FAIR ENVELOPE REGRESSION: PREDICTIVE-SENSITIVE SUBSPACE DECOMPOSITION FOR FAIRNESS-UTILITY TRADE-OFFS

Anonymous authors
Paper under double-blind review

ABSTRACT

Fair regression with multivariate sensitive attributes requires controlling sensitive association without unnecessarily discarding predictive signal. We study this problem in linear regression with a continuous response and vector-valued sensitive attributes, using the proportion of prediction variance attributable to the sensitive attributes as the fairness criterion. We introduce Fair Envelope Regression Models (FERM), which uses envelope methodology to organize feature variation according to its relevance to the response and its association with the sensitive attributes. Under a compatibility condition, this yields four orthogonal components: predictive-only, shared predictive-sensitive, sensitive-only, and residual variation. FERM excludes sensitive-associated variation that is immaterial for prediction and penalizes only the shared predictive-sensitive component, providing a tunable path between a sensitive-invariant predictor and an unconstrained predictive model. Under the envelope model and regularity conditions, we establish consistency, asymptotic fairness at the fully constrained endpoint, reduced asymptotic estimation variance from removing response-immaterial directions, and a closed-form characterization of the resulting fairness–utility interpolation. Simulations and real-data experiments show that FERM can improve predictive utility at comparable levels of the chosen unfairness measure relative to closely matched fair-regression baselines. Beyond prediction performance, the decomposition makes explicit which predictive directions also carry sensitive-associated information, providing subspace-level auditability of the fairness-utility trade-off.

1 INTRODUCTION

Machine learning is increasingly deployed in high-stakes domains such as healthcare, finance, and criminal justice (Das et al., 2021; Bogen & Rieke, 2018; Kourou et al., 2015; De Fauw et al., 2018; Raji & Buolamwini, 2019; Buolamwini & Gebru, 2018), where unfair predictions can reinforce or amplify social inequities. Ensuring fairness in predictive models is therefore essential. In such settings, predictive performance is not the only consideration: predictions may also depend on sensitive attributes, either explicitly or through features that are statistically associated with them (Barocas et al., 2023; Calmon et al., 2017; Feldman et al., 2015). Fair regression therefore requires controlling sensitive association while preserving as much useful predictive information as possible. This problem becomes particularly challenging when the sensitive attribute is multivariate, since sensitive information may be distributed across many correlated directions of the feature space.

A central difficulty is that not all feature variation associated with sensitive attributes plays the same role in prediction. Some sensitive-associated directions may carry little or no information about the response and can therefore be removed with minimal loss of predictive utility. Other directions may be simultaneously associated with the sensitive attributes and strongly predictive of the response. Removing these directions indiscriminately can substantially reduce predictive performance. Conversely, retaining them without control can lead to predictions that remain strongly associated with the sensitive attributes. The relevant fairness–utility trade-off therefore occurs primarily in this shared predictive–sensitive variation.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Existing approaches to fair regression typically control the association between predictions and sensitive attributes through constraints, residualization, or regularization. For example, coefficient-of-determination and moment-based approaches constrain the amount of prediction variation attributable to sensitive variables, while fair ridge methods penalize their contribution to the fitted model (Calders et al., 2013; Johnson et al., 2016; Komiyama et al., 2018; Scutari et al., 2022; Berk et al., 2021; Agarwal et al., 2019). These approaches provide useful mechanisms for fairness control, but they generally operate without first distinguishing whether sensitive-associated feature variation is useful for predicting the response. As a result, response-irrelevant variation may still contribute estimation noise, while predictive and non-predictive sensitive-associated directions are treated through the same broad fairness mechanism. This motivates a more structured question: *Can we first identify where predictive information and sensitive association overlap, and then concentrate fairness regularization on precisely that part of the feature space?*

We address this question through Fair Envelope Regression (FERM), which adapts envelope regression to organize feature variation according to both its predictive relevance to the response Y and its association with possibly multivariate sensitive attributes S. Under a compatibility condition, this yields predictive-only, shared predictive–sensitive, sensitive-associated but response-immaterial, and residual components. This structure isolates the part of the feature space where predictive information and sensitive association genuinely overlap, allowing fairness regularization to focus on that shared component rather than treating all sensitive-associated variation alike. We study FERM in linear regression using a multiple-R^2-type fairness criterion for multivariate S. Beyond fairness control, the decomposition removes response-immaterial variation that can inflate estimation uncertainty without contributing predictive signal. Under the envelope model and regularity conditions, we establish consistency, asymptotic fairness at the fully constrained endpoint, asymptotic variance reduction, and a closed-form characterization of the interpolation between fair and unconstrained predictors. Through simulations and real-data experiments, we evaluate FERM using held-out fairness–utility frontiers and compare it with closely related fair-regression methods.

Our Contributions

1. **Predictive–sensitive feature decomposition:** We introduce an envelope-based construction that, under an explicit compatibility condition, decomposes feature variation into predictive-only, predictive–sensitive, sensitive-only, and residual components. This identifies where sensitive association is irrelevant to prediction and where it genuinely overlaps with useful predictive information.

2. **Selective fairness control with statistical guarantees:** FERM removes response-immaterial variation and regularizes only the shared predictive–sensitive component. We establish consistency, asymptotic fairness at the fully constrained endpoint, asymptotic efficiency gains from removing response-immaterial directions, and an explicit characterization of the fairness–utility interpolation.

3. **Subspace-level auditability and empirical evaluation:** The decomposition makes explicit which parts of the feature space carry both predictive and sensitive-associated information. Through simulations and real regression datasets, we evaluate the resulting fairness–utility frontiers against closely matched fair-regression baseline.

2 Related Work

Fairness-aware regression. Fair regression methods commonly control statistical dependence between predictions and sensitive attributes through moment constraints, constrained optimization, or regularization. Early approaches enforce decorrelation or related moment conditions (Calders et al., 2013; Johnson et al., 2016), while Komiyama et al. (2018) introduced a constrained least-squares formulation that limits the proportion of prediction variance attributable to sensitive attributes. Building on this formulation, Scutari et al. (2022) proposed a ridge-based alternative that provides a simple mechanism for controlling the fairness-utility trade-off and extends to generalized and kernel regression. Other approaches consider convex fairness penalties (Berk et al., 2021), kernel-based dependence criteria (Pérez-Suay et al., 2017; Perez-Suay et al., 2023; Wei et al., 2023), and reduction-based formulations for fair regression (Agarwal et al., 2019).

Beyond in-processing. Related work has also studied post-processing and statistical limits of fair regression. Wasserstein-based post-processing methods (Chzhen et al., 2020), demographic-parity transformations (Taturyan et al., 2024; Divol & Gaucher, 2024), and methods accounting for sample-selection bias (Du et al., 2022) provide alternatives to modifying the regression objective directly. Complementary minimax analyses characterize the statistical cost of fairness under different model and group structures (Zhao, 2021; Chzhen & Schreuder, 2022; Fukuchi & Sakuma, 2023). These works address broader fairness settings and objectives, whereas our focus is on structured linear regression with multivariate sensitive attributes and an explicit fairness–utility control.

Subspace structure and our contribution. FERM is most closely related to fair-regression methods that control the contribution of multivariate sensitive attributes, particularly Komiyama et al. (2018) and Scutari et al. (2022). Their formulations regulate sensitive association at the level of the fitted predictor, but do not distinguish whether sensitive-associated feature variation is relevant for predicting the response. Our approach brings the material-immaterial decomposition of envelope regression into this setting: feature variation is organized jointly by its relevance to the response and its association with the sensitive attributes. This separates sensitive-associated variation that is response-immaterial from variation that is both sensitive-associated and predictive, allowing the former to be removed and fairness regularization to be concentrated on the latter. The contribution is therefore not a new fairness penalty, but a structured decomposition that identifies *where* the fairness-utility trade-off occurs in the feature space.

3 PRELIMINARIES

3.1 SETUP AND NOTATION

Let (X, S, Y) denote a random tuple, where $X \in \mathbb{R}^{d_X}$ is the feature vector, $S \in \mathcal{S} \subseteq \mathbb{R}^{d_S}$ is a possibly multivariate sensitive attribute, and $Y \in \mathbb{R}$ is a continuous response. Given observations $\{(X_i, S_i, Y_i)\}_{i=1}^n$, we write $\mathbf{X} = [X_1, \ldots, X_n]^\top \in \mathbb{R}^{n \times d_X}$, $\mathbf{S} = [S_1, \ldots, S_n]^\top \in \mathbb{R}^{n \times d_S}$, $\mathbf{Y} = (Y_1, \ldots, Y_n)^\top$. Throughout, X, S, and Y are assumed centered. Our objective is to learn an accurate predictor of Y while controlling its statistical association with S. In FERM, the sensitive attributes are used during training to identify sensitive-associated directions of the feature space, while prediction is based on X.

3.2 FAIRNESS CRITERION FOR MULTIVARIATE SENSITIVE ATTRIBUTES

We focus on a linear-association notion of fairness that is particularly natural when S is continuous and multivariate (Komiyama et al., 2018; Scutari et al., 2022). For a centered prediction $\hat{Y}$, define

$$R_S^2(\hat{Y}) = \frac{\mathrm{Cov}(\hat{Y}, S)\, \Sigma_S^{-1}\, \mathrm{Cov}(S, \hat{Y})}{\mathrm{Var}(\hat{Y})}, \qquad \Sigma_S = \mathrm{Var}(S), \tag{1}$$

assuming Σ_S is nonsingular. This is the coefficient of determination obtained by linearly projecting $\hat{Y}$ onto S, and therefore measures the proportion of prediction variance linearly attributable to the sensitive attributes. A fairness budget $r \in [0, 1]$ requires $R_S^2(\hat{Y}) \leq r$, with smaller values imposing stricter control of sensitive association. In particular, $R_S^2(\hat{Y}) = 0$ implies $\mathrm{Cov}(\hat{Y}, S) = 0$. This definition coincides with the criterion used by Komiyama et al. (2018) and Scutari et al. (2022). Indeed, under their residualized representation $\hat{Y} = \alpha^\top S + \beta^\top U$ with $\mathrm{Cov}(S, U) = 0$, $R_S^2(\hat{Y}) = \frac{\mathrm{Var}(\alpha^\top S)}{\mathrm{Var}(\hat{Y})}$. The scalar criterion jointly accounts for all coordinates of a multivariate S and provides a direct parameter for tracing the fairness-utility trade-off.

3.3 RESIDUALIZATION-BASED FAIR REGRESSION

A closely related class of fair-regression methods first separates the features into variation linearly explained by S and an S-orthogonal residual. Specifically,

$$\mathbf{X} = \mathbf{S}B + \mathbf{U}, \qquad \hat{B} = (\mathbf{S}^\top \mathbf{S})^{-1} \mathbf{S}^\top \mathbf{X}, \qquad \hat{\mathbf{U}} = \mathbf{X} - \mathbf{S}\hat{B}, \tag{2}$$

so that $\mathbf{S}^\top \hat{\mathbf{U}} = 0$. Using this decomposition, Komiyama et al. (2018) fit $\hat{Y} = \alpha^\top S + \beta^\top \hat{U}$ while constraining the contribution of the sensitive component through an R^2-type fairness budget. Scutari et al. (2022) replace the corresponding constrained optimization with a ridge penalty on α, yielding the Fair Ridge Regression Model (FRRM). Specifically, they solve the following problem: $\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{S}\boldsymbol{\alpha} - \hat{\mathbf{U}}\boldsymbol{\beta}\|_2^2 + \lambda(r)\|\boldsymbol{\alpha}\|_2^2$, where the ridge penalty $\lambda(r) \geq 0$ ensures $R_S^2(\hat{Y}) \leq r$.

