Auditing for Demographic Bias in Opaque Rankings

Antonio Ferrara
Intesa Sanpaolo AI Research, Turin, Italy
antonio.ferrara@intesasanpaolo.com

Carlo Abrate
Intesa Sanpaolo AI Research, Turin, Italy
carlo.abrate@intesasanpaolo.com

Fabio Vitale
Intesa Sanpaolo Innovation Center, Turin, Italy
fabio.vitale@intesasanpaolo.com

Francesco Bonchi
Intesa Sanpaolo AI Research, Turin, Italy
francesco.bonchi@intesasanpaolo.com

ABSTRACT

Auditing algorithmic fairness is a critical challenge in high-stakes domains like hiring and credit scoring, especially given the intrinsic opacity of algorithmic decision-making systems. In this paper, we tackle the following problem: given a ranking of individuals, how can we assess whether the order is driven by protected attributes (e.g., gender or race) rather than task-relevant features, under a strict black-box assumption where the ranking mechanism cannot be queried?

Building on kernel conditional independence and partial distance correlation, we introduce Condor, a model-agnostic audit framework. Condor first residualizes the ranking and protected attributes with respect to observables in a reproducing kernel Hilbert space. It then quantifies the remaining association via distance correlation on the residualized embeddings, returning a normalized effect-size score. This procedure captures general nonlinear dependencies without assuming access to latent scores, requires no hyperparameter fine-tuning, and naturally accommodates mixed continuous and categorical data. From Condor's effect-size score, we derive a hypothesis test for conditional independence. By combining this test with an unconditional independence test, auditors can achieve a comprehensive causal understanding of the protected attributes' influence.

We validate our proposal on real and semi-synthetic datasets with controlled influence of the protected attributes on the ranking: our method reliably detects the influence of protected attributes, outperforming established statistical auditing baselines.

PVLDB Reference Format:
Antonio Ferrara, Carlo Abrate, Fabio Vitale, and Francesco Bonchi. Auditing for Demographic Bias in Opaque Rankings. PVLDB, 14(1): XXX-XXX, 2020.
doi:XX.XX/XXX.XX

PVLDB Artifact Availability:
The source code, data, and/or other artifacts have been made available at https://github.com/Ambress92/Audit_ranking.

Proceedings of the VLDB Endowment, Vol. 14, No. 1 ISSN 2150-8097.
doi:XX.XX/XXX.XX

1 INTRODUCTION

Algorithmic rankings govern high-stakes decisions in hiring, credit scoring, and college admissions. Ensuring these rankings are not unduly influenced by protected demographic attributes (e.g., gender or race) is a critical requirement often mandated by fairness regulations [30, 43, 64]. However, auditing these systems is complicated by their opacity: auditors typically observe only the input features and the output ranking, without access to the internal scoring logic or training data. Several authors have explored this *black-box auditing* challenge in various forms. Adler et al. [1] have introduced a framework for detecting indirect influence in black-box models using controlled perturbations and influence functions. More recently, Cherian and Candès [9] have proposed a general statistical testing framework for model-agnostic fairness auditing, leveraging disparity metrics and bootstrapping techniques. In parallel, methods based on counterfactual reasoning – such as [52] – aim to explain individual predictions and identify fairness violations through feature perturbation, even without access to internal model parameters. Despite these advances, existing approaches tend to focus either on binary classification tasks or on testing specific hypotheses (e.g., the presence or absence of group disparity), rather than assessing whether a ranking depends on protected attributes.

In this paper, we tackle the problem of detecting whether an observed ranking R depends on protected attributes Z after accounting for legitimate, task-relevant attributes X (see Figure 1). Formally, this requires testing the conditional independence hypothesis:

$$H_0 : Z \perp\!\!\!\perp R \mid X$$

To solve this problem, we introduce Condor (*CONditional DistancecOrrelation for Rankings*), a model-agnostic audit framework that quantifies the residual dependence of a ranking on protected attributes (see Figure 1). Our primary contributions are:

(1) The Condor method: We combine kernel-based residualization with distance correlation. By projecting the ranking and protected attributes into a Reproducing Kernel Hilbert Space (RKHS; see, e.g., [48]), we "regress out" the influence of task-relevant features X. Measuring the association between the remaining residuals using distance correlation yields the Protected Attribution Score (PAS), a normalized effect size in $[0, 1]$ capable of capturing complex non-linear dependencies without hyperparameter tuning. We theoretically characterize the PAS showing that it ensures exact null alignment under conditional independence and exhibits predictable local sensitivity to emerging biases.

