (11) makes it possible to reduce the dimensionality and the complexity of the numeric optimisation compared to NCLM. We can estimate them as follows:

1. Apply (1) to $\mathbf{S}$, $\mathbf{X}$ to obtain $\mathbf{S}$, $\widehat{\mathbf{U}}$.
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:
   (a) If $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) \leqslant r$, set $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} = \widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}$.
   (b) Otherwise, find the value of $\lambda(r)$ that satisfies

$$\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} = \frac{r}{1-r}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} \tag{12}$$

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

As far as determining the value of $\lambda(r)$ that results in $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) \leqslant r$ is concerned, we can treat $\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$ as a constant that can be pre-computed from $\mathbf{U}$ independently of $\mathbf{S}$ and r. Furthermore, $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ is available as a closed-form function of r through $\lambda(r)$, and we know that $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\mathbf{S})\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}_{\mathrm{FRRM}} \to 0$ monotonically as $\lambda(r) \to \infty$ from the fundamental properties of ridge regression. As a result, (9) is guaranteed to have a single solution in $\lambda(r)$ which can be found with a simple, univariate root-finding algorithm. Selecting $\lambda(r)$ can be though of as model selection, since $\lambda(r)$ is a tuning parameter that affects the distribution of the $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$. Estimating the $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ given $\lambda(r)$ is then a separate model selection phase.

In the particular case that $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) \leqslant r$, a trivial solution to (12) is to set $\lambda(r) = 0$ and thus $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} = \widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}$: if the constraint is inactive because the bound we set is larger than the proportion of the overall variance that is attributable to the sensitive attributes, then the OLS estimate of $\boldsymbol{\beta}$ minimises the objective. This agrees with the behaviour

of NCLM shown in Fig. 2. On the other hand, if the constraint is active the objective is minimised when $\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha}$ takes is largest admissible value, which implies $R^2_{\mathbf{S}}(\boldsymbol{\alpha},\boldsymbol{\beta}) = r$. Rewriting (4) as an equality and moving all known terms to the right hand gives (12).

In the general case, the ridge penalty parameter is defined on $\mathbb{R}^+$. However, in (12) we can bound it above and below using the equality. From Lipovetsky (2006), we have that in a ridge regression with parameter λ

$$\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha} = \mathbf{y}^{\mathrm{T}}\mathbf{S}\mathbf{A}\,\mathrm{diag}\left(\frac{l_i + 2\lambda(r)}{(l_i + \lambda(r))^2}\right)\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{y}$$

where $\mathbf{A}\mathbf{3}\mathbf{A}^{\mathrm{T}} = \mathbf{A}\,\mathrm{diag}(l_i)\mathbf{A}^{\mathrm{T}}$ is again the eigenvalue decomposition of $\mathrm{VAR}(\mathbf{S})$. If we replace all the l_i on the right-hand side with the smallest (respectively, the largest) eigenvalue, we can bound $\boldsymbol{\alpha}^{\mathrm{T}}\,\mathrm{VAR}(\mathbf{S})\boldsymbol{\alpha}$ in

$$\left[\frac{l_{\min} + 2\lambda(r)}{(l_{\min} + \lambda(r))^2}\,d,\ \frac{l_{\max} + 2\lambda(r)}{(l_{\max} + \lambda(r))^2}\,d\right]$$

where $d = \mathbf{y}^{\mathrm{T}}\mathbf{S}\mathbf{A}\mathbf{A}^{\mathrm{T}}\mathbf{S}^{\mathrm{T}}\mathbf{y}$. We can then replace the bounds above and solve (12) as an equality in $\lambda(r)$ to obtain upper and lower bounds for the ridge penalty parameter. If we let $c = \frac{r}{1-r}\widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$, the resulting equations are

$$\frac{l_{\min} + 2\lambda(r)}{(l_{\min} + \lambda(r))^2}\,c = d \qquad \text{and} \qquad \frac{l_{\max} + 2\lambda(r)}{(l_{\max} + \lambda(r))^2}\,c = d$$

which are quadratic equations with one positive solution each. (Clearly, the respective negative solutions are not admissible since $\lambda(r) \geqslant 0$.)

Example 1 (continued) Consider one more time our example: the regression coefficients $(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}})$ are shown in Fig. 3. The regression coefficients $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ for S_1, S_2 and S_3 still converge to zero as $r \to 0$, ensuring that $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\mathbf{S})\widehat{\boldsymbol{\alpha}}^{\mathrm{T}}_{\mathrm{FRRM}} \to 0$, as in Fig. 2. However, the coefficients $\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$ for X_1, X_2 and X_3 do not change as r changes: they are equal to their OLS estimates for all values of r as implied by (11).

3.2 Analytical solution for independent S

In some instances, it is possible to solve (10) exactly and in closed form instead of relying on numerical optimisation. This allows us to better explore the behaviour of $\lambda(r)$ and $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$, as well as the effect of relaxing the constraint of statistical parity.

Assume that $\mathbf{S}$ is a $q \times n$ matrix of sensitive attributes which are mutually independent, that is, each S_j is independent of S_i for all $i, j = 1, \ldots, q$, $i \neq j$. Furthermore, assume that each S_j is scaled to $\mathrm{VAR}(S_j) = 1$ in addition to being centred. Let $\mathbf{X}$, instead, be a $p \times n$ matrix of predictors which are allowed

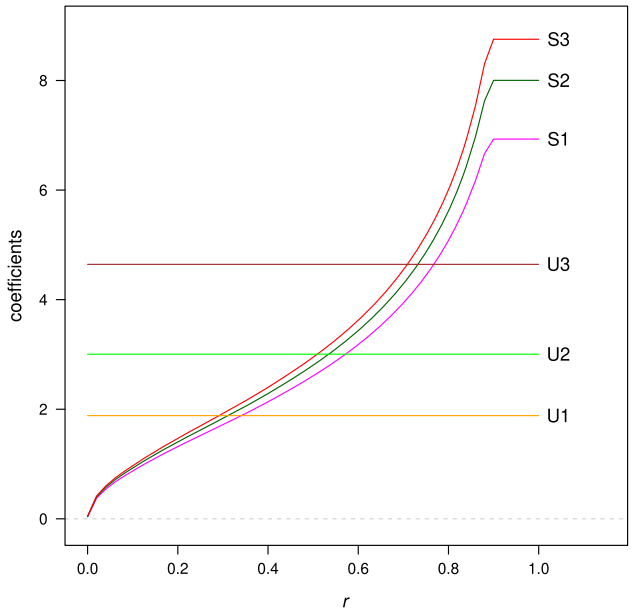

Fig. 3 Profile plot for the coefficients estimated by FRRM as a function of the bound r for the example in Sect. 3

to be correlated. Then, it is possible to write the solution of (12) with respect to λ in closed form.

Let c be defined as $c^2 := \widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{\mathrm{FRRM}}\,\mathrm{VAR}(\widehat{\mathbf{U}})\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$ (note that this is slightly different from the previous definition of c). Then (12) becomes

$$\left(\frac{1}{n+\lambda}\mathbf{I}_q\mathbf{S}^{\mathrm{T}}\mathbf{y}\right)^{\mathrm{T}}\mathbf{I}_q\left(\frac{1}{n+\lambda}\mathbf{I}_q\mathbf{S}^{\mathrm{T}}\mathbf{y}\right) = \frac{r}{1-r}c^2.$$

using the fact that $\mathbf{S}^{\mathrm{T}}\mathbf{S} = n\mathbf{I}_q$, where $\mathbf{I}_q$ is the identity matrix of size q. Solving the matrix products we get

$$\sum_{j=1}^{q}\left(S_j^{\mathrm{T}}\mathbf{y}\right)^2 - \frac{r}{1-r}c^2(n+\lambda)^2 = 0,$$

which has solution

$$\lambda(r) = -n + \frac{\|\mathbf{S}^{\mathrm{T}}\mathbf{y}\|_2^2}{c\sqrt{r/(1-r)}} \tag{13}$$

where $\|\mathbf{S}^{\mathrm{T}}\mathbf{y}\|_2^2 = \sum_{j=1}^{q}(S_j^{\mathrm{T}}\mathbf{y})^2$ is the squared Euclidean norm in $\mathbb{R}^n$. Plugging (13) into (12) gives

$$\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} = c\sqrt{\frac{r}{1-r}}\,\frac{\mathbf{S}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{S}^{\mathrm{T}}\mathbf{y}\|_2^2}. \tag{14}$$

From (14), we see that the $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ increase in magnitude (that is, move away from zero) as a function of r.