Residualization distinguishes S-associated from S-orthogonal variation, but it does not distinguish whether these feature directions are material for predicting Y. This distinction is central to our approach: FERM incorporates response relevance into the decomposition so that sensitive-associated variation that is immaterial for prediction can be separated from sensitive-associated variation that also carries predictive signal. Residualization therefore identifies which variation in X is associated with S, but not whether that variation is useful for predicting Y. We next use envelope structure to incorporate both relationships and formally construct the predictive-sensitive decomposition underlying FERM.

4 Fair Envelope Regression

4.1 Predictive-sensitive decomposition

FERM builds on the idea that feature variation should be characterized with respect to two distinct roles: its association with the sensitive attributes S and its relevance for predicting the response Y. Envelope regression provides a natural way to identify these two structures.

Let $\Gamma \in \mathbb{R}^{d_X \times m}$ span the S-associated envelope subspace of X, and let Γ_0 span its orthogonal complement. We assume $\Gamma_0^\top X \mid S \sim \Gamma_0^\top X, \quad \Gamma^\top X \perp\!\!\!\perp \Gamma_0^\top X \mid S$, so that variation along $\mathrm{span}(\Gamma_0)$ is invariant to S, whereas $\mathrm{span}(\Gamma)$ contains the feature variation associated with S. Similarly, let $\Phi \in \mathbb{R}^{d_X \times u}$ span the material predictor-envelope subspace for predicting Y from X, with orthogonal complement Φ_0. The predictor-envelope structure assumes $Y \perp\!\!\!\perp \Phi_0^\top X \mid \Phi^\top X, \quad \Phi^\top X \perp\!\!\!\perp \Phi_0^\top X$, so that $\Phi^\top X$ contains the information in X relevant for predicting Y, while $\Phi_0^\top X$ is response-immaterial. Let $P_S = \Gamma\Gamma^\top$, $P_Y = \Phi\Phi^\top$ denote the population projectors onto the S-associated and Y-material subspaces, respectively. We impose the compatibility condition: The population envelope projectors commute, i.e., $P_S P_Y = P_Y P_S$. Under this condition, the two decompositions can be simultaneously refined into four mutually orthogonal components with projection operators

$$P_{XSY} = P_S P_Y, P_{XY} = (I - P_S)P_Y, P_{XS} = P_S(I - P_Y), P_0 = (I - P_S)(I - P_Y). \quad (3)$$

These projectors are mutually orthogonal and satisfy $P_{XSY} + P_{XY} + P_{XS} + P_0 = I_{d_X}$. We denote their ranges by

$$\Pi_{XSY} = \mathrm{Range}(P_{XSY}), \quad \Pi_{XY} = \mathrm{Range}(P_{XY}), \quad \Pi_{XS} = \mathrm{Range}(P_{XS}), \quad \Pi_0 = \mathrm{Range}(P_0).$$

Equivalently,

$$\Pi_{XSY} = \mathrm{span}(\Gamma) \cap \mathrm{span}(\Phi) \quad \Pi_{XY} = \mathrm{span}(\Gamma_0) \cap \mathrm{span}(\Phi),$$
$$\Pi_{XS} = \mathrm{span}(\Gamma) \cap \mathrm{span}(\Phi_0) \quad \Pi_0 = \mathrm{span}(\Gamma_0) \cap \mathrm{span}(\Phi_0)$$

Consequently, every feature vector admits the decomposition $X = X_{XSY} + X_{XY} + X_{XS} + X_0, \quad X_A = P_A X$. The four components have distinct roles: X_{XY} is predictive of Y while lying in the S-invariant subspace; X_{XSY} is both Y-material and S-associated;X_{XS} is S-associated but immaterial for predicting Y;X_0 is neither Y-material nor S-associated. This separation identifies where the fairness-utility trade-off arises. The component X_{XS} carries sensitive-associated variation without contributing material predictive information and can therefore be excluded under the envelope model. In contrast, X_{XSY} contains predictive information that is also associated with S. Controlling this shared component may therefore improve fairness only at a potential cost in predictive utility. FERM concentrates its fairness regularization on precisely this component.

4.2 Estimating the decomposition

We estimate the four subspaces using standard envelope objectives (Cook et al., 2010; 2013; Cook & Zhang, 2015). For a positive-definite matrix $\widehat{M}$ and positive-semidefinite matrix $\widehat{U}$, an envelope basis of dimension q is obtained by

$$\widehat{G}_q = \underset{G \in \mathbb{R}^{p \times q}: G^\top G = I_q}{\arg\min} \left\{ \log|G^\top \widehat{M} G| + \log\left|G^\top (\widehat{M} + \widehat{U})^{-1} G\right| \right\}. \quad (4)$$

The optimization is carried out over the Grassmann manifold using standard envelope algorithms, such as the 1D or NIECE procedures (Cook & Zhang, 2016; Zhang et al., 2023).

Step 1: Estimate the Y-material subspace. We first estimate the predictor envelope for Y given X. Let $\widehat{\Sigma}_X$ and $\widehat{\Sigma}_{XY}$ denote the sample covariance and cross-covariance matrices. In equation 4, we use $\widehat{M}_Y = \widehat{\Sigma}_X$, $\widehat{U}_Y = \widehat{\Sigma}_{XY}\widehat{\Sigma}_{YX}$. For a selected dimension $\widehat{u}$, this yields an orthonormal basis $\widehat{\Phi} \in \mathbb{R}^{d_X \times \widehat{u}}$ for the estimated Y-material subspace and an orthogonal complement $\widehat{\Phi}_0$. The structural dimension $\widehat{u}$ is selected using BIC on the training data.

Step 2: Separate sensitive-associated variation within each Y-block. Define the projected features $Z_Y = \widehat{\Phi}^\top X$, $Z_0 = \widehat{\Phi}_0^\top X$. We then estimate an S-associated response envelope separately within each block. Within the Y-material block, let $\widehat{\Sigma}_{Z_Y} = \widehat{\Phi}^\top \widehat{\Sigma}_X \widehat{\Phi}$, , $\widehat{\Sigma}_{Z_Y S} = \widehat{\Phi}^\top \widehat{\Sigma}_{XS}$. We apply equation 4 with $\widehat{M}_{YS} = \widehat{\Sigma}_{Z_Y}$, $\widehat{U}_{YS} = \widehat{\Sigma}_{Z_Y S}\widehat{\Sigma}_{SZ_Y}$, obtaining an orthonormal basis $\widehat{G}_{YS}$ for the S-associated directions within the Y-material subspace and an orthogonal complement $\widehat{G}_Y$. Likewise, within the Y-immaterial block, $\widehat{\Sigma}_{Z_0} = \widehat{\Phi}_0^\top \widehat{\Sigma}_X \widehat{\Phi}_0$, $\widehat{\Sigma}_{Z_0 S} = \widehat{\Phi}_0^\top \widehat{\Sigma}_{XS}$, and the response-envelope objective yields $\widehat{G}_S$ and its orthogonal complement $\widehat{G}_0$.

Step 3: Construct the four estimated subspaces. Mapping the within-block envelope bases back to the original feature space gives

$$\begin{aligned} \widehat{\Pi}_{XSY} = \operatorname{span}(\widehat{\Phi}\,\widehat{G}_{YS}), \quad \widehat{\Pi}_{XY} = \operatorname{span}(\widehat{\Phi}\,\widehat{G}_Y), \\ \widehat{\Pi}_{XS} = \operatorname{span}(\widehat{\Phi}_0\,\widehat{G}_S), \quad \widehat{\Pi}_0 = \operatorname{span}(\widehat{\Phi}_0\,\widehat{G}_0). \end{aligned} \tag{5}$$

Let the corresponding orthogonal projection matrices be denoted by $\widehat{P}_{XSY}$, $\widehat{P}_{XY}$, $\widehat{P}_{XS}$, and $\widehat{P}_0$. The estimated feature components are then $\widehat{X}_A = \widehat{P}_A X, \quad A \in \{XSY, XY, XS, 0\}$.

4.3 Fairness-utility control

The decomposition in Sections 4.1–4.2 separates the response-material feature variation into two components: $\widehat{\Pi}_{XY}$, which is invariant to S, and $\widehat{\Pi}_{XSY}$, which is both predictive and sensitive-associated. FERM uses the former without penalization and controls the contribution of the latter.

For the training data, let $\mathbf{X}_{XY} = \mathbf{X}\widehat{P}_{XY}, \quad \mathbf{X}_{XSY} = \mathbf{X}\widehat{P}_{XSY}$. For $\lambda \geq 0$, we estimate

$$(\widehat{\beta}_{XY}(\lambda), \widehat{\beta}_{XSY}(\lambda)) = \underset{\beta_{XY}, \beta_{XSY}}{\arg\min} \left\{ \|\mathbf{Y} - \mathbf{X}_{XY}\beta_{XY} - \mathbf{X}_{XSY}\beta_{XSY}\|_2^2 + \lambda \|\beta_{XSY}\|_2^2 \right\}. \tag{6}$$

Thus, regularization is applied only to the shared predictive–sensitive component. The fitted predictor is $\widehat{\mathbf{Y}}_\lambda = \mathbf{X}_{XY}\widehat{\beta}_{XY}(\lambda) + \mathbf{X}_{XSY}\widehat{\beta}_{XSY}(\lambda)$. We select λ through the fairness budget introduced in Section 3.2. For a prescribed $r \in [0, 1]$, define $\lambda(r) = \inf\left\{\lambda \geq 0 : \widehat{R}_S^2(\widehat{\mathbf{Y}}_\lambda) \leq r\right\}$, where $\widehat{R}_S^2$ denotes the empirical multiple coefficient of determination between the fitted predictions and the sensitive attributes. The resulting FERM predictor is $\widehat{\mathbf{Y}}_r = \widehat{\mathbf{Y}}_{\lambda(r)}$. Hence, r specifies the desired level of sensitive association, while $\lambda(r)$ is induced by that choice rather than tuned independently.

The construction connects two natural endpoints. When $\lambda = 0$, FERM gives the unpenalized predictor based on the full Y-material subspace, $\widehat{\mathbf{Y}}_{\text{un}} = \mathbf{X}_{XY}\widehat{\beta}_{XY} + \mathbf{X}_{XSY}\widehat{\beta}_{XSY}$. As $\lambda \to \infty$, the shared component is suppressed and the predictor reduces to $\widehat{\mathbf{Y}}_{\text{fair}} = \mathbf{X}_{XY}\widehat{\beta}_{XY}^{\text{fair}}$. At the population level, X_{XY} lies in the S-invariant subspace; therefore the fully constrained predictor has zero sensitive association under the model assumptions. With estimated subspaces, this property is asymptotic, as formalized in Section 5.

FERM therefore does not regularize all feature variation associated with S. Sensitive-associated variation that is immaterial for predicting Y has already been removed through $\widehat{\Pi}_{XS}$, while predictive variation that is invariant to S is retained without penalty. The fairness parameter acts only on $\widehat{\Pi}_{XSY}$, the component for which reducing sensitive association may require sacrificing predictive information. This is the part of the feature space in which the fairness-utility trade-off arises.

5 THEORETICAL PROPERTIES

We now study the statistical consequences of the predictive-sensitive decomposition. The results address three questions: whether removing response-immaterial variation improves estimation efficiency, whether the estimated decomposition yields consistent and asymptotically fair predictors, and how penalizing the shared predictive-sensitive component moves the predictor between the unconstrained and fully constrained endpoints.