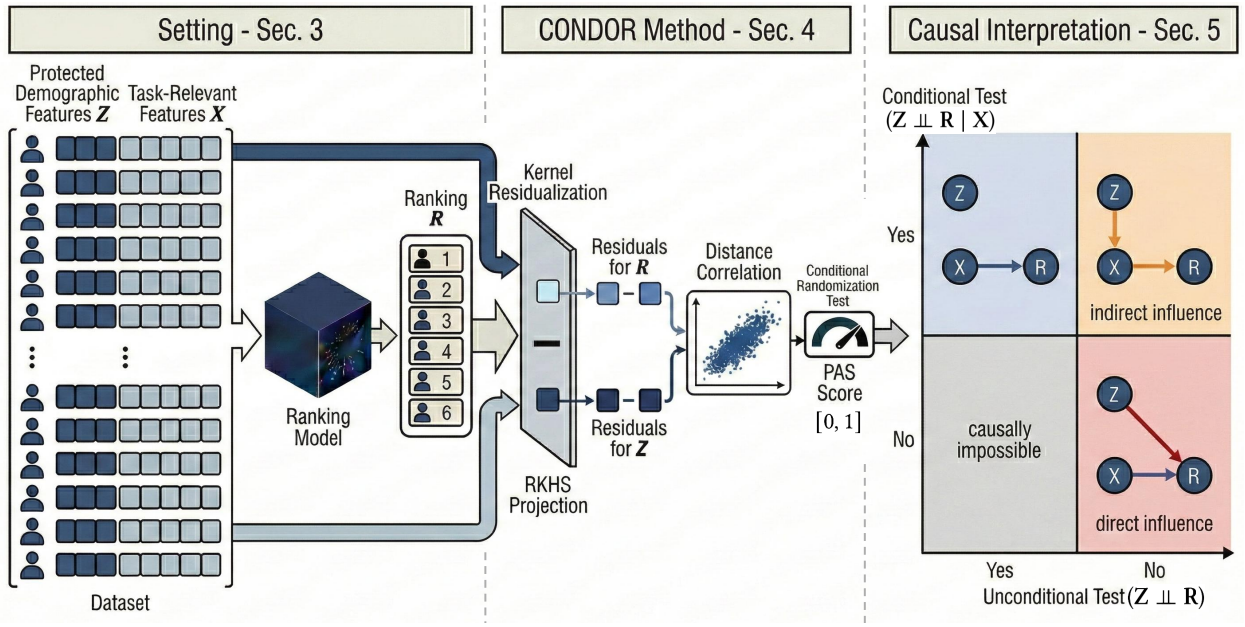

Figure 1: Condor Pipeline and Causal Diagnostic Framework. (Left) We observe a dataset of n individuals with protected attributes Z and task-relevant features X, alongside an opaque ranking R (§3). (Center) Condor audits this black box by projecting R and Z into a Reproducing Kernel Hilbert Space (RKHS) to residualize ("regress out") the legitimate influence of X. Distance correlation on these residuals yields the Protected Attribution Score (PAS), which is then calibrated via a Conditional Randomization Test to produce a robust p-value for the null hypothesis $Z \perp\!\!\!\perp R \mid X$ (§4). (Right) Pairing this conditional test with an unconditional baseline ($Z \perp\!\!\!\perp R$) creates a 2×2 causal matrix, allowing auditors to distinguish between a passed audit (no influence), indirect (mediated by task features) influence, and direct discrimination (§5).

As raw effect sizes depend heavily on dataset scale and variance, PAS values across different datasets are not directly comparable. To provide a definitive binary verdict, Condor calibrates the PAS against a "no residual effect" null hypothesis via a Conditional Randomization Test (CRT), converting the score into an interpretable one-sided p-value [12, 17, 35].

(2) Causal interpretation: By combining our conditional test ($Z \perp\!\!\!\perp R \mid X$) with an unconditional test ($Z \perp\!\!\!\perp R$), we provide a causal diagnostic framework. This allows auditors to distinguish between cases where the audit is passed (no influence), and cases where the protected attributes exert either a direct influence (violating conditional independence) or an indirect influence (mediated by task features) on the ranking.

(3) Empirical assessment: We validate Condor in 3 settings: (*i*) Both input data and ranking are syntectic, so that we can control the extent to which the protected information influences the ranking; (*ii*) Real-world datasets and grey-box ranking obtained by mixing a merit term with a tunable protected term, so the same protected-signal strength parameter governs *by design* the share of protected influence while preserving real-world feature distributions; (*iii*) Real-world dataset with black-box ranking.