As we noted in Sect. 3.1, $r = 0$ satisfies statistical parity exactly: $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} = 0$ and therefore $\widehat{\mathbf{y}}$ is independent of $\mathbf{S}$. As r

increases, $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$ grows and so does the correlation between $\widehat{\mathbf{y}}$ and $\mathbf{S}$:

$$\begin{aligned}\text{COR}(\widehat{\mathbf{y}}, \mathbf{S}) &= \frac{\text{COV}(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}\mathbf{S} + \widehat{\boldsymbol{\beta}}_{\text{FRRM}}\widehat{\mathbf{U}}, \mathbf{S})}{\sqrt{\text{VAR}(\widehat{\mathbf{y}})\mathbf{I}_q}} \\ &= \frac{\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}}{\sqrt{\widehat{\boldsymbol{\alpha}}^2_{\text{FRRM}}\,\text{VAR}(\mathbf{S}) + \widehat{\boldsymbol{\beta}}^2_{\text{FRRM}}\,\text{VAR}(\widehat{\mathbf{U}})}},\end{aligned}$$

which is again proportional to $\sqrt{r/(1-r)}$ and equal to 0 when $r = 0$.

If the sensitive attributes are not mutually independent, solving (12) exactly is not possible in general and we revert to the root-finding algorithm described in Sect. 3. Even with just two sensitive attributes $\mathbf{S} = [S_1, S_2]$, the right hand side of (12) becomes:

$$\begin{aligned}&\begin{bmatrix}A\\B\end{bmatrix}^{\text{T}}\begin{bmatrix}n+\lambda & C\\ C & n+\lambda\end{bmatrix}^{-1}\begin{bmatrix}1 & \frac{C}{n}\\ \frac{C}{n} & 1\end{bmatrix}\begin{bmatrix}n+\lambda & C\\ C & n+\lambda\end{bmatrix}^{-1}\begin{bmatrix}A\\B\end{bmatrix} \\ &= \frac{(A(n+\lambda)-BC)^2 + (B(n+\lambda)-AC)^2}{((n+\lambda)^2-C^2)^2} \\ &\quad + \frac{2C/n(A(n+\lambda)-BC)(B(n+\lambda)-AC)}{((n+\lambda)^2-C^2)^2} \\ &= \frac{(n+\lambda)^2\left(A^2+B^2+2ABC/n\right)}{((n+\lambda)^2-C^2)^2} \\ &\quad - \frac{2C(n+\lambda)\left(2AB+CA^2/n+CB^2/n\right)}{((n+\lambda)^2-C^2)^2} \\ &\quad + \frac{C^2\left(A^2+B^2+2ABC/n\right)}{((n+\lambda)^2-C^2)^2}.\end{aligned}$$

where $A = S_1^{\text{T}}\mathbf{y}$, $B = S_2^{\text{T}}\mathbf{y}$, $C = S_1^{\text{T}}S_2 = S_2^{\text{T}}S_1$ for brevity. Equating the expression above to $c^2 r/(1-r)$ will give a 4th-degree polynomial in λ. The solutions to this equation can be computed exactly, but the resulting expression for λ does not provide any immediate insights.

4 Possible extensions

FRRM has a simple and modular construction that can accommodate a wide range of extensions: some examples are modelling nonlinear relationships, incorporating different and more complex penalties, using different definition of fairness and handling different types of responses with generalised linear models. The separation between model selection (the choice of $\lambda(r)$) and model estimation (estimating $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$ and $\widehat{\boldsymbol{\beta}}_{\text{FRRM}}$) makes it possible to change how either or both are performing drawing extensively from established statistical literature.

4.1 Nonlinear regression models

We can incorporate kernels into FRRM by fitting the model in the transformed feature spaces $Z_{\mathbf{S}}(\mathbf{S})$ and $Z_{\widehat{\mathbf{U}}}(\widehat{\mathbf{U}})$ produced by some positive kernel function, as in Komiyama et al. (2018). Combining the kernel trick with a ridge penalty produces a kernel ridge regression model (Saunders et al. 1998), which can be estimated efficiently following Zhang et al. (2015). Furthermore, this approach suggests further extensions to Gaussian process regressions, since the two models are closely related as discussed in Kanagawa et al. (2018).

4.2 Different penalties

We may also want to regularise the $\boldsymbol{\beta}$ coefficients to improve predictive accuracy and to address any collinearity present in the data. One option is to add a ridge penalty to the $\boldsymbol{\beta}$ in addition to that on the $\boldsymbol{\alpha}$. Ideally, without making $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$ and $\widehat{\boldsymbol{\beta}}_{\text{FRRM}}$ dependent to preserve the intuitive behaviour of the regression coefficient estimates produced by FRRM. A simple way is to add a second penalty term to (10),

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

resulting in

$$\begin{bmatrix}\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}\\ \widehat{\boldsymbol{\beta}}_{\text{FRRM}}\end{bmatrix} = \begin{bmatrix}\left(\mathbf{S}^{\text{T}}\mathbf{S} + \lambda_1(r)\mathbf{I}\right)^{-1}\mathbf{S}^{\text{T}}\mathbf{y}\\ \left(\widehat{\mathbf{U}}^{\text{T}}\widehat{\mathbf{U}} + \lambda_2\mathbf{I}\right)^{-1}\widehat{\mathbf{U}}^{\text{T}}\mathbf{y}\end{bmatrix}. \tag{15}$$

This is sufficient to ensure there are no unaddressed collinearities as $\mathbf{S}$ and $\widehat{\mathbf{U}}$ are orthogonal by construction.

Example 1 (continued) Figure 4 shows the estimates $(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \widehat{\boldsymbol{\beta}}_{\text{FRRM}})$ obtained with $\lambda_2 = 10$ as a function of r. If we compare these new coefficients (solid lines) with those from Fig. 3 (dashed lines with the same colours), we can see that the $\widehat{\boldsymbol{\beta}}_{\text{FRRM}}$ are still independent from r. At the same time, they have been shrunk towards zero and that means that $\widehat{\boldsymbol{\beta}}_{\text{FRRM}}\,\text{VAR}(\widehat{\mathbf{U}})\widehat{\boldsymbol{\beta}}^{\text{T}}_{\text{FRRM}}$ is also smaller than before. As a result, we need a larger $\lambda(r)$ to produce estimates of $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$ small enough to satisfy the bound in (4). The $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$ in Fig. 4 are smaller than the corresponding $\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}$ in Fig. 3.

It is also interesting to note that (9) can be implemented with penalised models other than a ridge regression. Any model that can shrink the coefficients associated with the sensitive attributes towards zero, thus decreasing the proportion of variance they explain in the response, can control the value of the bound $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ as a function of the tuning parameter. One possibility is to replace the ridge penalty with a LASSO penalty in order to perform feature selection on the sensitive

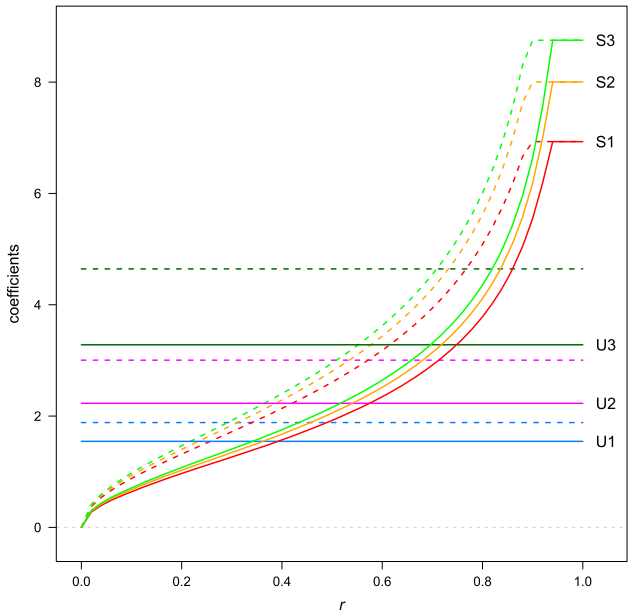

Fig. 4 Regression coefficients estimated by FRRM with $\lambda_2 = 10$ (solid lines) compared to those estimated without penalising the $\boldsymbol{\beta}$s (dashed lines, reported from Fig. 3). Lines in the same colour correspond to the same coefficient. (Color figure online)

attributes, a problem also investigated in Grgić-Hlača et al. (2018), Kazemi et al. (2018) and Khodadadian et al. (2021). Or we can combine it with the ridge penalty to obtain an elastic net model (Zou and Hastie 2005), which will often provide better predictive accuracy.