5.1 EFFICIENCY FROM REMOVING RESPONSE-IMMATERIAL VARIATION

Consider first the unpenalized problem. Since $\beta \in \mathrm{Range}(P_Y)$, directions in $P_Y^{\perp}$ contain no population predictive signal. Estimating coefficients along these directions can therefore increase sampling variability without improving the population regression function. We assume, the response satisfies

$$Y = \beta^\top X + \varepsilon, \qquad \mathbb{E}(\varepsilon \mid X) = 0, \qquad \mathrm{Var}(\varepsilon \mid X) = \sigma^2,$$

with $\beta \in \mathrm{Range}(P_Y)$. Thus, response-immaterial directions contain no population regression signal. Let $\widehat{\beta}_{\mathrm{OLS}}$ denote the unrestricted OLS estimator and define the oracle predictive projection $\widehat{\beta}_{\mathrm{env}} = P_Y \widehat{\beta}_{\mathrm{OLS}}$.

Theorem 5.1 (Variance reduction by predictive projection)**.** *Suppose* $\sqrt{n}(\widehat{\beta}_{\mathrm{OLS}} - \beta) \xrightarrow{d} N(0, \Sigma_\beta)$, *and* P_Y *is a reducing projector for* Σ_β, *i.e.,* $P_Y \Sigma_\beta = \Sigma_\beta P_Y$. *Then* $\sqrt{n}(\widehat{\beta}_{\mathrm{env}} - \beta) \xrightarrow{d} N(0, P_Y \Sigma_\beta P_Y)$, *and* $P_Y \Sigma_\beta P_Y \preceq \Sigma_\beta$. *Moreover,* $\Sigma_\beta - P_Y \Sigma_\beta P_Y$ *is nonzero whenever the response-immaterial subspace has positive asymptotic estimation variance.*

Theorem 5.1 formalizes the efficiency benefit of removing response-immaterial directions. Projection retains the population coefficient because $P_Y \beta = \beta$, while eliminating sampling variation in directions that do not contribute to the regression function.

Corollary 5.2 (Leading-order prediction efficiency)**.** *Let* X_{new} *be an independent test feature with covariance* Σ_X. *Under the conditions of Theorem 5.1, the leading-order estimation component of prediction risk satisfies* $\mathrm{tr}\,(\Sigma_X P_Y \Sigma_\beta P_Y) \le \mathrm{tr}\,(\Sigma_X \Sigma_\beta)$. *Hence removing response-immaterial directions cannot increase the leading-order prediction variance under the correctly specified envelope model.*

5.2 CONSISTENCY AND ASYMPTOTIC FAIRNESS

We next account for estimation of the envelope subspaces. Let $\widehat{P}_Y = \widehat{P}_{XY} + \widehat{P}_{XSY}$. Under standard envelope regularity conditions, $\widehat{P}_Y$ converges to P_Y at the parametric rate.

Proposition 5.3 (Consistency under estimated subspaces)**.** *Suppose* $\widehat{\beta}_{\mathrm{OLS}} - \beta = O_p(n^{-1/2})$ *and* $\|\widehat{P}_Y - P_Y\| = O_p(n^{-1/2})$. *Then the projected estimator* $\widehat{\beta}_{\mathrm{env}} = \widehat{P}_Y \widehat{\beta}_{\mathrm{OLS}}$ *satisfies* $\widehat{\beta}_{\mathrm{env}} - \beta = O_p(n^{-1/2})$.

Theorem 5.4 (Asymptotic fairness at the constrained endpoint)**.** *Let* $\widehat{Y}_{\mathrm{fair}} = \widehat{\beta}_{XY}^\top \widehat{P}_{XY} X$ *and* $\mathrm{Cov}(P_{XY} X, S) = 0$. *If* $\widehat{P}_{XY} \xrightarrow{p} P_{XY}$ *and* $\widehat{\beta}_{XY} \xrightarrow{p} \beta_{XY}$, *then* $\mathrm{Cov}(\widehat{Y}_{\mathrm{fair}}, S) \xrightarrow{p} 0$, *and consequently* $R^2_S(\widehat{Y}_{\mathrm{fair}}) \xrightarrow{p} 0$.

5.3 CHARACTERIZING THE FAIRNESS-UTILITY PATH

We finally characterize how ridge penalization of the shared predictive-sensitive component moves the fitted predictor between its two endpoints.

Theorem 5.5 (Interpolation along the shared subspace)**.** *Let* $\mathbf{Z}_1 = \mathbf{X}_{XY}$, $\mathbf{Z}_2 = \mathbf{X}_{XSY}$. *Assume that the two design blocks are sample-orthogonal and that the shared block is standardized:* $\mathbf{Z}_1^\top \mathbf{Z}_2 = 0$, $\mathbf{Z}_2^\top \mathbf{Z}_2 = nI$. *Let* $\widehat{\mathbf{Y}}_{\mathrm{un}}$ *denote the unpenalized FERM fit using both blocks, and let* $\widehat{\mathbf{Y}}_{\mathrm{fair}}$ *denote the fit using* $\mathbf{Z}_1$ *alone. Then the penalized FERM predictor in equation 6 satisfies*

$$\widehat{\mathbf{Y}}_\lambda = \omega(\lambda)\widehat{\mathbf{Y}}_{\mathrm{un}} + \{1 - \omega(\lambda)\}\widehat{\mathbf{Y}}_{\mathrm{fair}}, \quad \omega(\lambda) = \frac{n}{n+\lambda}. \tag{7}$$

Thus, $\lambda = 0 \implies \widehat{\mathbf{Y}}_\lambda = \widehat{\mathbf{Y}}_{\mathrm{un}}$, *while* $\lambda \to \infty \implies \widehat{\mathbf{Y}}_\lambda \to \widehat{\mathbf{Y}}_{\mathrm{fair}}$.

Corollary 5.6 (Utility cost of penalization)**.** *Under the conditions of Theorem 5.5,*

$$\mathrm{E}\left\|\mathbf{Y}-\widehat{\mathbf{Y}}_{\lambda}\right\|_2^2=\mathrm{E}\left\|\mathbf{Y}-\widehat{\mathbf{Y}}_{\mathrm{un}}\right\|_2^2+\{1-\omega(\lambda)\}^2\,\mathrm{E}\left\|\widehat{\mathbf{Y}}_{\mathrm{un}}-\widehat{\mathbf{Y}}_{\mathrm{fair}}\right\|_2^2. \tag{8}$$

Because $\lambda(r)$ in Section 4.3 is chosen to satisfy the prescribed fairness budget, equations equation 7-equation 8 show how stronger suppression of the shared predictive-sensitive component moves the predictor toward the fully constrained endpoint. The result identifies precisely where the utility cost arises: the difference between the predictive signal available in the shared subspace and that retained by the S-invariant predictive subspace.

6 Simulation Study

We use simulations to evaluate the fairness-utility behavior of FERM and to isolate the benefit of incorporating response relevance into the feature decomposition. We compare the full method, denoted **FERM-predictive**, with the Fair Ridge Regression Model (FRRM) of Scutari et al. (2022) and with an ablated version of FERM, denoted **FERM-decorrelated**. FERM-decorrelated uses the full S-invariant feature block, $X_{XY}+X_0$, together with the shared predictive–sensitive component X_{XSY}. FERM-predictive additionally removes the response-immaterial component X_0 and therefore uses only $X_{XY}+X_{XSY}$. In both variants, the shared component X_{XSY} is ridge-penalized to control the prescribed fairness budget, while the S-invariant component is left unpenalized. Thus, their comparison isolates the effect of removing response-immaterial variation, which is the efficiency mechanism studied in Section 5.1.

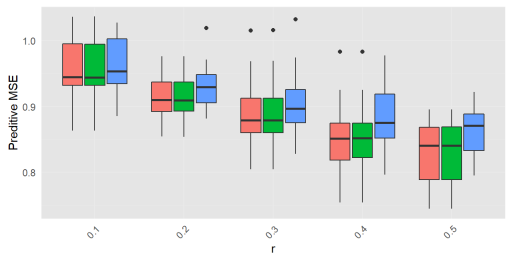
Data generation. We generate $\{(X_i,S_i,Y_i)\}_{i=1}^n$ from $Y=\alpha^\top S+\beta^\top X+\epsilon,\quad \epsilon\sim\mathcal{N}(0,0.5^2)$, where $\alpha\in\mathbb{R}^{d_S}$ and $\beta\in\mathbb{R}^{d_X}$. The sensitive attributes are generated as $S\sim\mathcal{N}(0_{d_S},I_{d_S})$. To induce sensitive-associated feature variation, we partition $X=[X_{\mathrm{corr}},X_{\mathrm{indep}}]$, where $X_{\mathrm{corr}}\in\mathbb{R}^{d_{\mathrm{corr}}}$ is correlated with S and $X_{\mathrm{indep}}\in\mathbb{R}^{dX-d_{\mathrm{corr}}}$ is independent of S. Specifically, $X_{\mathrm{corr}}=\beta_{SX}^\top S_Q S+\eta$, where $S_Q\in\mathbb{R}^{d_S\times d_S}$ is a projection matrix of rank q_S, $\beta_{SX}\in\mathbb{R}^{d_S\times d_{\mathrm{corr}}}$, and η is independent of S. The remaining coordinates are generated independently from a standard multivariate normal distribution.

All variables are standardized before model fitting, and each generated dataset is randomly divided into 80% training and 20% test observations. Table 1 summarizes the three simulation settings. Settings (1)-(2) use $\eta\sim\mathcal{N}(0,I)$, while Setting (3) replaces the Gaussian noise by independent centered Poisson variables, $\eta_j\sim\mathrm{Poi}(1)-1$, to assess robustness to non-Gaussian feature noise. Results for Setting (3) are reported in Appendix F. We evaluate all methods at fairness budgets $r\in\{0.1,0.2,0.3,0.4,0.5\}$. For each fitted model, we report predictive mean squared error (MSE) and the attained fairness measure on the held-out test set. Each configuration is repeated 50 times.

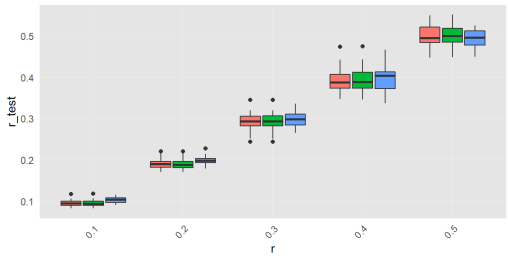
Table 1: Simulation settings.