We compare our method against three established baselines: a kernel-based conditional test (KCI [65]), a distance correlation based technique (pdCor [53]), and an information-theoretic conditional mutual information estimator (CMI – KSG k-NN [32]). Our experiments confirm the suitability of Condor as a conditional independence test and its superiority over the baselines.

2 RELATED WORK

Besides the topic of black-box auditing, which we already covered in Section 1, this work is related – on the technical level – to non-parametric measures of conditional dependence, while – at higher level – it collocates within the literature about fairness in ranking.

Measures of conditional independence. Kernel-based tests of conditional independence (e.g., KCI [65]) first residualize Z and R with respect to X in an RKHS via a regularized projection, and then quantify the remaining association using the Hilbert–Schmidt Independence Criterion [25], a kernel-based dependence measure that equals zero under independence [26]. Condor draws inspiration from this kernel-based framework but translates the kernel–residualization idea of KCI into a distance–correlation setting, thus producing a normalized effect size.

A complementary line relies on distance correlation dCor[54], which characterizes independence for multivariate vectors; its partial variant pdCor [53] provides a closed-form, distance-based effect size for residual association between R and Z after accounting for X. Building on this line, Condor augments pdCor with an RKHS

residualization step and kernel–induced distances. This hybrid design preserves the geometric invariances and closed-form interpretability of PDCOR, while adding the conditioning strength of kernel methods to capture complex (nonlinear, interaction) effects of the observables. Other conditional independence approaches, such as conditional distance correlation (CDCor) [57] and conditional-mutual-information (CMI) estimators [33], yield measures that are zero *if and only if* conditional independence holds, but they require estimating conditional densities/expectations and introduce multiple tuning knobs (kernels, bandwidths, neighborhood sizes), which can be brittle in black-box audits.

Relative to these conditional independence tools, CONDOR sidesteps the estimation of conditional objects altogether and limits its tuning to robust defaults (kernel family and a small ridge). In practice, this yields a normalized effect size (PAS) that is more sensitive to residual dependence than plain PDCOR when the influence of the observables is nonlinear, and it accommodates mixed continuous/categorical data through product kernels with stable defaults.

Fair ranking. Fairness in algorithmic ranking has attracted a growing attention [30, 36, 41, 43, 64]. Much of such literature frames the problem as one of equitable resource allocation. With a few exceptions that focus on individual fairness [5, 15, 46], or that tackle individual and group fairness jointly [22, 63], the bulk of the literature on fairness in ranking [4, 8, 24, 50, 60–62] and learning-to-rank [16, 37, 51] deals with *group fairness* along the lines of *demographic parity* [15] or *equal opportunity* [28]. This typically requires defining *a priori* demographic groups that have to be protected from discrimination, as enforced by legal norms [18, 19]. This is typically casted as the problem of maximizing utility under a *group-fairness constraint*, that requires a minimum fraction of individuals from the protected group to be included in the top-k positions [8, 50, 62]. Such methods are inherently prescriptive: they aim to devise algorithms that produce fair rankings by balancing utility and fairness, often relying on explicit models of attention, such as the position-based discount functions used in Discounted Cumulative Gain (DCG) [31]. They face key challenges in defining the individual merit, the strength of group-fairness constraints, or even which groups should be protected in the first place [40].

Our work adopts a fundamentally different, *descriptive* posture: we do not prescribe how to rank fairly but instead provide a post-hoc, forensic tool to audit an *existing*, opaque ranking. The objective is not to manage exposure allocation but to quantify the *residual influence* of protected attributes on the observed ranking.

3 PRELIMINARIES

We are given n individuals each described by an attribute vector $\mathbf{v}_i = (\mathbf{x}_i, \mathbf{z}_i)$ for $i \in \{1, \dots, n\}$, where $\mathbf{x}_i$ are *observable task-relevant* attributes and $\mathbf{z}_i$ are *protected demographic* attributes. We are also given a deterministic ranking of the n individuals: i.e., a bijective function $r : [n] \to [n]$, such that $r(i)$ is the position of the individual i in the observed ranking, with $r(i) = 1$ being the top (best) position. The positions in the ranking are hereafter viewed as stacked into the vector $\mathbf{r} := (r(1), \dots, r(n))^\top$. We represent task-relevant and protected attributes by the matrices $\mathbf{X} \in \mathbb{R}^{n \times d_o}$ and $\mathbf{Z} \in \mathbb{R}^{n \times d_p}$, whose i-th rows are $\mathbf{x}_i^\top$ and $\mathbf{z}_i^\top$, respectively. As our target is a property of the underlying data-generating mechanism, not of a single dataset, we adopt a *population-level* viewpoint, i.e., we use random variables X, Z, R for task-relevant attributes, protected attributes, and the ranking respectively, and study the conditional–independence hypothesis $H_0 : Z \perp\!\!\!\perp R \mid X$.