4.3 Different definitions of fairness

The modular approach to fairness used in (9) makes it possible to change the definition of fairness and its implementation in the bound used in model selection without affecting the estimation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

For instance, (10) uses $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r$ as a bound to enforce fairness as defined by statistical parity. But we can replace it with a similar bound for equality of opportunity, such as

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

where $\widehat{\mathbf{y}}$ is defined as before and $\boldsymbol{\phi}$, ψ are the coefficients of the regression model

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

If equality of opportunity holds exactly, $\widehat{\mathbf{y}}$ is independent from $\mathbf{S}$ given $\mathbf{y}$ and $\mathrm{COV}(\widehat{\mathbf{y}}, \mathbf{S} \mid \mathbf{y}) = 0$. Then all $\boldsymbol{\phi}$ are equal to zero and $R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi) = 0$. FRRM can achieve that asymptotically as $\lambda(r) \to \infty$ because $\widehat{\mathbf{y}} \to \widehat{\mathbf{U}}\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}$ and $\widehat{\mathbf{U}}$ is orthogonal to $\mathbf{S}$. If, on the other hand, $\lambda(r) \to 0$ the constraint becomes inactive because the $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ converge to the corresponding $\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}$. For finite, positive values of $\lambda(r)$, we have that $\widehat{\mathbf{y}} = \widehat{\mathbf{U}}\widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}} + \mathbf{S}\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ and therefore $|\,\mathrm{COV}(\widehat{\mathbf{y}}, \mathbf{S} \mid \mathbf{y})|$ will decrease as $\lambda(r)$ increases. This allows us to control $R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi)$ in the same way as we did $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ in Sect. 3.1.

A further advantage of enforcing fairness in this way is that we can control both statistical parity and equality of opportunity as a function of r (through $\lambda(r)$) at the same time. So, for instance, we could replace the constraint in (9) with $\max\{R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}), R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi)\}$ or a convex combination $wR^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) + (1-w)R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi)$, $w \in (0, 1)$. Few approaches in the literature combine different definitions of fairness in the same model; one example is Berk et al. (2017).

We can also choose to enforce individual fairness. Following along the lines of Berk et al. (2017), we can start by defining a penalty function

$$f(\boldsymbol{\alpha}, \mathbf{y}, \mathbf{S}) = \sum\nolimits_{i,j} d(y_i, y_j)(\mathbf{s}_i\boldsymbol{\alpha} - \mathbf{s}_j\boldsymbol{\alpha})^2$$

that penalises models in which individuals i and j with profiles $(y_i, \mathbf{u}_i, \mathbf{s}_i)$ and $(y_j, \mathbf{u}_j, \mathbf{s}_j)$ receive differential treatment in proportion to $(\mathbf{s}_i\boldsymbol{\alpha} - \mathbf{s}_j\boldsymbol{\alpha})^2$. If two individuals take the same values for the sensitive attributes, $\mathbf{s}_i = \mathbf{s}_j$ and their term vanish from the sum. If $\mathbf{s}_i \neq \mathbf{s}_j$, the corresponding term increases with both the difference in the outcomes, measured by some distance $d(y_i, y_j)$, and with the difference in their sensitive attributes $\mathbf{s}_i$ and $\mathbf{s}_j$.

If $\lambda(r) \to \infty$, then $(\mathbf{s}_i\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}} - \mathbf{s}_j\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}})^2 \to 0$ because all coefficients in $\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}$ are shrunk towards zero. As a result, $f(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \mathbf{y}, \mathbf{S})$ converges to zero as well. On the other hand, if $\lambda(r) \to 0$ then $f(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \mathbf{y}, \mathbf{S}) \to f(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \mathbf{y}, \mathbf{S})$ to take its maximum value.

For consistency with $R^2_{\mathbf{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ and $R^2_{\mathrm{EO}}(\boldsymbol{\alpha}, \boldsymbol{\beta})$, we then construct the constraint to use in (9) by normalising $f(\boldsymbol{\alpha}, \mathbf{y}, \mathbf{S})$ as

$$D_{\mathrm{IF}} = \frac{f(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \mathbf{y}, \mathbf{S})}{f(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \mathbf{y}, \mathbf{S})}$$

so that the bound r is defined in $[0, 1]$ as before. This is convenient for interpretation and to include D_{IF} in a convex combination with other fairness definitions.

Example 1 (continued) Consider the example from Sect. 2.2 one last time. Figure 5 shows the estimates of $R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi)$ and D_{IF} as a function of $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}}) = r \in [0, 1]$. For the sake of the example, we choose $d(y_i, y_j) = |y_i - y_j|$ in D_{IF}. As r increases, that is, as $\lambda(r) \to 0$, all of $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{FRRM}}, \widehat{\boldsymbol{\beta}}_{\mathrm{FRRM}})$, $R^2_{\mathrm{EO}}(\boldsymbol{\phi}, \psi)$ and D_{IF} increase monotonically. Hence any function that preserves their joint monotonicity can be used to enforce a user-specified combination of statistical parity, equality of opportunity and individual fairness.

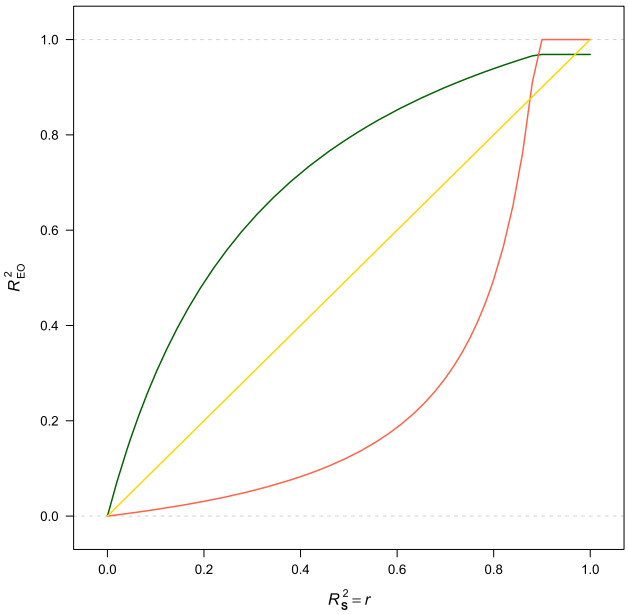

Fig. 5 $R^2_{\text{EQ}}(\boldsymbol{\phi}, \psi)$ (green) and D_{IF} (orange) as a function of $R^2_{\text{S}}(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \widehat{\boldsymbol{\beta}}_{\text{FRRM}}) = r$ (yellow). (Color figure online)

4.4 Generalised linear models

Another possible extension of FRRM is to adapt (10) to generalised linear models (GLMs; McCullagh and Nelder 1989) including Cox's proportional hazard models (Cox 1972). This makes it possible to introduce fair modelling in the extensive range of applications in which GLMs are a *de facto* standard while still being able to follow the best practices developed in the literature for those applications (significance testing, model comparison, confidence intervals for parameters, meta analysis, etc.).