Setting	d_X	d_{corr}	q_S	d_S	n	η
(1)	40	20	5	10	5000	Gaussian
(2)	100	50	15	20	20000	Gaussian
(3)	40	20	5	10	5000	centered Poisson

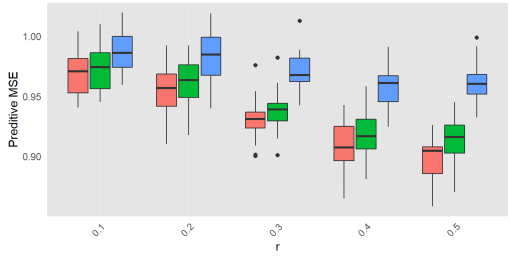
Results. Figure 1 shows the results under the two Gaussian settings. Across the considered fairness budgets, both decomposition-based methods achieve lower test MSE than FRRM while attaining similar levels of held-out sensitive association. The gain is more pronounced in the higher-dimensional Setting (2). The comparison between the two FERM variants isolates the role of response-material dimension reduction. FERM-decorrelated retains the response-immaterial component X_0, whereas FERM-predictive removes it before regression. Their performance is similar in Setting (1), while FERM-predictive achieves a clearer reduction in prediction error in the higher-dimensional setting. This behavior is consistent with the efficiency result in Section 5.1: removing response-immaterial directions reduces estimation variability without removing population predictive signal under the envelope model. The right panels show that the attained test fairness levels generally track the prescribed budgets for all methods. The same qualitative behavior is observed



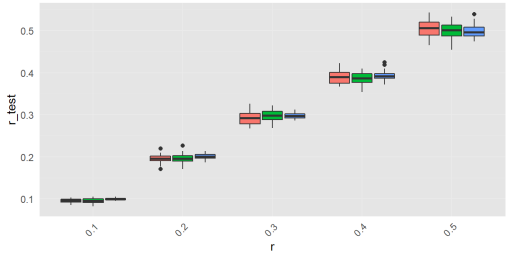
(1a) Simulation setting 1: Predictive MSE



(1b) Simulation setting 1: Unfairness on test data



(2a) Simulation setting 2: Predictive MSE



(2b) Simulation setting 2: Unfairness on test data

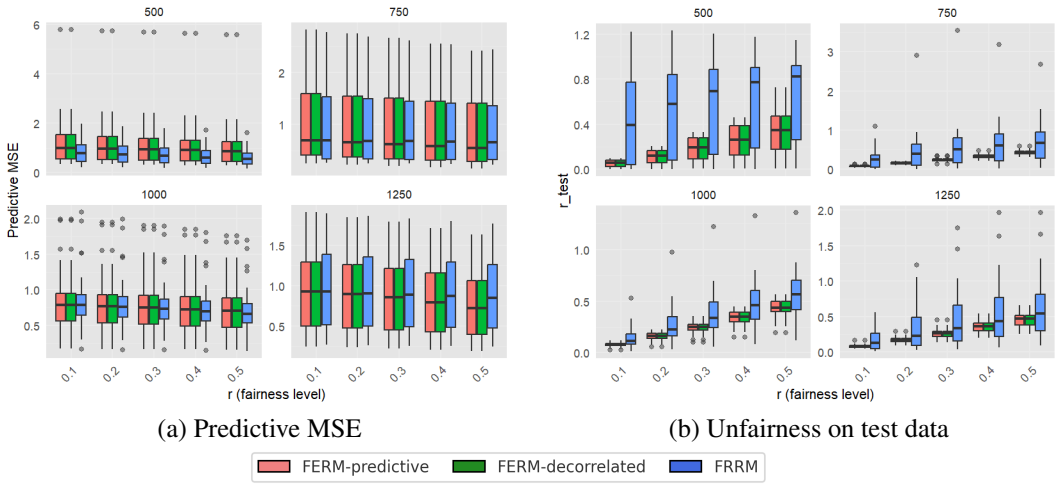


Figure 1: Left panel: Predictive MSE for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) for various unfairness levels r; lower values are better. Right panel: Unfairness levels on test data (r_{test}) for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) at varying unfairness levels r. Simulation settings are described in Table 1.

under centered Poisson noise (Appendix F, Figure 8), indicating that the observed advantage is not restricted to Gaussian feature noise. Overall, the simulations support the proposed role of the predictive-sensitive decomposition: separating response-relevant from response-immaterial variation can improve predictive efficiency while maintaining comparable control of sensitive association.

6.1 Real-world data

We evaluate FERM on the Full-Year Consolidated files from the *Medical Expenditure Panel Survey (MEPS)*. The continuous response Y is total annual medical expenditure (`totexp`). The multivariate sensitive attribute S contains eight standardized socioeconomic and payment-related variables, including age, family income, poverty level, and payments from major insurance sources. The feature vector X contains 100 continuous health, demographic, income, and expenditure variables after removing identifiers, low-variance fields, and categorical-like variables. We consider fairness budgets $r \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ and sample sizes $n \in \{500, 750, 1000, 1250\}$. For each configuration, the data are subsampled as described in the experimental setup, and the experiment is repeated 30 times. Figure 2 reports held-out predictive MSE and attained sensitive association for FRRM, FERM-decorrelated, and FERM-predictive. Across most settings, the two FERM variants achieve predictive error comparable to or lower than FRRM while providing tighter control of the attained fairness level. In several configurations, FRRM attains similar or slightly lower MSE but exceeds the prescribed fairness budget, whereas FERM remains closer to the target level. These results illustrate the importance of comparing predictive utility at the fairness levels actually achieved rather than considering MSE alone. An additional application to the U.S. Health Insurance dataset is provided in Appendix F.2.



(a) Predictive MSE (b) Unfairness on test data

Figure 2: Left panel: Predictive MSE for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) for unfairness levels $r = \{0.1, 0.2, 0.3, 0.4, 0.5\}$; lower values are better. Right panel: Unfairness levels on test data (r_{test}) for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) at unfairness levels $r = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. The plots for sample sizes $n = 500, 750, 1000, 1250$ are shown here.

7 Conclusion

We introduced Fair Envelope Regression (FERM), a structured approach to fair regression that organizes feature variation according to both its relevance for predicting Y and its association with multivariate sensitive attributes S. The resulting decomposition distinguishes sensitive-associated variation that is immaterial for prediction from variation that is simultaneously predictive and sensitive-associated. FERM removes the former and concentrates fairness regularization on the latter, where the fairness–utility trade-off arises. Under the envelope model and regularity conditions, we establish consistency, asymptotic fairness at the fully constrained endpoint, efficiency gains from removing response-immaterial variation, and an explicit characterization of the interpolation between constrained and unconstrained predictors. Simulations and real-data experiments show that the decomposition can improve predictive efficiency while maintaining comparable or tighter control of sensitive association than the matched fair-regression baseline. The current framework relies on linear envelope structure and a linear-association fairness criterion. Extending the predictive-sensitive decomposition to nonlinear representations, alternative fairness criteria, and high-dimensional settings provides a natural direction for future work.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on the conceptual framing and research methodology, assist in refining mathematical claims and their proof structure, and support the interpretation and presentation of experimental results. We have not used generative AI tools to generate synthetic datasets, implement the proposed methods, clean or reformat datasets, translate research material, or conduct qualitative or thematic data analysis; the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools to draft and edit portions of the manuscript for clarity and concision, suggest manuscript structure, titles and keywords, summarize and organize relevant literature, and assist with brainstorming and identifying potential gaps in presentation. All AI-assisted mathematical statements and methodological suggestions were independently checked by the authors against the underlying derivations and established literature, and all descriptions of experimental results were verified against the actual outputs of our experiments. The authors reviewed and revised all AI-assisted text and take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work concerns fairness-aware regression and therefore directly engages with questions of bias and discrimination in predictive systems. FERM controls a specific notion of linear association between predictions and multivariate sensitive attributes and should not be interpreted as guaranteeing broader, causal, or deployment-level fairness. The choice of sensitive attributes and fairness thresholds is application dependent and may involve normative considerations. Our experiments use publicly available datasets, involve no new human-subject data collection, and do not deploy the proposed method in consequential decision-making systems. No new personal data are released. As with other fairness methods, inappropriate specification of the fairness criterion or sensitive attributes, or use outside the model assumptions, may lead to misleading conclusions about fairness.

REPRODUCIBILITY STATEMENT

Complete proofs are provided in Appendix C, algorithmic and computational details in Appendices A and G, and code for reproducing the simulation and real-data experiments is included in the supplementary material. Random seeds and experimental settings are fixed for reproducibility.

REFERENCES

Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pp. 120–129. PMLR, 2019.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Miranda Bogen and Aaron Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn, December*, 7, 2018.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.

Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pp. 71–80. IEEE, 2013.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Silvia Chiappa. Path-specific counterfactual fairness. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 7801–7808, 2019.

Evgeny Chzhen and Nicolas Schreuder. A minimax framework for quantifying risk-fairness trade-off in regression. *The Annals of Statistics*, 50(4):2416–2442, 2022.

Evgeny Chzhen, Clément Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In *Advances in Neural Information Processing Systems*, volume 33, pp. 7321–7331, 2020.

R Dennis Cook and Xin Zhang. Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611, 2015.

R Dennis Cook and Xin Zhang. Algorithms for envelope estimation. *Journal of Computational and Graphical Statistics*, 25(1):284–300, 2016.

R Dennis Cook and Xin Zhang. Fast envelope algorithms. *Statistica Sinica*, 28(3):1179–1197, 2018.

R Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, pp. 927–960, 2010.

R Dennis Cook, IS Helland, and Zoey Su. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(5):851–877, 2013.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pp. 1436–1445. PMLR, 2019.

Sanjiv Das, Michele Donini, Jason Gelman, Kevin Haas, Mila Hardt, Jared Katzman, Krishnaram Kenthapadi, Pedro Larroy, Pinar Yilmaz, and Bilal Zafar. Fairness measures for machine learning in finance. *The Journal of Financial Data Science*, 2021.

Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

Victor Divol and Simon Gaucher. Demographic parity in regression and classification within the unawareness framework. arXiv preprint arXiv:2409.02471, 2024.

Wei Du, Xintao Wu, and Hanghang Tong. Fair regression under sample selection bias. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 1435–1444. IEEE, 2022.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Kazuto Fukuchi and Jun Sakuma. Demographic parity constrained minimax optimal regression under linear model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8653–8689, 2023.

Kory D Johnson, Dean P Foster, and Robert A Stine. Impartial predictive modeling: Ensuring fairness in arbitrary models. *arXiv preprint arXiv:1608.00528*, 2, 2016.

Motonobu Kanagawa, Philipp Hennig, Dino Sejdinovic, and Bharath K Sriperumbudur. Gaussian processes and kernel methods: A review on connections and equivalences. *arXiv preprint arXiv:1807.02582*, 2018.

Junpei Komiyama, Akiko Takeda, Junya Honda, and Hajime Shimao. Nonconvex optimization for regression with fairness constraints. In *International conference on machine learning*, pp. 2737–2746. PMLR, 2018.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1088–1097, 2022.

Weishen Pan, Sen Cui, Jiang Bian, Changshui Zhang, and Fei Wang. Explaining algorithmic fairness through fairness-aware causal path decomposition. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 1287–1297, 2021.

Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-García, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 339–355. Springer, 2017.

Alberto Perez-Suay, Pablo Gordaliza, Jean-Michel Loubes, Dino Sejdinovic, and Gustau Camps-Valls. Fair kernel regression through cross-covariance operators. *Transactions on Machine Learning Research*, 2023.

Inioluwa Deborah Raji and Joy Buolamwini. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, 2019.

Craig Saunders, Alexander Gammerman, and Volodya Vovk. Ridge regression learning algorithm in dual variables. In *Proceedings of the Fifteenth International Conference on Machine Learning*, pp. 515–521, 1998.

Marco Scutari, Francesca Panero, and Manuel Proissl. Achieving fairness with a simple ridge penalty. *Statistics and Computing*, 32(5):77, 2022.

Grigory Taturyan, Evgeny Chzhen, and Mohamed Hebiri. Regression under demographic parity constraints via unlabeled post-processing. In *Advances in Neural Information Processing Systems*, volume 37, pp. 117917–117953, 2024.

Shaokui Wei, Jiayin Liu, Bing Li, and Hongyuan Zha. Mean parity fair regression in rkhs. In *International conference on artificial intelligence and statistics*, pp. 4602–4628. PMLR, 2023.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Xin Zhang, Kai Deng, and Qing Mai. Envelopes and principal component regression. *Electronic Journal of Statistics*, 17(2):2447–2484, 2023.

Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.