For sake of completeness and clarity of treatment, we next recall the fundamental definition of a kernel function and its associated spaces, which will be used in the next section, pointing the reader to standard texts for comprehensive treatments [34, 48, 49].

DEFINITION 1 (KERNEL FUNCTION). *Let $\mathcal{X}$ be a non-empty set. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a **kernel** if there exists a Hilbert space $\mathcal{H}$ (a feature space) and a feature map $\phi : \mathcal{X} \to \mathcal{H}$ such that for all $\mathbf{u}, \mathbf{u}' \in \mathcal{X}$, the kernel computes the inner product in $\mathcal{H}$:*

$$k(\mathbf{u}, \mathbf{u}') = \langle \phi(\mathbf{u}), \phi(\mathbf{u}') \rangle_{\mathcal{H}}. \tag{1}$$

In our setting, $\mathcal{X}$ can be, e.g., the space of task attributes $\mathbb{R}^{d_o}$ or protected attributes $\mathbb{R}^{d_p}$. To rigorously ground this feature space, we define a Reproducing Kernel Hilbert Space, which ensures that point evaluations are well-behaved.

DEFINITION 2 (REPRODUCING KERNEL HILBERT SPACE (RKHS)). *Let $\mathcal{H}$ be a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. $\mathcal{H}$ is a **Reproducing Kernel Hilbert Space** if, for all $\mathbf{x} \in \mathcal{X}$, the linear evaluation functional $L_{\mathbf{x}} : \mathcal{H} \to \mathbb{R}$ defined by $L_{\mathbf{x}}(f) = f(\mathbf{x})$ is bounded. By the Riesz representation theorem [34], this implies that for every $\mathbf{x} \in \mathcal{X}$, there exists a unique function $k_{\mathbf{x}} \in \mathcal{H}$ (the representer of evaluation) such that the reproducing property holds:*

$$f(\mathbf{x}) = \langle f, k_{\mathbf{x}} \rangle_{\mathcal{H}}, \quad \forall f \in \mathcal{H}. \tag{2}$$

By the Moore-Aronszajn theorem [48], every positive-definite kernel k is associated with a unique RKHS where $k(\mathbf{u}, \mathbf{u}') = \langle k_{\mathbf{u}}, k_{\mathbf{u}'} \rangle_{\mathcal{H}}$. A standard choice for continuous variables or rankings is the radial basis function (RBF) kernel:

$$k(\mathbf{u}, \mathbf{u}') = \exp\left(-\frac{\|\mathbf{u} - \mathbf{u}'\|^2}{2\sigma^2} \right), \tag{3}$$

where $\sigma > 0$ is the bandwidth parameter. For categorical variables, a common choice is the exponentiated Hamming kernel:

$$k(\mathbf{u}, \mathbf{u}') = \exp\left(-\gamma \cdot d_H(\mathbf{u}, \mathbf{u}') \right), \tag{4}$$

where $d_H(\mathbf{u}, \mathbf{u}')$ is the Hamming distance and $\gamma > 0$ is an inverse width scaling parameter.

DEFINITION 3 (GRAM MATRIX). *Given a kernel function k and a dataset of n individuals with attributes $\{\mathbf{u}_1, \dots, \mathbf{u}_n\} \subseteq \mathcal{X}$, the **Gram matrix** (or kernel matrix) $\mathbf{K} \in \mathbb{R}^{n \times n}$ is the symmetric, positive semi-definite matrix defined by the pairwise kernel evaluations [49]:*

$$\mathbf{K}_{ij} = k(\mathbf{u}_i, \mathbf{u}_j) = \langle \phi(\mathbf{u}_i), \phi(\mathbf{u}_j) \rangle_{\mathcal{H}}, \quad \forall i, j \in \{1, \dots, n\}. \tag{5}$$

Next, given a Gram matrix $\mathbf{K}$, we define the **Ridge residualizer matrix** $\mathbf{R} \in \mathbb{R}^{n \times n}$ as:

$$\mathbf{R} = \mathbf{I}_n - \mathbf{K}(\mathbf{K} + n\varepsilon \mathbf{I}_n)^{-1}, \tag{6}$$

where $\varepsilon > 0$ is a regularization parameter and $\mathbf{I}_n$ is the $n \times n$ identity matrix. Furthermore, we define the **centering matrix** $\mathbf{H} \in \mathbb{R}^{n \times n}$ as:

$$\mathbf{H} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}\mathbf{1}^\top, \tag{7}$$

where $\mathbf{1} \in \mathbb{R}^n$ is the column vector of all ones.