Minimising the sum of squared residuals in a linear regression is a particular case of minimising the deviance $D(\cdot)$ (that is, -2 times the log-likelihood) of a generalised linear model, which we can constrain to ensure that we achieve the desired level of fairness. Starting from the general formulation of a GLM

$$\mathrm{E}(\mathbf{y}) = \boldsymbol{\mu}, \qquad \boldsymbol{\mu} = g^{-1}(\boldsymbol{\eta}), \qquad \boldsymbol{\eta} = \mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta},$$

where $g(\cdot)$ is the link function, we can draw on Friedman et al. (2010) and replace (10) with

$$(\widehat{\boldsymbol{\alpha}}_{\text{FRRM}}, \widehat{\boldsymbol{\beta}}_{\text{FRRM}}) = \underset{\boldsymbol{\alpha},\boldsymbol{\beta}}{\operatorname{argmin}}\, D(\boldsymbol{\alpha}, \boldsymbol{\beta}) + \lambda(r)\|\boldsymbol{\alpha}\|_2^2. \tag{16}$$

The ridge penalty $\lambda(r)$ can then be estimated to give

$$\frac{D(\boldsymbol{\alpha}, \boldsymbol{\beta}) - D(\mathbf{0}, \boldsymbol{\beta})}{D(\boldsymbol{\alpha}, \boldsymbol{\beta}) - D(\mathbf{0}, \mathbf{0})} \leqslant r. \tag{17}$$

We call this approach a *fair generalised ridge regression model* (FGRRM).

For a Gaussian GLM, (16) is identical to (10) because the deviance is just the residual sum of squares, and (17) simplifies to $R^2_{\text{S}}(\boldsymbol{\alpha}, \boldsymbol{\beta}) \leqslant r$. For a Binomial GLM with the canonical link function $\boldsymbol{\eta} = \log(\boldsymbol{\mu}/(1-\boldsymbol{\mu}))$, that is, a logistic regression, (17) bounds the difference made by $\mathbf{S}\boldsymbol{\alpha}$ in the classification odds. For a Poisson GLM with the canonical link $\boldsymbol{\eta} = \log \boldsymbol{\mu}$, that is, a log-linear regression, (17) bounds the difference in the intensity (that is, the expected number of arrivals per unit of time).

In the case of Cox's proportional hazard model for survival data, we can write the hazard function as

$$h\left(t; \widehat{\mathbf{U}}, \mathbf{S}\right) = h_0(t) \exp\left(\mathbf{S}\boldsymbol{\alpha} + \widehat{\mathbf{U}}\boldsymbol{\beta}\right)$$

where $h_0(t)$ is the baseline hazard at time t. The corresponding deviance can be used as in (16) and (17) to enforce the desired level of fairness, bounding the ratio of hazards through the difference in the effects of the sensitive attributes. The computational details of estimating this model are described in Simon et al. (2011).

Friedman et al. (2010) and Simon et al. (2011) describe how to fit GLMs and Cox's proportional hazard models with an elastic net penalty, which is a further extension to the application of FRRM to this class of models. We may also consider adapting one of the several pseudo-R^2 coefficients available in the literature, such as Nagelkerke (1991)'s or Tjur (2009)'s, to replace (17).

Finally, we note that the $\widehat{\boldsymbol{\beta}}_{\text{FGRRM}}$ are not constant over r in GLMs with a fixed scale factor (such as logistic and log-linear regressions): their values depend on the residual deviance, which changes as a function of r through the $\widehat{\boldsymbol{\alpha}}_{\text{FGRRM}}$. This phenomenon is described in detail in Mood (2010), and we illustrate it with the example below.

Example 2 Consider again the $\mathbf{X}$ and $\mathbf{S}$ from Example 1, this time in the context of a logistic regression with linear component

$$\boldsymbol{\eta} = 1 + 0.5X_1 + 0.6X_2 + 0.7X_3 + 0.8X_4 + 0.9X_0.5 + X_6.$$

The estimates of the regression coefficients given by FGRRM over $r \in [0, 1]$ are shown in Fig. 6. The $\widehat{\boldsymbol{\alpha}}_{\text{FGRRM}}$ are all equal to zero when $r = 0$, and they gradually increase to reach corresponding $\widehat{\boldsymbol{\alpha}}_{\text{OLS}}$ as in Fig. 3. In doing that, they gradually explain more and more of the deviance of the model, which forces the $\widehat{\beta}_{\text{FGRRM}}$ to increase as well. However, they increase monotonically, unlike the $\widehat{\boldsymbol{\beta}}_{\text{NCLM}}$, with a speed that matches that of the $\widehat{\boldsymbol{\alpha}}_{\text{FGRRM}}$. The change in scale is driven by the implicit constraint that a standard logistic distribution has a fixed variance: any increases in the variance explained by the $\widehat{\boldsymbol{\alpha}}_{\text{FGRRM}}$ also affect the variance of the residuals, thus forcing

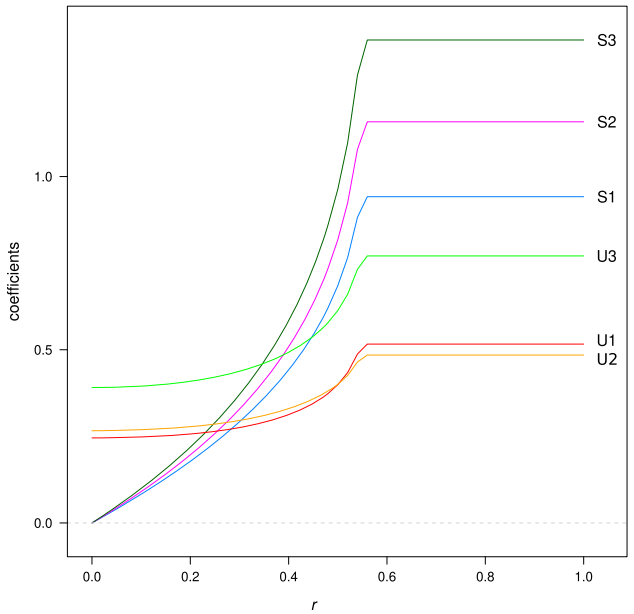

Fig. 6 Profile plot for the coefficients estimated by FGRRM as a function of the bound r for the model in Example 2

a rescaling of all coefficients to satisfy the constraint. We find this behaviour more intuitive to explain than NCLM's because it closely matches that of an unconstrained logistic regression as described in Mood (2010). The estimates of the $\boldsymbol{\beta}$ produced by the fair logistic regression model proposed by Zafar et al. (2019), which we will use in the experimental validation in Sect. 5.2, change non-monotonically in r like the $\widehat{\boldsymbol{\beta}}_{\text{NCLM}}$ (figure not shown for brevity). Note that using a quasi-Binomial GLM would remove the constraint on the scale factor and thus allow the $\widehat{\boldsymbol{\beta}}_{\text{FGRRM}}$ to be constant with respect to r.

5 Experimental evaluation

We evaluate the performance of F(G)RRM using NCLM and the fair regression models from Zafar et al. (2019) as baselines. We will label the latter as ZLM (*Zafar's linear model*) and ZLRM (*Zafar's logistic regression model*) in the following. All six data sets used in this section are available in the *fairml* R package (Scutari 2021); we refer the reader to its documentation for further details on each data set including how they have been preprocessed. F(G)RRM, NCLM and ZL(R)M are also implemented in *fairml*.

We choose ZL(R)M because it is a current, strong baseline (Zafar et al. 2019 shows that it outperforms four other methods from recent literature) and because it uses a definition of fairness that is comparable to that in (17):

- ZLRM controls the effect of the sensitive attributes on the response by bounding $|\,\text{COV}(\widehat{\boldsymbol{\eta}}, S_i)|$ marginally for each S_i;
- ZLM equivalently bounds $|\,\text{COV}(\widehat{\mathbf{y}}, S_i)|$, since $\widehat{\boldsymbol{\eta}} = \widehat{\mathbf{y}}$ in a linear regression model.

If $|\,\text{COV}(\widehat{\boldsymbol{\eta}}, S_i)| = 0$, then $\widehat{\mathbf{y}}$ is independent from S_i, giving statistical parity. If $|\,\text{COV}(\widehat{\boldsymbol{\eta}}, S_i)| > 0$, then its magnitude controls the proportion of the variance of $\widehat{\boldsymbol{\eta}}$ explained in the simple regression model of $\widehat{\boldsymbol{\eta}}$ against S_i. This proportion maps to the proportion of explained variance directly in ZLM, and to the proportion of explained deviance through the link function $g(\cdot)$ in ZLRM. The key difference between ZL(R)M and F(G)RRM is that ZL(R)M controls the overall proportion of variance or deviance explained by the sensitive attributes marginally for each S_i, while F(G)RRM controls it jointly for all S_i.