Han Zhao. Costs and benefits of fair regression. arXiv preprint arXiv:2106.08812, 2021.

A ALGORITHMS

Algorithm 1 Estimating the Predictive–Sensitive Decomposition

1: **Input:** Training data $\{(X_i, S_i, Y_i)\}_{i=1}^n$
2: **Output:** $\widehat{P}_{XY}$, $\widehat{P}_{XSY}$, $\widehat{P}_{XS}$, $\widehat{P}_0$
3: Center $\mathbf{X}$, $\mathbf{S}$, and $\mathbf{Y}$.
4: Estimate the predictor envelope of Y given X and select its dimension $\widehat{u}$ by BIC, yielding bases $\widehat{\Phi}$ and $\widehat{\Phi}_0$.
5: Form the Y-material and Y-immaterial coordinates $Z_Y = \mathbf{X}\widehat{\Phi}, \quad Z_0 = \mathbf{X}\widehat{\Phi}_0$.
6: Within Z_Y, estimate the response envelope relative to S, yielding bases $\widehat{G}_{YS}$ and $\widehat{G}_Y$.
7: Within Z_0, estimate the response envelope relative to S, yielding bases $\widehat{G}_S$ and $\widehat{G}_0$.
8: Construct $\widehat{\Theta}_{XSY} = \widehat{\Phi}\widehat{G}_{YS}, \quad \widehat{\Theta}_{XY} = \widehat{\Phi}\widehat{G}_Y, \widehat{\Theta}_{XS} = \widehat{\Phi}_0\widehat{G}_S, \quad \widehat{\Theta}_0 = \widehat{\Phi}_0\widehat{G}_0$.
9: Set $\widehat{P}_A = \widehat{\Theta}_A\widehat{\Theta}_A^\top, \quad A \in \{XSY, XY, XS, 0\}$.
10: **Return:** $\widehat{P}_{XY}, \widehat{P}_{XSY}, \widehat{P}_{XS}, \widehat{P}_0$.

Algorithm 2 Training FERM at a Target Fairness Budget

1: **Input:** Training data $(\mathbf{X}, \mathbf{S}, \mathbf{Y})$; projectors $\widehat{P}_{XY}$ and $\widehat{P}_{XSY}$ from Algorithm 1; fairness budget $r \in [0, 1]$
2: **Output:** FERM predictor $\widehat{f}_r$
3: Construct $\mathbf{X}_{XY} = \mathbf{X}\widehat{P}_{XY}, \quad \mathbf{X}_{XSY} = \mathbf{X}\widehat{P}_{XSY}$.
4: For$\lambda \geq 0$ Fit

$$(\widehat{\beta}_{XY}(\lambda), \widehat{\beta}_{XSY}(\lambda)) = \underset{\beta_{XY}, \beta_{XSY}}{\arg\min} \ \|\mathbf{Y} - \mathbf{X}_{XY}\beta_{XY} - \mathbf{X}_{XSY}\beta_{XSY}\|_2^2 + \lambda\|\beta_{XSY}\|_2^2.$$

Compute $\widehat{R}_S^2(\widehat{\mathbf{Y}}_\lambda)$.
5: Choose the smallest $\lambda(r)$ satisfying $\widehat{R}_S^2(\widehat{\mathbf{Y}}_{\lambda(r)}) \leq r$.
6: Define $\widehat{f}_r(x) = (\widehat{P}_{XY}x)^\top\widehat{\beta}_{XY}(\lambda(r)) + (\widehat{P}_{XSY}x)^\top\widehat{\beta}_{XSY}(\lambda(r))$.
7: **Return:** $\widehat{f}_r$.

B ENVELOPE REGRESSION MODELS

In this section, we illustrate the envelope regression methods, which aim to improve estimation efficiency by identifying relevant subspaces in the response or predictor space Cook et al. (2010; 2013), forming the foundation of our proposed methodology.

Definition B.1 (M-Envelopes, (Cook et al., 2010))**.** For $M \in \mathbf{R}^{d\times d}$ and $\mathcal{B} \subseteq$ span(M), the $M-$envelope of $\mathcal{B}$, denoted as $\mathcal{E}_M(\mathcal{B})$, is defined as the intersection of all reducing subspaces of M that contains $\mathcal{B}$.

Response Envelope Regression Model We employ the response envelope model, which assumes that a part of the predictors $\mathbf{X}$ remain stochastically constant as the sensitive attributes $\mathbf{S}$ vary. Consider the model:

$$X = B^\top S + e, \tag{9}$$

where e is zero-mean normal noise with covariance matrix $Cov(e) = \Sigma_X$. The response envelope model assumes that the distribution of certain linear combinations of X is invariant to S. This decomposition is formalized using two orthogonal matrices $\Gamma \in \mathbb{R}^{d_X\times m}$ and $\Gamma_0 \in \mathbb{R}^{d_X\times(d_X-m)}$, where $[\Gamma\ \Gamma_0]$ is an orthogonal matrix. Then we have:

(a) $\Gamma_0^\top X \mid S \sim \Gamma_0^\top X$, indicating invariance of $\Gamma_0^\top X$ to S.
(b) $\Gamma^\top X$ is uncorrelated with $\Gamma_0^\top X$ given S.

These conditions imply that $\Gamma_0^\top X$ carries no information about S. Furthermore, Cook et al. (2010) showed that conditions (a) and (b) are equivalent to:

(a') $\text{span}(B) \subset \text{span}(\Gamma)$.

(b') $\Sigma_X = \Sigma_1 + \Sigma_2 = P_\Gamma \Sigma_X P_\Gamma + Q_\Gamma \Sigma_X Q_\Gamma$,

where Σ_X is the covariance of X, P_Γ and $Q_\Gamma = I - P_\Gamma$ are projection operators onto $\text{span}(\Gamma)$ and $\text{span}(\Gamma_0)$, respectively. Based on these conditions, the response envelope model can be expressed as:

$$\begin{aligned} X &= \Gamma \zeta S + e, \\ \Sigma_X &= \Gamma \Omega \Gamma^\top + \Gamma_0 \Omega_0 \Gamma_0^\top, \end{aligned} \tag{10}$$

where $\zeta \in \mathbf{R}^{m \times d_S}$ denotes the coefficients, the columns of Γ is an orthogonal basis for the Σ_X-envelope of $\text{span}(B)$, denoted by $\mathcal{E}_{\Sigma_X}(\text{span}(B))$, and $m = \dim\left(\mathcal{E}_{\Sigma_X}(\text{span}(B))\right)$. The matrices $\Omega \in \mathbb{R}^{m \times m}$ and $\Omega_0 \in \mathbb{R}^{(d_X - m) \times (d_X - m)}$ provide the coordinates of Σ_X with respect to Γ.

The visual representation (Figure 3) illustrates this decomposition for $X \in \mathbb{R}^2$ and $S \in \mathbb{R}$. Each point represents an observation $X = (X_1, X_2)$, colored according to its value of the sensitive attribute S.

- The blue arrow, labeled $\text{Span}(\Gamma)$, indicates the direction in the $X-$space along which the values of X change significantly with S. Observations with similar S values tend to cluster or vary predominantly along this direction.
- The red arrow, labeled $\text{Span}(\Gamma_0)$, represents the direction in the $X-$space along which the values of $\mathbf{X}$ are invariant of S. Variations along this direction show no systematic change with the sensitive attribute.

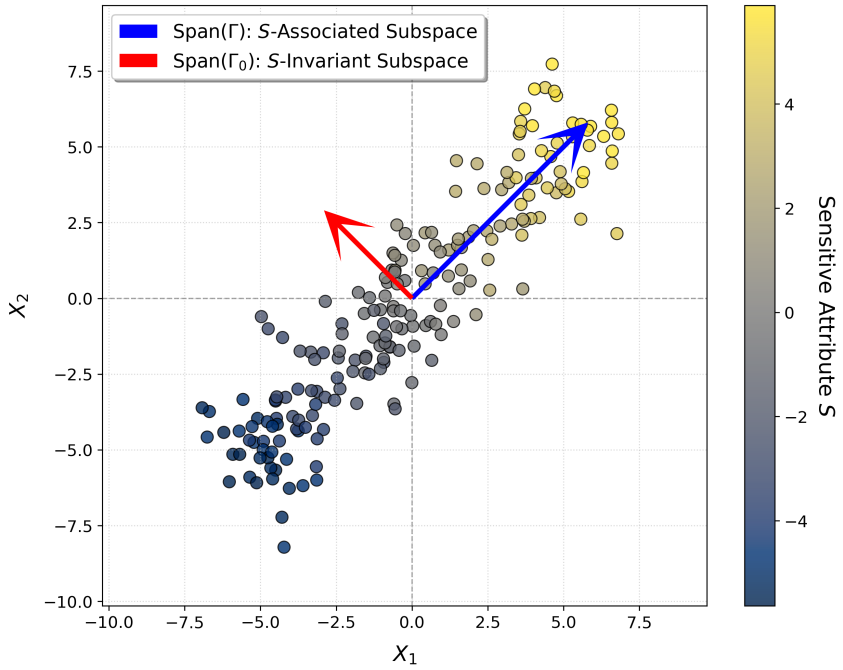


Figure 3: Response X with Sensitive Attribute S and Envelope Subspaces. The data points are colored by the value of S. The blue arrow indicates the direction of the sensitive-associated subspace $\text{Span}(\Gamma)$, while the red arrow indicates the sensitive-invariant subspace $\text{Span}(\Gamma_0)$.

Predictor Envelope Regression Model The predictor envelope model Cook et al. (2013) identifies material and immaterial parts of the predictors X in the regression of Y on X. The model is defined as:

$$Y = \mu_Y + \beta^\top (X - \mu_X) + \epsilon, \tag{11}$$

where ϵ is zero-mean noise. The predictor envelope method seeks a dimension reduction for X by finding the Σ_X-envelope of $\text{span}(\beta^\top)$, denoted by $\varepsilon_{\Sigma_X}(\text{span}(\beta^\top))$. This decomposition divides X into material and immaterial parts, denotes as $P_M X$ and $Q_M X$, which satisfies

(i) $Q_M X$ and $P_M X$ are uncorrelated.

(ii) $Q_M X$ is uncorrelated with Y given $P_M X$.

Let $u = \dim\left(\varepsilon_{\Sigma_X}(\text{span}(\beta^\top))\right)$, and let $\Phi \in \mathbb{R}^{p\times u}$ and $\Phi_0 \in \mathbb{R}^{p\times(p-u)}$ be orthogonal bases for $\varepsilon_{\Sigma_X}(\text{span}(\beta^\top))$ and its orthogonal complement, respectively. The predictor envelope model can then be written as:

$$\begin{aligned} Y &= \mu_Y + \psi^\top \Phi^\top (X - \mu_X) + \epsilon, \\ \Sigma_X &= \Phi \Delta \Phi^\top + \Phi_0 \Delta_0 \Phi_0^\top, \end{aligned} \tag{12}$$

where $\beta = \Phi\psi$, and $\Delta \in \mathbb{R}^{u\times u}$ and $\Delta_0 \in \mathbb{R}^{(p-u)\times(p-u)}$ provide the coordinates of Σ_X with respect to Φ.