Table 1 summarizes the main notations of the paper.

4 CONDOR

We next present in detail the CONDOR method (Algorithm 1), which builds upon both Partial Distance Correlation (PDCOR) [53] and the residualization machinery of kernel-based conditional independence tests, such as KCI [65].

Intuition. CONDOR begins by constructing Gram matrices (kernel similarity matrices) that capture the pairwise similarities between individuals regarding their task-relevant attributes (X), protected attributes (Z), and the observed ranking (r). It then residualizes the ranking and protected attributes using a projection operator, effectively "subtracting" the variation that is already explained by the task-relevant attributes (X). Next, it computes the distance correlation between these residualized embeddings to produce a Protected Attribution Score (PAS) $\in [0,1]$, where higher values indicate a residual dependence on protected attributes. Finally, it performs a Conditional Randomization Test (CRT) to calibrate the PAS, generating a p-value to determine if the result is statistically significant against the null hypothesis $H_0 : Z \perp\!\!\!\perp R \mid X$.

Formalization. First, given positive semi-definite kernel functions k_X, k_Z, k_R, CONDOR computes the *Gram matrices* $\mathbf{K}_X, \mathbf{K}_Z, \mathbf{K}_R$ as:

$$(\mathbf{K}_X)_{ij} = k_X(\mathbf{x}_i, \mathbf{x}_j);\ (\mathbf{K}_Z)_{ij} = k_Z(\mathbf{z}_i, \mathbf{z}_j);\ (\mathbf{K}_R)_{ij} = k_R\big(r(i),\ r(j)\big)\,.$$

Then, it uses $\mathbf{R}_X$, the Ridge residualizer matrix derived from $\mathbf{K}_X$, to remove the variation from $\mathbf{K}_Z$ and $\mathbf{K}_R$ that is already explained by the task-relevant attributes X. This is achieved by left- and right-multiplying $\mathbf{K}_Z$ and $\mathbf{K}_R$ by $\mathbf{R}_X$. Hence, the resulting matrices $\widetilde{\mathbf{K}}_R = \mathbf{R}_X\mathbf{K}_R\mathbf{R}_X$ and $\widetilde{\mathbf{K}}_Z = \mathbf{R}_X\mathbf{K}_Z\mathbf{R}_X$ represent the remaining pairwise similarities once the effect of $\mathbf{X}$ has been removed. These similarities are transformed into pairwise distances, obtaining:

$$\widetilde{\mathbf{D}}^{(r)}_{ij} = \widetilde{\mathbf{K}}_{R,ii} + \widetilde{\mathbf{K}}_{R,jj} - 2\,\widetilde{\mathbf{K}}_{R,ij}, \qquad \widetilde{\mathbf{D}}^{(z)}_{ij} = \widetilde{\mathbf{K}}_{Z,ii} + \widetilde{\mathbf{K}}_{Z,jj} - 2\,\widetilde{\mathbf{K}}_{Z,ij}\,.$$

Next, to ensure that comparisons depend only on relative distances, CONDOR applies double-centering by left- and right-multiplying by the centering matrix $\mathbf{H}$:

$$\overline{\mathbf{D}}^{(z)} = \mathbf{H}\,\widetilde{\mathbf{D}}^{(z)}\,\mathbf{H}, \qquad \overline{\mathbf{D}}^{(r)} = \mathbf{H}\,\widetilde{\mathbf{D}}^{(r)}\,\mathbf{H}\,.$$

From these centered residual distances, we compute the distance-correlation Protected Attribute Score (PAS):

$$\text{PAS} \;:=\; \text{DCOR}\big(\overline{\mathbf{D}}^{(r)},\, \overline{\mathbf{D}}^{(z)}\big) \in [0,1]\,,$$

where DCOR is calculated as a normalized measure of association between the two double-centered distance matrices $\overline{\mathbf{D}}^{(r)}$ and $\overline{\mathbf{D}}^{(z)}$ [54]. Formally, letting $\langle \mathbf{A}, \mathbf{B}\rangle_F = \sum_{i,j} A_{ij}B_{ij}$ denote the Frobenius inner product, the sample *distance covariance* is computed via the centered matrices as:

$$\mathrm{dCov}(\mathbf{D}^{(r)}, \mathbf{D}^{(z)}) \;=\; \frac{1}{n^2}\,\langle \overline{\mathbf{D}}^{(r)},\, \overline{\mathbf{D}}^{(z)}\rangle_F,$$

and the corresponding *distance variances* are $\mathrm{dVar}(\mathbf{D}^{(r)}) = \mathrm{dCov}(\mathbf{D}^{(r)}, \mathbf{D}^{(r)})$ and $\mathrm{dVar}(\mathbf{D}^{(z)}) = \mathrm{dCov}(\mathbf{D}^{(z)}, \mathbf{D}^{(z)})$. The *distance correlation* is then obtained as:

$$\text{DCOR}\big(\overline{\mathbf{D}}^{(r)},\, \overline{\mathbf{D}}^{(z)}\big) \;=\; \frac{\mathrm{dCov}(\mathbf{D}^{(r)}, \mathbf{D}^{(z)})}{\sqrt{\mathrm{dVar}(\mathbf{D}^{(r)})\,\mathrm{dVar}(\mathbf{D}^{(z)})}}\,, \tag{8}$$

which takes values in $[0,1]$. CONDOR's pseudocode is reported in Algorithm 1.

Table 1: Notation

Notation	Description
$\mathbf{x}_i, \mathbf{z}_i$	Feature vectors for individual i
$\mathbf{X}, \mathbf{Z}$	Matrices of task-relevant and protected attributes
$\mathbf{r}$	Observed ranking vector
X, Z, R	Population random variables
k_X, k_Z, k_R	Positive-semidefinite kernel functions
$\mathbf{K}_X, \mathbf{K}_Z, \mathbf{K}_R$	Gram (similarity) matrices from k_X, k_Z, k_R
$\mathbf{R}_X$	Ridge residualizer matrix
$\widetilde{\mathbf{K}}_R, \widetilde{\mathbf{K}}_Z$	Residualized kernel matrices (removing X influences)
$\widetilde{\mathbf{D}}^{(r)}, \widetilde{\mathbf{D}}^{(z)}$	Kernel-induced distance matrices from $\widetilde{\mathbf{K}}_R$ and $\widetilde{\mathbf{K}}_Z$
$\mathbf{H}$	Centering matrix, $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$
$\overline{\mathbf{D}}^{(r)}, \overline{\mathbf{D}}^{(z)}$	Double-centered residual distance matrices
dCov, dVar	Distance covariance and distance variance
DCOR	Distance correlation metric

Algorithm 1 CONDOR

Input: Observed data $\mathbf{X} \in \mathbb{R}^{n\times d_o}$ (task attributes), $\mathbf{Z} \in \mathbb{R}^{n\times d_p}$ (protected attributes), ranking map $r : [n] \to [n]$

Output: PAS$\in [0,1]$ — residual dependence of $\mathbf{r}$ on $\mathbf{Z}$ given $\mathbf{X}$

1: **Step 1: Build kernel matrices.** Select positive-semidefinite kernels k_X, k_Z, k_R. Compute Gram matrices:

$$(\mathbf{K}_X)_{ij} = k_X(\mathbf{x}_i, \mathbf{x}_j),\ (\mathbf{K}_Z)_{ij} = k_Z(\mathbf{z}_i, \mathbf{z}_j),\ (\mathbf{K}_R)_{ij} = k_R(r(i), r(j))\,.$$

2: **Step 2: Remove predictable variation (residualization).** Build the ridge residualizer in the RKHS of $\mathbf{X}$:

$$\mathbf{R}_X = \mathbf{I}_n - \mathbf{K}_X(\mathbf{K}_X + n\varepsilon\,\mathbf{I}_n)^{-1},$$

with a small regularization parameter $\varepsilon > 0$ (default 10^{-3}) to ensure numerical stability. Apply it to $\mathbf{K}_Z$ and $\mathbf{K}_R$:

$$\widetilde{\mathbf{K}}_Z = \mathbf{R}_X\mathbf{K}_Z\mathbf{R}_X, \quad \widetilde{\mathbf{K}}_R = \mathbf{R}_X\mathbf{K}_R\mathbf{R}_X\,.$$

These residual kernels retain only the variation in $\mathbf{Z}$ and $\mathbf{r}$ that cannot be explained by $\mathbf{X}$.