Overall, we find that F(G)RRM is at least as good as the best between NCLM and ZL(R)M in terms of both predictive accuracy and goodness of fit. In particular:

- FRRM outperforms NCLM for all but one data set when $r > 0$.
- F(G)RRM outperforms ZL(R)M for all considered data sets and for low values of r, that is, for models that have strong fairness constraints like those we may find in practical applications.

5.1 Fair linear regression models

We compare FRRM with NCLM and ZLM using the four real-world data sets that were also used in Komiyama et al. (2018) as well as the German Credit data set (Dua and Graff 2017). Our results for NCLM differ from those in Komiyama et al. (2018) due to the bias issue described in Sect. 2.2, although they do largely agree overall.

The Communities and Crime data set (C&C, 810 observations, 101 predictors) comprises socio-economic data and crime rates in communities in the USA: we take the normalised crime rate as the response variable, and the proportion of African American people and foreign-born people as the sensitive attributes. The COMPAS data set (COMPAS, 5855 observations, 13 predictors) comprises demographic and criminal records of offenders in Florida: we take recidivating within two years as the response variable and the offender's gender and race as the sensitive attributes. The National Longitudinal Survey of Youth data set (NLSY, 4908 observations, 13 predictors) is a collection of statistics from the U.S. Bureau of Labour Statistics on the labour market activities and life events of several groups: we take income in 1990 as the response variable, and gender and age as the sensitive attributes. The Law School Admissions Council data set (LSAC) is a survey among U.S. law school students: we take the GPA score of each student as the response variable, and the race and the age as the sensitive attributes. The German Credit data set (GCR, 1000 observations, 42 predictors)

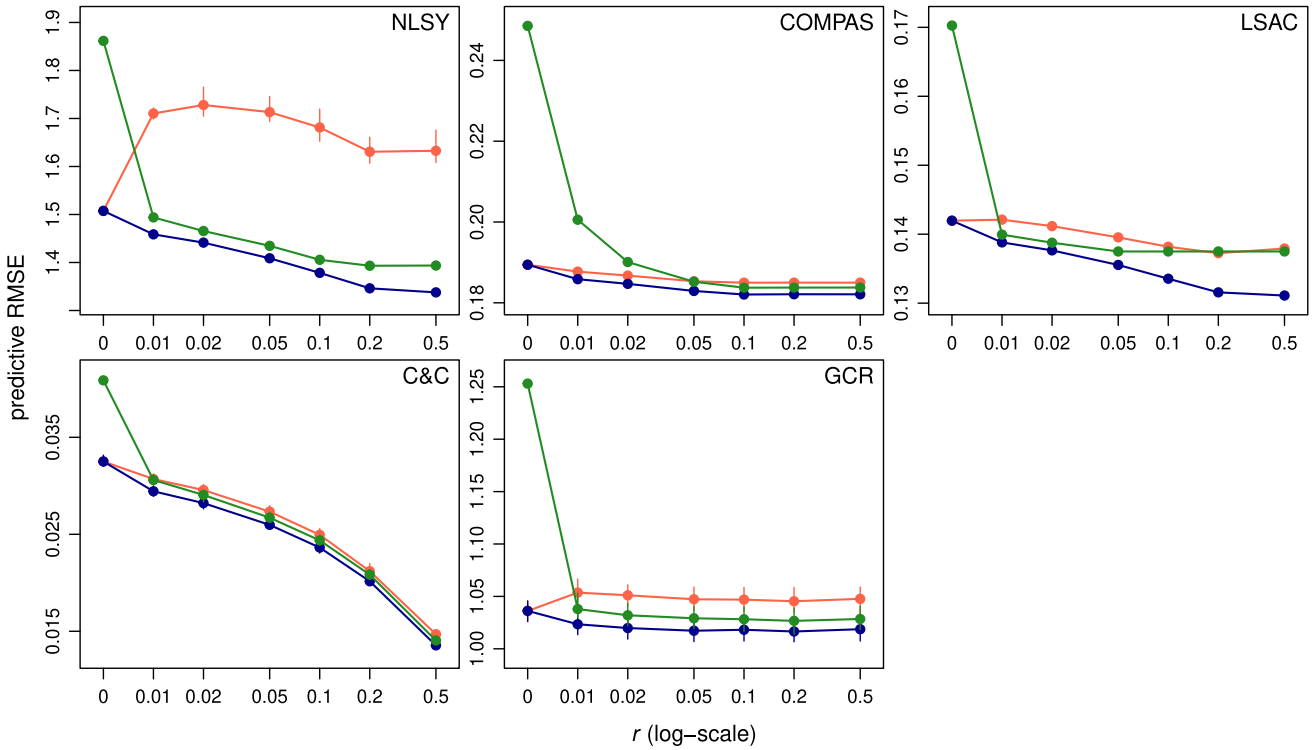

Fig. 7 Predictive RMSE for NCLM (orange), FRRM (blue) and ZLM (green) on the data sets described in Sect. 5.1. Bars show 90% confidence intervals. Lower values are better. (Color figure online)

is a collection of 700 "good" loans and 300 "bad" loans with a set of attributes that can be used to classify them as good or bad credit risks. We take the rate as the response variable, and the age, gender and foreign-born status as the sensitive attributes.

We evaluate both NCLM and FRRM using 50 runs of 10-fold cross-validation with constraint values $r = \{0, 0.01, 0.02, 0.05, 0.10, 0.20, 0.50\}$. We then measure the largest resulting $|\,\mathrm{COV}(\widehat{\mathbf{y}}, S_i)|$ for each r and we use that as the bound in ZLM to compare the accuracy of all models for the same level of fairness. To reduce simulation variability, in each run of cross-validation we use the same folds for all algorithms. We measure performance with:

- The predictive root-mean-square error (RMSE) produced by the model on the validation sets in the cross-validation;
- The training RMSE produced by the model on the training sets in the cross-validation.

The predictive RMSE is shown in Fig. 7. FRRM consistently achieves a smaller RMSE than NCLM across all data sets and $r > 0$. FRRM also achieves a smaller RMSE than ZLM in NLSY, COMPAS and LSAC for $r > 0$. In the case of C&C and GCR, FRRM achieves a lower RMSE than NCLM and ZLM but the difference is negligible for practical purposes.

For $r = 0$, FRRM and NCLM estimate the same model containing only the decorrelated predictors $\widehat{\mathbf{U}}$ and therefore have the same predictive RMSE. On the other hand, ZLM has a much higher RMSE than FRRM and NCLM because it estimates a model that only includes those predictors that are orthogonal to all sensitive attributes simultaneously ($|\,\mathrm{COV}(\widehat{\mathbf{y}}, S_i)| \propto |\,\mathrm{COV}(\mathbf{X}, S_i)| = 0$ for all S_i). However, the empirical covariances between predictors and sensitive attributes are usually numerically different from zero even when their theoretical counterparts are not. Hence ZLM ends up dropping more and more predictors as $r \to 0$ and estimates an intercept-only model for $r = 0$.