B.1 Envelope Estimation Objective

Given the dimension u of an envelope subspace, envelope estimation reduces to solving the constrained optimization problem

$$\hat{\Gamma} = \arg\min_{\Gamma\in\mathbb{R}^{p\times u}\,:\,\Gamma^\top\Gamma = I_u} J_n(\Gamma), \qquad J_n(\Gamma) = \log\left|\Gamma^\top \hat{M}\,\Gamma\right| + \log\left|\Gamma^\top(\hat{M}+\hat{U})^{-1}\Gamma\right|, \tag{13}$$

where $\hat{M} \succ 0$ and $\hat{U} \succeq 0$ are finite-sample estimators of the population matrices M and U. The envelope estimator is then defined as

$$\widehat{\mathcal{E}}_M(U) = \text{span}(\hat{\Gamma}).$$

The constraint $\Gamma^\top\Gamma = I_u$ ensures that Γ has orthonormal columns. Consequently, optimization problem (13) is *non-convex* and takes place on a Stiefel manifold. If we treat the subspace $\mathcal{S} = \text{span}(\Gamma)$ as the argument rather than the basis matrix Γ, then the problem becomes equivalent to a non-convex optimization over the Grassmann manifold (the set of u-dimensional subspaces of $\mathbb{R}^p$).

This formulation highlights that envelope estimation is inherently a problem of manifold optimization. Almost all envelope methods, including those used in regression, prediction, and our fairness-aware adaptation (FERM), are connected through this shared objective structure.

Choice of $\hat{M}$ and $\hat{U}$: The specific form of $\hat{M}$ and $\hat{U}$ depends on whether one is estimating a *response envelope* or a *predictor envelope*. Let $\hat{\Sigma}_Y$, $\hat{\Sigma}_X$, $\hat{\Sigma}_S$, $\hat{\Sigma}_{XY}$, $\hat{\Sigma}_{XS}$and $\hat{\Sigma}_{YX} = \hat{\Sigma}_{XY}^\top$, $\hat{\Sigma}_{SX} = \hat{\Sigma}_{XS}^\top$ denote the (centered) sample covariance and cross-covariance matrices computed from $\{(X_i, S_i, Y_i)\}_{i=1}^n$. As described in Cook & Zhang (2018); Zhang et al. (2023), Then the empirical counterparts used in equation 13 are:

- Response envelope model in (10) :

$$\hat{M} = \hat{\Sigma}_X, \qquad \hat{U} = \hat{\Sigma}_{XS}\,\hat{\Sigma}_{SX}.$$

- Predictor envelope model in (12) :

$$\hat{M} = \hat{\Sigma}_X, \qquad \hat{U} = \hat{\Sigma}_{XY}\,\hat{\Sigma}_{YX}.$$

C Proofs of Results in Section 5

Proof of Theorem 5.1:

Recall that $\hat{\beta}_{\text{env}} := P\hat{\beta}_{\text{OLS}}$, where P is the orthogonal projector onto the subspace spanned by Π_{XSY} and Π_{XY}. Let $\Theta \in \mathbb{R}^{d_X \times k}$ have orthonormal columns forming a basis for this subspace, and let $\Theta_0 \in \mathbb{R}^{d_X \times (d_X - k)}$ form an orthonormal basis for its orthogonal complement, where $k = \dim(\text{span}\{\Pi_{XSY}, \Pi_{XY}\})$.

Since $\beta \in \text{range}(P)$, we have $P\beta = \beta$. Thus,

$$\sqrt{n}\,(\hat{\beta}_{\text{env}} - \beta) = \sqrt{n}\,P(\hat{\beta}_{\text{OLS}} - \beta).$$

Because $v \mapsto Pv$ is linear and continuous, the continuous mapping theorem combined with the assumed asymptotic normality of $\hat{\beta}_{\text{OLS}}$ gives

$$\sqrt{n}\,(\hat{\beta}_{\text{env}} - \beta) \xrightarrow{d} N\big(0,\, P\,\Sigma\,P\big),$$

where $\Sigma := avar(\sqrt{n}\,\hat{\beta}_{\text{OLS}})$. Therefore,

$$avar(\sqrt{n}\,\hat{\beta}_{\text{env}}) = P\Sigma P.$$

Next, express P and $I - P$ in the chosen orthonormal basis:

$$P = \Theta\Theta^\top, \qquad I - P = \Theta_0\Theta_0^\top.$$

With respect to this basis, Σ has the block form

$$\Sigma = [\Theta \quad \Theta_0] \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} [\Theta \quad \Theta_0]^\top, \quad \text{where } \Sigma_{ij} = \Theta_i^\top \Sigma \Theta_j.$$

By construction of the envelope subspace, $\Sigma_{12} = \Sigma_{21}^\top = 0$, so Σ is block-diagonal with respect to $P \oplus P^\perp$. Hence,

$$P\Sigma P = \Theta\,\Sigma_{11}\,\Theta^\top, \qquad \Sigma - P\Sigma P = \Theta_0\,\Sigma_{22}\,\Theta_0^\top.$$

Since $\Sigma \succeq 0$ implies $\Sigma_{22} \succeq 0$, we obtain $\Sigma - P\Sigma P \succeq 0$, i.e.,

$$P\Sigma P \preceq \Sigma.$$

Finally, if Θ_0 is nontrivial and $\Sigma_{22} \succ 0$, then

$$\Sigma - P\Sigma P = \Theta_0\,\Sigma_{22}\,\Theta_0^\top \succ 0,$$

which establishes the strict inequality $P\Sigma P \prec \Sigma$. □

Proof of Proposition 5.3. The result follows by combining the asymptotic normality of $\hat{\beta}_{\text{OLS}}$ with the $\sqrt{n}$-consistency of envelope subspace estimators.

First, by standard linear model theory,

$$\sqrt{n}(\hat{\beta}_{\text{OLS}} - \beta) \xrightarrow{d} N(0, \Sigma),$$

so $\hat{\beta}_{\text{OLS}}$ is $\sqrt{n}$-consistent for β.

Second, we estimate the envelope projector $\hat{P}$ using either the 1D algorithm of Cook & Zhang (2016) or the NIECE algorithm of Zhang et al. (2023). Proposition 6 in Cook & Zhang (2016) and Theorem 1 in Zhang et al. (2023) both establish that the estimated envelope subspace converges at $\sqrt{n}$-rate to the population envelope subspace, i.e.,

$$\|\hat{P} - P\| = O_p(n^{-1/2}).$$

Now decompose

$$\hat{\beta}_{\text{env}}(\hat{P}) - \beta = (\hat{P} - P)\hat{\beta}_{\text{OLS}} + P(\hat{\beta}_{\text{OLS}} - \beta).$$

Since $\hat{\beta}_{\text{OLS}} = O_p(1)$ and $\hat{P} - P = O_p(n^{-1/2})$. The second term is asymptotically normal with mean zero and covariance $P\Sigma P$ by Theorem 5.1.

Therefore,

$$\sqrt{n}\big(\hat{\beta}_{\text{env}}(\hat{P}) - \beta\big) = O_p(1),$$

showing that $\hat{\beta}_{\text{env}}$ is $\sqrt{n}$-consistent. □

Proof of Theorem 5.4. For an independent test pair (X, S), conditional on the training data, $\mathrm{Cov}(\widehat{Y}_{\text{fair}}, S) = \widehat{\beta}_{XY}^{\top}\widehat{P}_{XY}\mathrm{Cov}(X, S)$. Since $\widehat{P}_{XY} \xrightarrow{p} P_{XY}$ and $\widehat{\beta}_{XY} \xrightarrow{p} \beta_{XY}$, the continuous mapping theorem gives

$$\mathrm{Cov}(\widehat{Y}_{\text{fair}}, S) \xrightarrow{p} \beta_{XY}^{\top} P_{XY}\mathrm{Cov}(X, S) = \beta_{XY}^{\top}\mathrm{Cov}(P_{XY}X, S) = 0.$$

Hence, assuming $\mathrm{Var}(\widehat{Y}_{\text{fair}})$ converges to a positive limit and Σ_S is nonsingular, $R_S^2(\widehat{Y}_{\text{fair}}) \xrightarrow{p} 0$. □

Proof of Theorem 5.5 and Corollary 5.6:

Note that

$$\begin{aligned}
\hat{\beta}_{XY}^{\text{fair}} &= \arg\min_{\beta_{XY}} \|\mathbf{Y} - \mathbf{X}_{XY}\beta_{XY}\|_2^2, \\
(\hat{\beta}_{XY}^{\text{OLS}}, \hat{\beta}_{XSY}^{\text{OLS}}) &= \arg\min_{\beta_{XY}, \beta_{XSY}} \|\mathbf{Y} - \mathbf{X}_{XY}\beta_{XY} - \mathbf{X}_{XSY}\beta_{XSY}\|_2^2, \\
(\hat{\beta}_{XY}(\lambda), \hat{\beta}_{XSY}(\lambda)) &= \arg\min_{\beta_{XY}, \beta_{XSY}} \|Y - X_{XY}\beta_{XY} - X_{XSY}\beta_{XSY}\|_2^2 + \lambda\|\beta_{XSY}\|_2^2,
\end{aligned}$$

with penalty $\lambda \geq 0$ applied only to the shared component. Thus,

$$\begin{aligned}
\hat{\beta}_{XY}(\lambda) &= \hat{\beta}_{XY}^{\text{OLS}} = \left(\hat{\Pi}_{XY}^{\top}\mathbf{X}^{\top}\mathbf{X}\hat{\Pi}_{XY}\right)^{-1}\hat{\Pi}_{XY}^{\top}\mathbf{X}^{\top}\mathbf{Y}, \\
\hat{\beta}_{XSY}(\lambda) &= \left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY} + \lambda(r)I\right)^{-1}\mathbf{X}_{XSY}^{\top}\mathbf{Y}.
\end{aligned}$$

Using Sherman-Morrison-Woodbury identity for matrices, we have

$$\left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY} + \lambda(r)I\right)^{-1} = \left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY}\right)^{-1} - \left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY}\right)^{-1}\left(\frac{1}{\lambda(r)}\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY} + I\right)^{-1}.$$

Thus,

$$\begin{aligned}
\hat{\beta}_{XSY}(\lambda) &= \left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY}\right)^{-1}\mathbf{X}_{XSY}^{\top}\mathbf{Y} - \left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY}\right)^{-1}\left(\frac{1}{\lambda(r)}\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY} + I\right)^{-1}\mathbf{X}_{XSY}^{\top}\mathbf{Y} \\
&= \hat{\beta}_{XSY}^{\text{OLS}} - \left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY}\right)^{-1}\left(\frac{1}{\lambda(r)}\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY} + I\right)^{-1}\mathbf{X}_{XSY}^{\top}\mathbf{Y}.
\end{aligned}$$

This allows us to obtain the following representation, using $\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY} = nI$,

$$\begin{aligned}
\hat{Y}_{\lambda}(X, S) &= \hat{\beta}_{XY}(\lambda)^{\top}X_{XY} + \hat{\beta}_{XSY}^{\top}(\lambda)X_{XSY} \\
&= (\hat{\beta}_{XY}^{\text{OLS}})^{\top}X_{XY} + (\hat{\beta}_{XSY}^{\text{OLS}})^{\top}X_{XSY} - (\left(\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY}\right)^{-1}\left(\frac{1}{\lambda(r)}\mathbf{X}_{XSY}^{\top}\mathbf{X}_{XSY} + I\right)^{-1}\mathbf{X}_{XSY}^{\top}\mathbf{Y})^{\top}X_{XSY} \\
&= (\hat{\beta}_{XY}^{\text{OLS}})^{\top}X_{XY} + (\hat{\beta}_{XSY}^{\text{OLS}})^{\top}X_{XSY} - \frac{1}{n}\frac{\lambda(r)}{n + \lambda(r)}(\mathbf{X}_{XSY}^{\top}\mathbf{Y})^{\top}X_{XSY} \\
&= (\hat{\beta}_{XY}^{\text{OLS}})^{\top}X_{XY} + (\hat{\beta}_{XSY}^{\text{OLS}})^{\top}X_{XSY} - \frac{\lambda(r)}{n + \lambda(r)}(\hat{\beta}_{XSY}^{\text{OLS}})^{\top}X_{XSY}
\end{aligned}$$