3: **Step 3: Convert residual kernels into distances.** For all pairs i, j, convert similarities to kernel-induced distances:

$$\widetilde{\mathbf{D}}^{(z)}_{ij} = \widetilde{\mathbf{K}}_{Z,ii} + \widetilde{\mathbf{K}}_{Z,jj} - 2\widetilde{\mathbf{K}}_{Z,ij}, \quad \widetilde{\mathbf{D}}^{(r)}_{ij} = \widetilde{\mathbf{K}}_{R,ii} + \widetilde{\mathbf{K}}_{R,jj} - 2\widetilde{\mathbf{K}}_{R,ij}\,.$$

This expresses distance once the influence of $\mathbf{X}$ is removed.

4: **Step 4: Double-center the distance matrices.** Let $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Apply centering on both sides:

$$\overline{\mathbf{D}}^{(z)} = \mathbf{H}\,\widetilde{\mathbf{D}}^{(z)}\,\mathbf{H}, \qquad \overline{\mathbf{D}}^{(r)} = \mathbf{H}\,\widetilde{\mathbf{D}}^{(r)}\,\mathbf{H}\,.$$

This operation removes global offsets (location effects) so that dependence is measured only through relative distances.

5: **Step 5: Compute distance-correlation dependence.** Define:

$$\mathrm{dCov} \;=\; \frac{1}{n^2}\,\langle \overline{\mathbf{D}}^{(r)}, \overline{\mathbf{D}}^{(z)}\rangle_F, \qquad \mathrm{dVar}_r \;=\; \frac{1}{n^2}\,\langle \overline{\mathbf{D}}^{(r)}, \overline{\mathbf{D}}^{(r)}\rangle_F,$$

$$\mathrm{dVar}_z \;=\; \frac{1}{n^2}\,\langle \overline{\mathbf{D}}^{(z)}, \overline{\mathbf{D}}^{(z)}\rangle_F\,.$$

Then set:

$$\text{PAS} \;=\; \text{DCOR}\big(\overline{\mathbf{D}}^{(r)}, \overline{\mathbf{D}}^{(z)}\big) = \frac{\mathrm{dCov}}{\sqrt{\mathrm{dVar}_r\,\mathrm{dVar}_z}} \;\in [0,1]\,.$$

6: **return** PAS.

4.1 PAS Characterization

We next provide the intuition of two key properties establishing the correctness of the PAS. For readability sake, the corresponding theorems are provided in the Appendix.

Null Alignment: If the ranking does not use Z beyond what X already explains (i.e., under the null hypothesis $Z \perp\!\!\!\perp R \mid X$), the population PAS score is exactly zero, and the sample estimate converges to zero as n grows (Theorem 2 in the Appendix). This ensures the score conceptually aligns with the conditional-independence null tested by the CRT (see Section 4.2).

Local Sensitivity: When the ranking R begins to leverage Z slightly beyond what X already explains, the population PAS departs from 0 linearly in relation to the small signal (Theorem 3 in the Appendix).

Taken together, these calibrated behaviors—at the null and under weak effects—justify reading PAS as a simple, model-agnostic effect size for the residual dependence of the ranking on Z beyond X.

4.2 Statistical Calibration

To translate the PAS into a binary decision with finite-sample error control, we use conditional randomization [12, 17, 35] to test the null hypothesis $H_0 : Z \perp\!\!\!\perp R \mid X$. This is how the Conditional Randomization Calibration (CRT) works in practice. We first fix the observed task attributes $\mathbf{X}$ and the ranking $\mathbf{r}$, we regenerate protected attributes under a model that enforces H_0, and we ask how often the resulting score exceeds the observed one. Concretely, given an audit mapping $\mathrm{PAS}(\mathbf{X}, \mathbf{Z}, \mathbf{r}) \in [0, 1]$, compute $s_{\text{obs}} = \mathrm{PAS}(\mathbf{X}, \mathbf{Z}, \mathbf{r})$; fit a conditional model $\widehat{P}(\mathbf{Z} \mid \mathbf{X})$ (e.g., logistic/multinomial models, tree-based conditionals, or CTGAN [59]); draw B independent resamples $\mathbf{Z}^{(b)} \sim \widehat{P}(\mathbf{Z} \mid \mathbf{X})$ and form $s^{(b)} = \mathrm{PAS}(\mathbf{X}, \mathbf{Z}^{(b)}, \mathbf{r})$ for $b = 1, \ldots, B$. The one-sided Monte Carlo p-value is

$$p = \frac{\left|\{\, b : s^{(b)} \geq s_{\text{obs}} \,\}\right|}{B}. \tag{9}$$

For a pre-set level α (e.g., $\alpha = 0.05$), reject when $p \leq \alpha$; otherwise report "no evidence of residual use". When benchmarking multiple audits on the same dataset, reuse the same $\{\mathbf{Z}^{(b)}\}_{b=1}^{B}$ to reduce Monte Carlo variability and ensure fair comparison. Larger B increases p-value resolution at linear computational cost; we used $B = 1000$ in the experiments. CRT is finite-sample exact if $P(\mathbf{Z} \mid \mathbf{X})$ is known; with an estimate $\widehat{P}$, sample-splitting or cross-fitting improves robustness [7]. The pseudocode of the CRT calibration is provided in Algorithm 2.