The training RMSE is shown in Fig. 8, and follows a similar pattern to the predictive RMSE in Fig. 7. However, it is notable that both FRRM and ZLM outperform NCLM, despite its theoretical optimality guarantees, for $r > 0$ in NLSY, LSAC and GCR, and for $r > 0.05$ in COMPAS. This is possible because the assumptions made in Yamada and Takeda (2018) and Komiyama et al. (2018) do not hold. Firstly, Yamada and Takeda (2018) assume that the constraint must be active, which is not the case whenever $R^2_{\mathbf{S}}(\widehat{\boldsymbol{\alpha}}_{\mathrm{OLS}}, \widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) \geqslant r$. Secondly, both Yamada and Takeda (2018) and Komiyama et al. (2018) assume that both $\mathrm{VAR}(\mathbf{S})$ and $\mathrm{VAR}(\widehat{\mathbf{U}})$ are full rank. While this is technically true for all data sets, we note that $\mathrm{VAR}(\mathbf{S})$ has at least one eigenvalue smaller than 10^{-6} in each of COMPAS, LSAC, NLSY and GCR. The fact that FRRM outperforms NCLM for all these

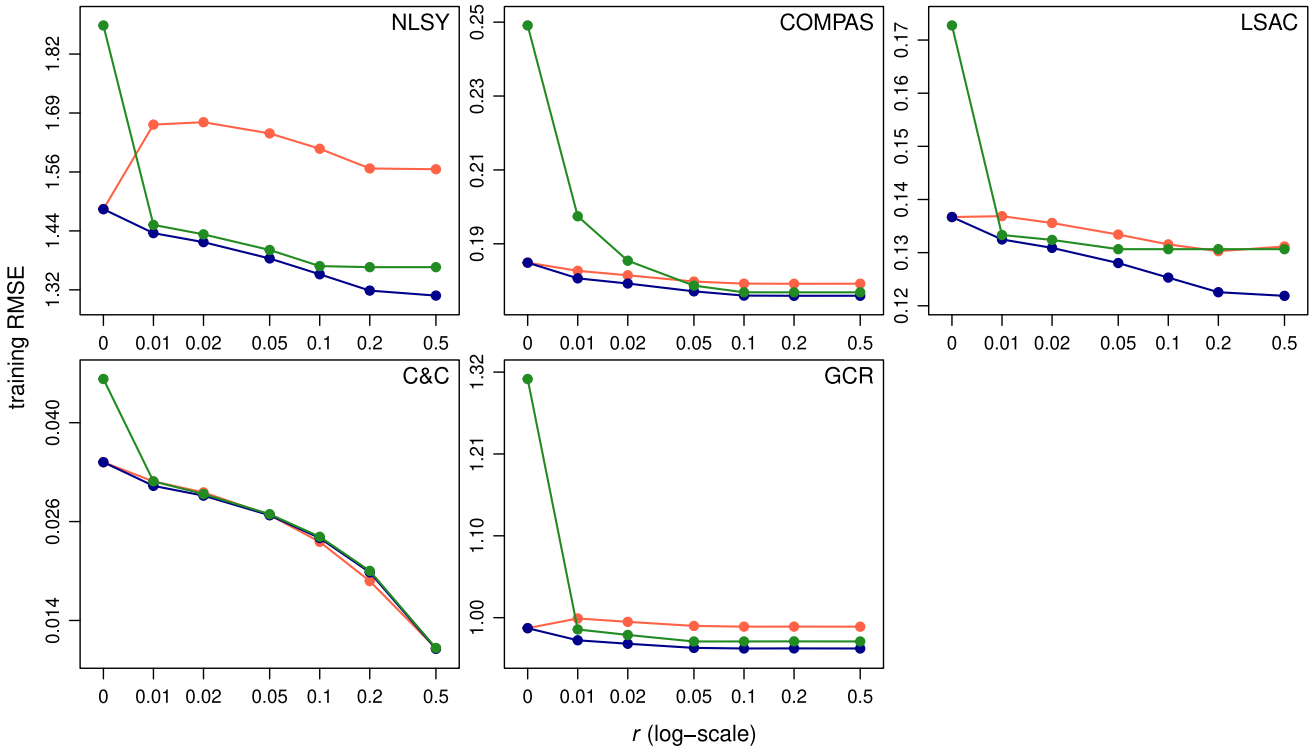

Fig. 8 Training RMSE for NCLM (orange), FRRM (blue) and ZLM (green) on the data sets described in Sect. 5.1. Bars showing 90% confidence intervals are too small to be visible. Lower values are better. (Color figure online)

data sets, but not for C&C, suggests that it is more numerically robust in this respect; which is expected since ridge regression does not make any assumption on the nature of $\mathbf{S}$.

All algorithms achieve the desired level of fairness on both on the training and the validation sets in the cross-validation whenever the bound r is active, that is, when the estimated model does not revert to an OLS regression. In particular, the level of fairness observed in the predictions for the validation sets is matches that required when training the models.

5.2 Fair logistic regression models

We now compare FGRRM with ZLRM, following the same steps in Sect. 5.1. However, we measure both predictive accuracy and goodness of fit with the F1 score (the harmonic average between precision and recall).

For this purpose, we will use the ADULT and BANK data sets that were also used in Zafar et al. (2019) as well as COMPAS. The ADULT data set (30162 observations, 14 predictors) contains a set of answers to the U.S. 1994 Census that are relevant for predicting whether a respondent's income exceeds \$50K. We take the binary income indicator (whether income is above or below \$50K) as the response variable, and sex and age as the sensitive attributes. Note that we enforce fairness for both sensitive attributes simultaneously, while Zafar et al. (2019) only considered them individually in separate models. The BANK data set (41,188 observations, 19 variables) contains information on the phone calls conducted by a Portuguese banking institution's direct marketing campaigns to convince prospective clients to subscribe a term deposit. We take the age as the sensitive attribute and whether the call resulted in a subscription as the response. The COMPAS data set is the same as in Sect. 5.1, but we now treat the response variable as a discrete binary variable.

The results for predictive accuracy are shown in Fig. 9. FGRRM systematically outperforms ZLRM for $r \leqslant 0.05$ for both ADULT and COMPAS, and the two models have equivalent performance for $r > 0.1$. In the first case the bounds are active; in the latter they are not, and both FGRRM and ZLRM revert back to unconstrained logistic regression models. As for the BANK data set, FGRRM and ZLRM have equivalent performance for all values of r because BANK contains just one sensitive attribute. Therefore, controlling the proportion of deviance explained by sensitive attributes marginally (in ZLRM) is the same as controlling it jointly (in FGRRM). For $r = 0$, ZLRM suffers a catastrophic loss in predictive accuracy for the same reasons as ZLM.

The observed goodness of fit follows the same patterns as predictive accuracy for all data sets, save for the fact that the F1 scores are higher by up to 0.02. Furthermore, both FGRRM and ZLM achieve the desired level of fairness as was the case for the models in Sect. 5.1.

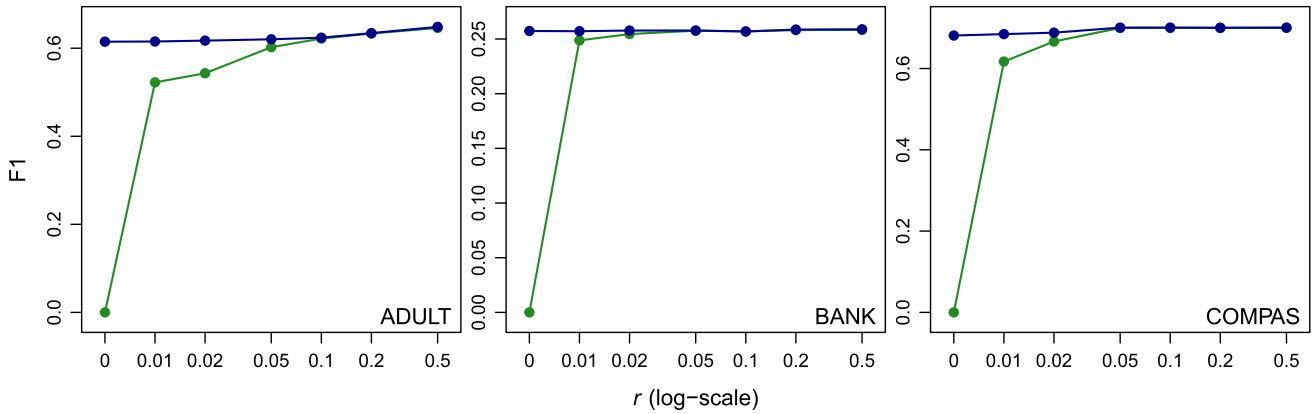

Fig. 9 Predictive F1 score for ZLRM (green) and FGRRM (blue) on the data sets described in Sect. 5.2. Bars showing 90% confidence intervals are too small to be visible. Higher values are better. (Color figure online)

5.3 Comparison with other fair models in the literature

The limitations intrinsic to other fair models in the literature prevent us from comparing them with F(G)RRM as thoroughly as we did for NCLM and ZL(R)M. Nevertheless, we can draw some limited results to get a partial view of how their performance relates to that of F(G)RRM. Here we consider the models built on statistical parity proposed by Steinberg et al. (2020) and Agarwal et al. (2018).