On rearranging, we get

$$\hat{Y}_{\lambda} = \frac{n}{n + \lambda(r)}\hat{Y}_{\text{OLS}} + \left(1 - \frac{n}{n + \lambda(r)}\right)\hat{Y}_{\text{fair}}, \qquad \lambda(r) \geq 0$$

Let $\omega(n, r) = \frac{n}{n + \lambda(r)}$. Thus,

$$Y - \hat{Y}_{\lambda} = Y - \left[\omega(n, r)\,\hat{Y}_{\text{OLS}} + (1 - \omega(n, r))\,\hat{Y}_{\text{fair}}\right] = (Y - \hat{Y}_{\text{OLS}}) + (1 - \omega(n, r))(\hat{Y}_{\text{OLS}} - \hat{Y}_{\text{fair}})$$

Hence

$$\begin{aligned}
E\left[Y - \hat{Y}_{\lambda}\right]^2 = E\left[Y - \hat{Y}_{\text{OLS}}\right]^2 &+ (1 - \omega(n, r))^2 E\left[\hat{Y}_{\text{OLS}} - \hat{Y}_{\text{fair}}\right]^2 \\
&+ 2(1 - \omega(n, r))E\left[(Y - \hat{Y}_{\text{OLS}})(\hat{Y}_{\text{OLS}} - \hat{Y}_{\text{fair}})\right].
\end{aligned}$$

But note that $(Y - \hat{Y}_{\mathrm{OLS}})$ and $(\hat{Y}_{\mathrm{OLS}} - \hat{Y}_{\mathrm{fair}})$ are orthogonal, hence uncorrelated. Since, $E(e) = 0$, the cross-term above vanishes.

That is,

$$\mathrm{MSE}_\lambda = \mathrm{MSE}_{\mathrm{OLS}} + (1 - \omega(n, r))^2 \, \mathrm{MSE}_{\mathrm{fair}}.$$

showing exactly how FERM interpolates between the two models via the tuning parameter $\lambda(r)$.

D When Are Linear Envelopes Appropriate? Nonlinear Generalizations of FERM

The FERM framework is built on an explicit four-way linear decomposition of the predictor space into orthogonal components capturing predictive, sensitive, shared, and residual variation. This decomposition, combined with the closed-form characterization of the fairness-utility trade-off, is only achievable under a linear model for Y. Linear envelopes therefore serve as the mathematically tractable foundation on which our theoretical guarantees rest.

This structure mirrors the classical development of PCA: the linear PCA formulation provides interpretability, closed-form solutions, and a clear geometric foundation; nonlinear variants such as kernel PCA arise only after the linear theory is firmly established. Our contribution plays an analogous role in fairness-aware regression: it formalizes the linear envelope setting as a principled, interpretable baseline with guarantees on efficiency and fairness.

Nevertheless, linear envelopes are not the end point. FERM is constructed with a modular and flexible design, which enables a wide range of possible extensions. These include incorporating nonlinear modeling frameworks, introducing alternative or more sophisticated penalization schemes, applying different definitions of fairness, and adapting the method to various types of response variables using generalized linear models. A key strength of FERM lies in the clear separation between model selection (i.e., the choice of $\lambda(r)$) and model estimation, which facilitates independent modification of either stage. This separation allows FERM to draw from and integrate with a broad body of established techniques in statistical modeling, offering adaptability to different fairness-aware regression settings.

D.1 Nonlinear Regression Models

FERM can be extended to handle nonlinear relationships by incorporating kernel methods, in a similar fashion as suggested by Komiyama et al. (2018) and Scutari et al. (2022). Specifically, the model can be fitted into transformed feature spaces $Z_\Gamma(\hat{\Gamma}^T X)$ and $Z_{\Gamma_0}(\hat{\Gamma}_0^T X)$, obtained through positive definite kernel functions, similar to the approach in Komiyama et al. (2018). Applying the kernel trick in conjunction with ridge penalization leads to a kernel ridge regression variant of FERM (cf. Saunders et al. (1998)), which can be estimated efficiently using techniques such as those proposed by Zhang et al. (2015). Moreover, since kernel ridge regression is closely related to Gaussian process regression, this extension naturally opens the door to Bayesian nonparametric variants of FERM using Gaussian processes, as discussed in Kanagawa et al. (2018).

D.2 Extensions to Deep Learning Models

The FERM framework can also be extended to deep learning architectures by leveraging the envelope decomposition as a structured preprocessing step. Specifically, the decomposition of the predictor space into orthogonal components – associated with the response (Y), the sensitive attribute (S), their shared variation, and residuals – provides a principled way to separate and control different sources of variation prior to feeding them into a neural network. One approach is to project the raw inputs onto the learned subspaces, and then use only the subspace orthogonal to the sensitive attributes as input to the downstream model. This can help ensure that the representations learned by the network are less entangled with sensitive information, improving fairness in predictions.

Taken together, the linear envelope provides the theoretical scaffold, while the nonlinear extensions offer paths toward greater flexibility. This Appendix clarifies when the linear formulation is appropriate (e.g., when interpretability, closed-form decomposition, or theoretical guarantees are desired) and how the nonlinear generalizations can be systematically constructed.

E Empirical validation of Theoretical results

In this section, we present additional simulation results to support our theoretical guarantees. We compare three regression settings:

$$\begin{aligned}
\mathcal{M}_1 &: Y \sim X_{XY} + X_{XSY} \qquad \text{(Env_Xy_Xys)},\\
\mathcal{M}_2 &: Y \sim X_{XY} + X_0 + X_{XSY} \qquad \text{(Env_XyX0_Xys)},\\
\mathcal{M}_3 &: Y \sim \underbrace{X_{XY} + X_0}_{U} + \underbrace{X_{XSY} + X_{XS}}_{S} \qquad \text{(FRRM)},
\end{aligned}$$

where "$\sim$" indicates that Y is regressed on the specified terms.

Our evaluation focuses on three key objectives:

- **Coefficient convergence in the unconstrained case.** When $r = 1$, Proposition 5.3 shows that both envelope estimators ($\mathcal{M}_1$ and $\mathcal{M}_2$) converge to the ordinary least-squares estimator (FRRM in $\mathcal{M}_3$). We verify this by computing the L_2 distance between the envelope coefficients and the FRRM coefficients.
- **Fairness: convergence of covariance to zero.** When $r = 0$, Theorem 5.4 implies that the covariance between predictions and the sensitive component converges to zero in probability for all three models. We evaluate this by computing $\mathrm{Cov}(\widehat{Y}, S)$ on held-out test data.
- **Efficiency gain through subspace reduction.** The envelope model removes the redundant component X_0, which has no impact on Y. Eliminating this noise-inflating subspace reduces the variance of the estimated coefficients, as shown in Theorem 5.1, demonstrating the efficiency benefit of envelope-based dimension reduction.

We conduct experiments under two settings:

$$\begin{aligned}
d_X = 30, \qquad d_S = 20, \qquad \dim(\Pi_{XY}) = \dim(\Pi_{XSY}) = \dim(\Pi_{XS}) = 5,\\
d_X = 100, \qquad d_S = 20, \qquad \dim(\Pi_{XY}) = \dim(\Pi_{XSY}) = \dim(\Pi_{XS}) = 10.
\end{aligned}$$

For each setting, we vary the sample size

$$n \in \{200, 500, 1000, 2000, 5000, 10000, 20000\},$$

and the unfairness control level

$$r \in \{0, 0.1, 0.3, 0.5, 0.7, 0.9, 1\}.$$

All configurations are replicated 50 times. The corresponding results are summarized below.

From Figure 4, we observe that in both settings ($d_X = 30$ and $d_X = 100$) and in the unconstrained case ($r = 1$), the estimated coefficients from $\mathcal{M}_1$ and $\mathcal{M}_2$ converge toward those of $\mathcal{M}_3$, confirming the consistency result established in Proposition 5.3.

Figures 5 and 6 further show that when $r = 0$, all three models yield fully fair estimators: the covariance between the predicted value and the sensitive component decreases steadily as the sample size increases, in accordance with Lemma **??**. In contrast, this covariance does not vanish when $r = 0.5$, since the model is not required to be fair and the predictions may legitimately depend on the sensitive direction.

Finally, Figure 7 demonstrates the efficiency gain obtained from removing the redundant subspace X_0. Across all dimensions and all fairness levels, the average variance ratio between $\mathcal{M}_2$ (which includes X_0) and $\mathcal{M}_1$ (which excludes X_0) is consistently greater than one. This confirms that eliminating components that do not contribute to the response reduces estimator variance and yields a more efficient model.

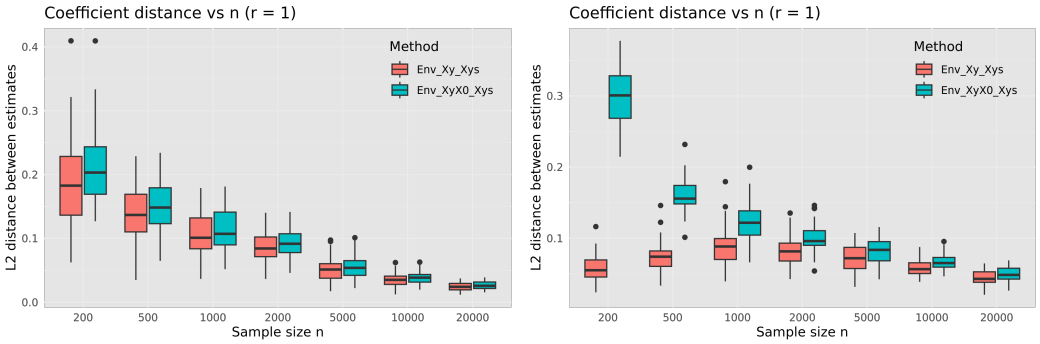


Figure 4: **Coefficient convergence under the unconstrained case** ($r = 1$)**.** The plots show the L_2 distance between the estimated coefficients of the two envelope models ($\mathcal{M}_1$ and $\mathcal{M}_2$) and the FRRM estimator $\mathcal{M}_3$ across different sample sizes, under two dimensional settings ($d_X = 30$ and $d_X = 100$).

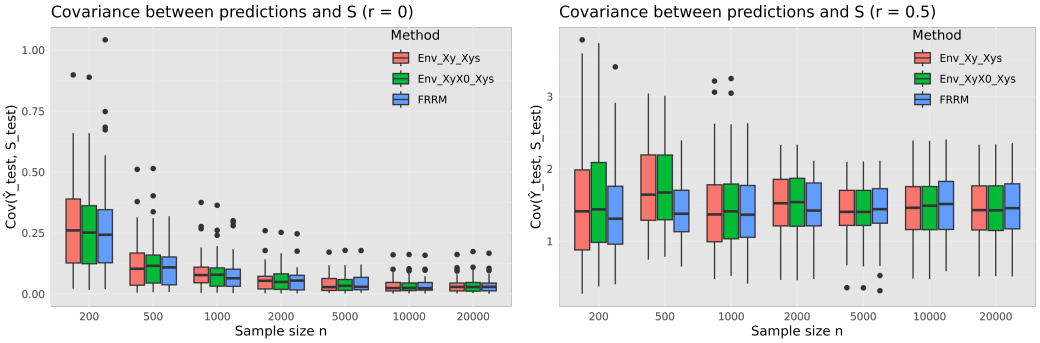


Figure 5: **Covariance between predictions and the sensitive component when** $d_X = 30$**.** Left: When $r = 0$, all three models produce fully fair estimators, and the empirical covariance $\mathrm{Cov}(\widehat{Y}, S)$ decreases toward zero as the sample size increases. Right: When $r = 0.5$, the covariance does not converge to zero since the model is only partially fair.