The fair regression model proposed by Steinberg et al. (2020) uses an auxiliary logistic regression model to control the effect of a single binary sensitive attribute on $\mathbf{y}$. The optimal regression is chosen as the model that maximises a penalised loglikelihood score computed as follows for a given penalty γ:

- They estimate the main regression model $\mathbf{y} = \mathbf{S}\boldsymbol{\beta}_{\mathbf{S}} + \mathbf{X}\boldsymbol{\beta}_{\mathbf{X}}$ to obtain $\widehat{\mathbf{y}}$.
- They approximate an auxiliary logistic regression of $\mathbf{S}$ on $\widehat{\mathbf{y}}$ and then approximate the mutual information between $\mathbf{S}$ and its fitted values $\widehat{\mathbf{S}}$.
- They add a penalty term equal to γ times the mutual information above to the loglikelihood of the main regression model to promote fairness.

Steinberg et al. (2020) do not provide an implementation of their proposed linear regression model. In our own implementation, we extend it in two ways to allow for a meaningful comparison with FRRM:

- We allow more than one binary sensitive attribute in the model by adding a separate penalty term for each, all with the same coefficient γ;
- We allow sensitive attributes with more than two values by using a multinomial logistic regression as the auxiliary model.

Even so, we are limited to the COMPAS data (same sensitive attributes as before), the NLSY data (gender as the only sensitive attribute) and the LSAC data (race as the only sensitive attribute). Furthermore, we are unable to control r exactly due to the highly nonlinear relationship between γ and r. The predictive RMSE for FRRM and for the model from Steinberg et al. (2020) are shown in Fig. 10: the former dominates the latter for $r < 0.1$ for all three data sets. The right-most point in the curves for the model from Steinberg et al. (2020) corresponds to $\gamma = 0$, that is, the regression model where the constraint is inactive. No further reductions of the penalty encoding the fairness are possible: since the model from Steinberg et al. (2020) does not outperform FRRM even then, we conclude that FRRM dominates it even for larger values of r.

Agarwal et al. (2018) estimate a fair classifier by choosing an optimal model over the set Δ of randomised classifiers subject to an inequality constraint that enforces fairness:

$$\min_{Q \in \Delta} \operatorname{err}(Q) \qquad \text{subject to} \qquad \mathbf{M}\boldsymbol{\mu}(Q) \leq \mathbf{c} + \boldsymbol{\varepsilon} \tag{18}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{K}| \times |\mathcal{I}|}$ and $\mathbf{c} \in \mathbb{R}^{|\mathcal{K}|}$ describe the linear constraints for the chosen definition of fairness, $\mathbf{M}(Q) \in \mathbb{R}^{|\mathcal{I}|}$ is a vector of conditional moments of functions of the classifier Q and $\boldsymbol{\epsilon} \in \mathbb{R}^{|\mathcal{K}|}$ controls the level of fairness. The choice of $\mathbf{M}$, $\boldsymbol{\mu}$ and $\mathbf{c}$ is subject to the chosen definition of fairness. The solution to (18) is found through a series of cost-sensitive classification problems by rewriting it as a saddle point problem with a Lagrangian multiplier and applying the exponentiated gradient reduction proposed by Freund and Schapire (1997) and Kivinen and Warmuth (1997).

Agarwal et al. (2018) provide an implementation of their approach in *Fairlearn* (Bird et al. 2020). It supports both regression and classification but it does not support statistical parity for regression: hence we compare it only with FGRRM. To match *fairml*, we use as base classifier the logis-

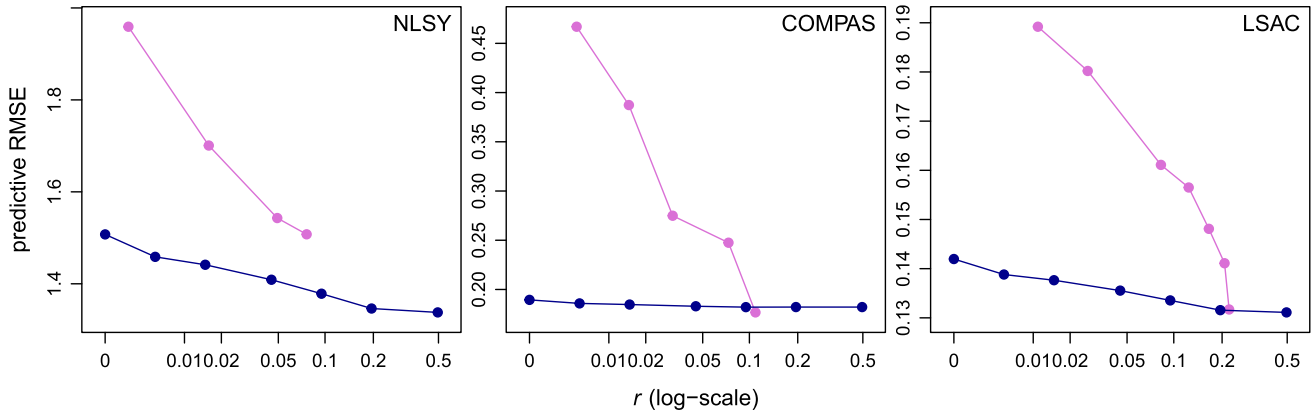

Fig. 10 Predictive RMSE for FRRM (blue) and the approach from Steinberg et al. (2020, violet) on the data sets to which both apply. Bars showing 90% confidence intervals are too small to be visible. Lower values are better. (Color figure online)

tic regression of *scikit-learn* with no penalty and "newton-cg" solver. For computational reasons, we only evaluate $\epsilon = 0.1$, which was also considered in Agarwal et al. (2018), and we fix all other parameters to their default values. Furthermore, *Fairlearn* only allows a single categorical sensitive attribute: this is not a limitation for the COMPAS data (we merged the two sensitive attributes into a single variable), but it is for the the ADULT data (we used sex as the only sensitive attribute). Finally, *Fairlearn* did not converge for the BANK data. The average F1 we obtain for both data sets ($F1 = 0.6787$ for COMPAS and $F1 = 0.5805$ for ADULT) is smaller than those produced by FGRRM ($[0.6891, 0.7029]$ for COMPAS and $[0.6141, 0.6489]$ for ADULT) over all the considered values of r .

6 Conclusions

In this paper, we presented a general framework for learning fair regression models that comprises both linear and generalised linear models. Our proposal, which we call F(G)RRM for *fair (generalised) ridge regression models*, uses a ridge penalty to reduce the proportion of variance (deviance) explained by the sensitive attributes over the total variance (deviance) explained by the model. Unlike most other approaches in the literature, it F(G)RRM can handle arbitrary types and combinations of predictors and sensitive attributes and different types of response variables.

Compared to the other approaches we have considered, we show that F(G)RRM achieve a better predictive accuracy and a better goodness of fit for the same level of fairness. (This is despite the optimality guarantees of NCLM, which we show may not hold in practical applications.) In addition, we argue that F(G)RRM produces regression coefficient estimates whose behaviour is more intuitive than the other models we investigated in this paper.

F(G)RRM compares favourably with NLCM and ZL(R)M in two other respects as well. Firstly, it is mathematically simpler and easier to implement since the only numeric optimisation it requires is root finding in a single variable bounded in a finite interval; the coefficient estimates are either available in closed form (for FRRM) or can be estimated with standard software (for FGRRM). Secondly, F(G)RRM is more modular than NLCM and ZL(R)M: it can be extended to use kernels for modelling nonlinear relationships, different penalties, and different definitions of fairness. It can accommodate multiple definitions of fairness simultaneously as well.