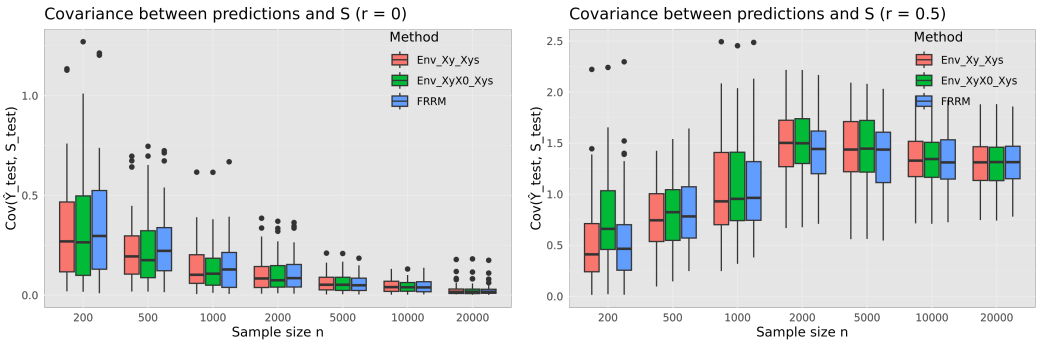


Figure 6: **Covariance between predictions and the sensitive component when** $d_X = 100$**.**

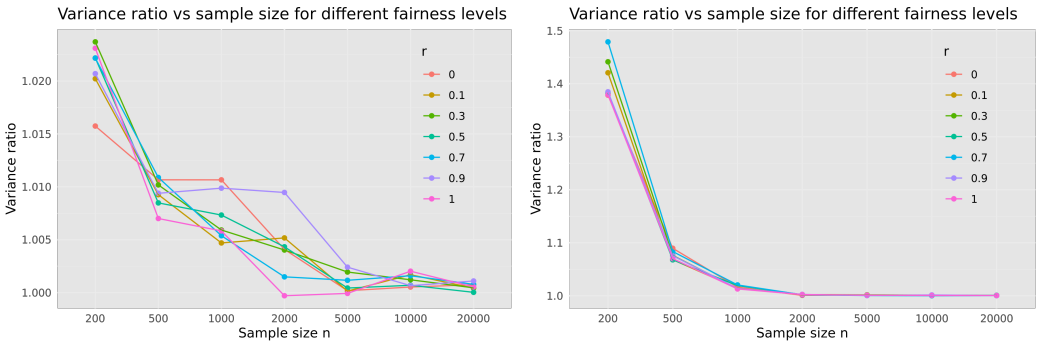
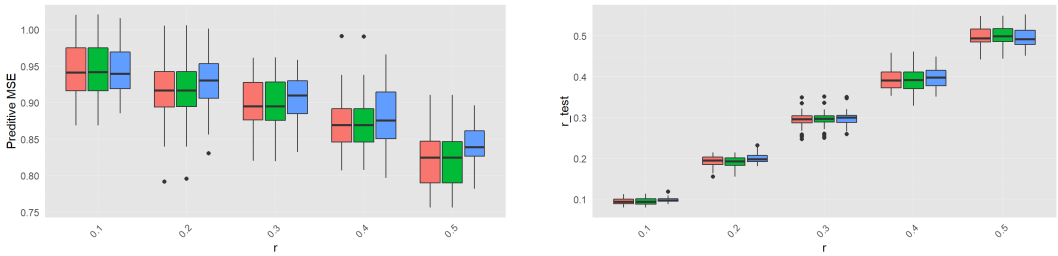


Figure 7: **Efficiency comparison via variance ratios of envelope estimators.** For each dimension setting ($d_X = 30$ (left panel) and $d_X = 100$ (right panel)), the plots show the ratio of the average coordinatewise variance of the estimator from $\mathcal{M}_2$ (which includes the redundant subspace X_0) to that from $\mathcal{M}_1$ (which excludes X_0). Across all fairness levels and sample sizes, this ratio is consistently greater than 1.



(3a) Simulation setting 3: Predictive MSE

(3b) Simulation setting 3: Unfairness on test data

Figure 8: Left panel: Predictive MSE for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) for various unfairness levels r; lower values are better. Right panel: Unfairness levels on test data (r_{test}) for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) at varying unfairness levels r. Simulation settings are described in Table 1.

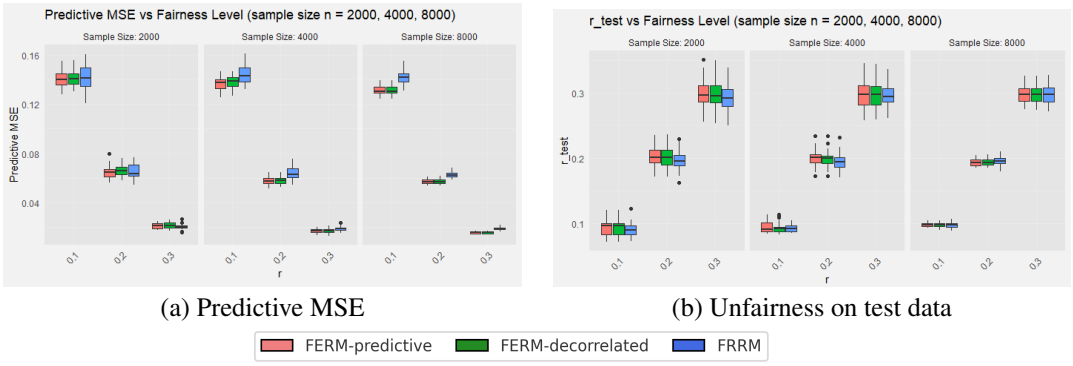
F Experiments

F.1 Poisson Setting

In Figure 8, we provide simulation results for the Poisson setting (3) given in Table 1.

F.2 Additional Real-world data

In another real data application, we use U.S. Health Insurance Dataset (available here). In this analysis, the sensitive attributes include gender, age, medical history, family medical history, and region. After creating dummy variables for all categorical variables, we finally have 15 features in S and 15 features in X. We evaluated the performance under various settings with $r \in 0.1, 0.2, 0.3$ and sample sizes $n \in 2000, 4000, 8000$. The real data was perturbed and each configuration was replicated 30 times. The results are provided in Figure 9. Results demonstrate that FERM models are superior to FRRM approach in most of the settings.



(a) Predictive MSE (b) Unfairness on test data

Figure 9: Left panel: Predictive MSE for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) for unfairness levels $r = \{0.1, 0.2, 0.3\}$; lower values are better. Right panel: Unfairness levels on test data (r_{test}) for FRRM (in blue), FERM-predictive (in red), and FERM-decorrelated (in green) at unfairness levels $r = \{0.1, 0.2, 0.3\}$.

G Implementation Details of Algorithms

G.1 Baseline Implementation Details: Fair Ridge Regression Model (FRRM)

For the Fair Ridge Regression Model (FRRM; Scutari et al. (2022)), we follow the authors' implementation exactly as recommended. All FRRM results were obtained using the `frrm()` function from the `fairml` package in `R`, with default settings unless otherwise specified. FRRM requires the user to specify a target unfairness budget r, and the algorithm internally determines the corresponding ridge penalty λ to satisfy this constraint. No additional hyperparameter tuning is exposed to the user. In all our experiments (Section 6), we evaluated FRRM at unfairness levels $r = \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For reproducibility, we ensured that FRRM was fit using the same training splits, sensitive variable encoding, and preprocessing steps applied to FERM. No post-hoc adjustments were made to FRRM outputs.

G.2 Envelope-Based Decomposition

Numerical optimization: The envelope objective is non-convex and solved on Stiefel/Grassmann manifolds. In practice, we use manifold optimization routines from the `R` package `Renvlp`.

Practical computation of λ: In practice, λ is found via a line search (e.g., bisection) over a grid of candidate values until the fairness constraint $R^2_S(\lambda) \leq r$ is met. This ensures the fitted model interpolates smoothly between the fair and unconstrained extremes.

Software: All algorithms are implemented in `R`. For cross-validation and BIC, we rely on existing libraries for model selection. Ridge penalties are implemented using standard linear algebra solvers.

G.3 Computational Details:

We conduct all our experiments on a Gentoo Linux server with an Intel Xeon E5-2683 v4 @ 2.10GHz CPU and 251GB of RAM. No GPU is involved.

H Additional Discussion of Related Work

For completeness, we briefly situate FERM relative to two areas of the fairness literature that also discuss direct and indirect effects.

Causal literature. Works in the causal literature defines "direct" and "indirect" effects in terms of causal pathways under a structural causal model (e.g., Chiappa (2019); Pan et al. (2021)). These approaches rely on strong assumptions (e.g., valid mediators, no hidden confounding, identifiable counterfactuals) that are not available in our supervised learning setting with multivariate continuous S. Our decomposition is geometric - based on orthogonal subspaces of the X -space - rather than causal.

Fair representation learning. Methods from the fair representation–learning literature (e.g., adversarial learning, VFAE, pre-processing decorrelation) are almost exclusively developed for binary or categorical sensitive attributes and rely on iterative optimization procedures (e.g., Creager et al. (2019); Zemel et al. (2013); Liu et al. (2022)). Their loss functions and fairness objectives (e.g., adversarial equality of odds, demographic parity constraints, mutual-information penalties) are tailored to discrete protected groups and are not readily extendable to the multivariate continuous S that we consider. Moreover, these approaches do not yield explicit fairness–utility trade-offs or closed-form decompositions, making them unsuitable as baselines for our theoretical framework.

I Limitations

While FERM provides a principled framework for fairness-aware regression, several limitations should be noted:

Linearity of subspace decomposition: FERM assumes that both material and immaterial variation can be captured through *linear* subspaces of the predictor space. In many real-world applications, sensitive attributes may influence Y through nonlinear interactions with X, in which case a purely linear envelope decomposition may be too restrictive. Extending the methodology to nonlinear settings (e.g., kernelized or deep envelope methods) remains an important direction for future work.

Dependence on subspace estimation: The validity of FERM relies on accurately estimating the envelope subspaces. Although the 1D and NIECE algorithms provide $\sqrt{n}$-consistent estimators, in finite samples the estimated subspaces may deviate from their population counterparts, especially when d_X is large relative to n. This may lead to imperfect fairness guarantees in small samples.

Choice of tuning parameters: FERM requires selecting the envelope dimensions and the interpolation parameter. While BIC and cross-validation are standard, their stability can vary across datasets. Automatic and robust criteria for selecting these parameters remain an open area of research.

Fairness notion: Our fairness control is expressed through linear-association criterion measured by R^2_S between predictions and sensitive attributes. Other fairness notions (e.g., counterfactual fairness, equalized odds) are not directly captured. Extending the envelope framework to these broader definitions requires additional methodological development.

Computational cost: Envelope estimation involves non-convex optimization on Stiefel or Grassmann manifolds. Although efficient algorithms like NIECE exist, the procedure is more computationally intensive than standard regression, and scalability to very high-dimensional predictors may be challenging without additional structural assumptions.

Overall, while these limitations highlight directions for future work, FERM demonstrates that subspace decomposition provides a tractable and interpretable pathway to fairness-aware regression.