Funding Open access funding provided by SUPSI - University of Applied Sciences and Arts of Southern Switzerland. Marco Scutari and Manuel Proissl acknowledge the UBS-IDSIA research collaboration for the advancement of financial services with Artificial Intelligence, which served as the host for this joint work on algorithmic fairness. Francesca Panero was supported by the EPSRC and MRC Centre for Doctoral Training in Statistical Science, University of Oxford (Grant EP/L016710/1)

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., Wallach, H.: A reductions approach to fair classification. In: Proceedings of

Machine Learning Research. 35th International Conference on Machine Learning (ICML), vol. 80, pp. 60–69 (2018)
Agarwal, A., Dudik, M., Wu, Z.S.: Fair regression: quantitative definitions and reduction-based algorithms. In: Proceedings of Machine Learning Research. 36th International Conference on Machine Learning (ICML), vol. 97, pp. 120–129 (2019)
Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., Roth, A.: A convex framework for fair regression. In: Fairness, Accountability, and Transparency in Machine Learning (FATML) (2017)
Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art. Sociol. Methods Res. **50**(1), 3–44 (2021)
Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., Sameki, M., Wallach, H., Walker, K.: Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. Tech. Rep. MSR-TR-2020-32, Microsoft (2020)
Calders, T., Karim, A., Kamiran, F., Ali, W., Zhang, X.: Controlling attribute effect in linear regression. In: Proceedings of the 13th IEEE International Conference on Data Mining, pp. 71–80 (2013)
Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., Floridi, L.: Artificial intelligence and the "Good Society": the US, EU, and UK approach. Sci. Eng. Ethics **24**(2), 505–528 (2018)
Choraś, M., Pawlicki, M., Puchalski, D., Kozik, R.: Machine learning—the results are not the only thing that matters! What about security, explainability and fairness? In: Proceedings of the International Conference on Computational Science (ICCS), pp. 615–628 (2020)
Chzhen, E., Denis, C., Hebiri, M., Oneto, L., Pontil, M.: Fair regression via plug-in estimator and recalibration with statistical guarantees. Adv. Neural Inf. Process. Syst. **33**, 19137–19148 (2020)
Cox, D.R.: Regression models and life-tables. J. R. Stat. Soc. Ser. B **34**(2), 187–220 (1972)
Del Barrio, E., Gordaliza, P., Loubes, J.M.: Review of Mathematical Frameworks for Fairness in Machine Learning. arXiv:2005.13755 (2020)
Dua, D., Graff, C.: UCI Machine Learning Repository. http://archive.ics.uci.edu/ml (2017)
European Commission: Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (2021)
Freund, Y., Schapire, R.E.: A decision-theoretic generalization of online learning and an application to boosting. J. Comput. Syst. Sci. **55**(1), 119–139 (1997)
Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for generalized linear models via coordinate descent. J. Stat. Soft. **33**(1), 1–22 (2010)
Fukuchi, K., Sakuma, J., Kamishima, T.: Prediction with model-based neutrality. In: Proceedings of the joint European conference on machine learning and knowledge discovery in databases (ECML PKDD), pp. 499–514. Springer (2013)
Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., Walther, A.: Predictably Unequal? The Effects of Machine Learning on Credit Markets. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3072038 (2020)
Grgić-Hlača, N., Zafar, M.B., Gummadi, K.P., Weller, A.: Beyond distributive fairness in algorithmic decision making: feature selection for procedurally fair learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI) (2018)
Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer (2009)
Hoerl, A.E., Kennard, R.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
Kanagawa, M., Hennig, P., Sejdinovic, D., Sriperumbudur, B.K.: Gaussian Processes and Kernel Methods: A Review on Connections and Equivalences. arXiv:1807.02582 (2018)
Kazemi, E., Zadimoghaddam, M., Karbasi, A.: Scalable deletion-robust submodular maximization: data summarization with privacy and fairness constraints. In: Proceedings of Machine Learning Research. 35th International Conference on Machine Learning (ICML), vol. 80, pp. 2544–2553 (2018)
Khodadadian, S., Nafea, M., Ghassami, A., Kiyavash, N.: Information Theoretic Measures for Fairness-aware Feature Selection. arXiv:2106.00772 (2021)
Kivinen, J., Warmuth, M.K.: Exponentiated gradient versus gradient descent for linear predictors. Inform. Comput. **132**(1), 1–63 (1997)
Komiyama, J., Takeda, A., Honda, J., Shimao, H.: Nonconvex optimization for regression with fairness constraints. In: Proceedings of Machine Learning Research. 35th International Conference on Machine Learning (ICML), vol. 80, pp. 2737–2746 (2018)
Lipovetsky, S.: Two-parameter ridge regression and its convergence to the eventual pairwise model. Math. Comput. Modell. **44**(3–4), 204–318 (2006)
Mary, J., Calauzenes, C., El Karoui, N.: Fairness-aware learning for continuous attributes and treatments. In: Proceedings of Machine Learning Research. 36th International Conference on Machine Learning (ICML), vol. 97, pp. 4382–4391 (2019)
McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. CRC Press, Boca Raton (1989)
Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Comput. Surv. **54**(6), 115 (2021)
Mood, C.: Logistic regression: why we cannot do what we thin kWe can do, and what we can do about it. Eur. Sociol. Rev. **26**(1), 67–82 (2010)
Nagelkerke, N.J.D.: A note on a general definition of the coefficient of determination. Biometrika **78**(3), 691–692 (1991)
Pérez-Suay, A., Laparra, V., Mateo-García, G., Muñoz-Marí, J., Gómez-Chova, L., Camps-Valls, G.: Fair kernel learning. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD), pp. 339–355. Springer (2017)
Pessach, D., Shmueli, E.: Algorithmic Fairness. arXiv:2001.09784 (2020)
Raghavan, M., Barocas, S., Kleinberg, J., Levy, K.: Mitigating bias in algorithmic hiring: evaluating claims and practices. In: Proceedings of the 3rd Conference on Fairness, Accountability and Transparency, pp. 469–481 (2020)
Russell, C., Kusner, M.J., Loftus, J.R., Silva, R.: When worlds collide: integrating different countefactual assumptions in fairness. Adv. Neural Inf. Process. Syst. **30**, 6414–6423 (2017)
Saunders, C., Gammerman, A., Vovk, V.: Ridge regression learning algorithm in dual variables. In: Proceedings of the 15th International Conference on Machine Learning (ICML), pp. 515–521. Morgan Kauffmann (1998)
Scutari, M.: fairml: Fair Models in Machine Learning. R package version 0.7 (2021)
Simon, N., Friedman, J., Hastie, T., Tibshirani, R.: Regularization paths for Cox's proportional hazards model via coordinate descent. J. Stat. Softw. **39**(5), 1–13 (2011)
Steinberg, D., Reid, A., O'Callaghan, S., Lattimore, F., McCalman, L., Caetano, T.: Fast Fair Regression via Efficient Approximations of Mutual Information. arXiv:2002.06200 (2020)
Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
Tjur, T.: Coefficients of determination in logistic regression models–a new proposal: the coefficient of discrimination. Am. Stat. **63**(4), 366–372 (2009)
van Wieringen, W.N.: Lecture Notes on Ridge Regression. arXiv:1509.09169 (2018)

Wachter, S., Mittelstadt, B., Russell, C.: Bias preservation in machine learning: the legality of fairness metrics under EU non-discrimination law. W. Va. Law Rev. **123**, 735 (2021)

Woodworth, B., Gunasekar, S., Ohannessian, M.I., Srebro, N.: Learning non-discriminatory predictors. In: Proceedings of Machine Learning Research. Conference on Learning Theory (COLT), vol. 65, pp. 1920–1953 (2017)

Yamada, S., Takeda, A.: Successive Lagrangian relaxation algorithm for nonconvex quadratic optimization. J. Glob. Optim. **71**(2), 313–319 (2018)

Zafar, M.B., Valera, I., Gomez-Rodriguez, M., Gummadi, K.P.: Fairness constraints: a flexible approach for fair classification. J. Mach. Learn. Res. **20**, 1–42 (2019)

Zhang, Y., Duchi, J., Wainwright, M.: Divide and conquer kernel ridge regression: a distributed algorithm with minimax optimal rates. J. Mach. Learn. Res. **16**, 3299–3340 (2015)

Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B **67**(2), 301–320 (2005)