Algorithm 2 Conditional–Randomization Test

Input: Observed $(\mathbf{X}, \mathbf{Z}, \mathbf{r})$, function $\mathrm{PAS}(\cdot)$, number of resamples B
Output: One–sided p–value testing $H_0 : R \perp\!\!\!\perp Z \mid X$

1: **Step 1: Compute observed score.** $s_{\text{obs}} \leftarrow \mathrm{PAS}(\mathbf{X}, \mathbf{Z}, \mathbf{r})$.
2: **Step 2: Generate null resamples.** For $b = 1, \ldots, B$: sample $\mathbf{Z}^{(b)} \sim \widehat{P}(\mathbf{Z} \mid \mathbf{X})$ (fitted from data), then compute $s^{(b)} = \mathrm{PAS}(\mathbf{X}, \mathbf{Z}^{(b)}, \mathbf{r})$.
3: **Step 3: Compute empirical tail probability** p as in Equation 4.2
4: **return** (s_{obs}, p).

4.3 Computational Complexity

We analyze the time complexity of Condor w.r.t. the sample size n, the dimension of task-relevant (d_o) and protected attributes (d_p), and the number of bootstrap resamples B used for calibration.

Algorithm 1 - Step 1: Kernel Matrix Construction. Constructing the Gram matrices requires computing pairwise similarities for all n individuals.

- For the task-relevant attributes K_X (using inputs $X \in \mathbb{R}^{n \times d_o}$), the complexity is $O(n^2 d_o)$. Similarly, for the protected attributes K_Z, the complexity is $O(n^2 d_p)$.
- For the ranking K_R, the complexity is $O(n^2)$.

Algorithm 1 - Step 2: Ridge Residualization. This step involves two computationally intensive operations: Computing the ridge residualizer $R_X = I_n - K_X(K_X + n\epsilon I_n)^{-1}$ requires inverting a dense $n \times n$ matrix, which scales as $O(n^3)$; applying the residualizer to obtain $\tilde{K}_Z = R_X K_Z R_X$ and $\tilde{K}_R = R_X K_R R_X$ requires dense matrix-matrix multiplication, scaling as $O(n^3)$ for each kernel.

Algorithm 1 - Steps 3–5: Distance Conversion and Correlation. Converting kernel matrices to distances $(\tilde{D}^{(z)}, \tilde{D}^{(r)})$, performing double-centering via H, and computing the final distance correlation involve element-wise operations on $n \times n$ matrices. The total complexity for these steps is $O(n^2)$.

Statistical Calibration (Algorithm 2). CRT repeats the score calculation for B resamples to generate a p-value. Since X and the ranking r remain fixed throughout the permutation test, the residualizer R_X and the residualized ranking kernel $\tilde{K}_R$ can be precomputed only once at the cost of $O(n^3 + n^2 d_o)$. Then, for each $b \in \{1, \ldots, B\}$ (bootstrap loop):

- *Resampling (Step 1):* Generating new protected attributes $Z^{(b)}$ and building the corresponding kernel $K_{Z^{(b)}}$ takes $O(n^2 d_p)$.
- *Projection (Step 2):* The new protected kernel must be projected into the residual space: $\tilde{K}_{Z^{(b)}} = R_X K_{Z^{(b)}} R_X$. This matrix multiplication is the bottleneck of the loop, scaling as $O(n^3)$.
- *Scoring (Steps 3–5):* Re-evaluating the PAS takes $O(n^2)$.

Total Complexity. Aggregating the costs from Algorithm 1 and Algorithm 2, the total time complexity is: $O\left(n^2 d_o + B \cdot (n^3 + n^2 d_p)\right)$.

While the algorithm scales linearly with the feature dimension d_p, d_o and the number of resamples B, it is cubic with respect to the sample size n due to matrix multiplications required during the residualization of Z.

Scalable Approximations. This $O(n^3)$ bottleneck can be effectively circumvented using standard scalable kernel techniques. Replacing direct matrix factorizations with Krylov iterative solvers (e.g., Conjugate Gradient or MINRES) [29, 38, 45] reduces the inversion cost to a sequence of T matrix-vector multiplications. By further combining this with low-rank approximations, such as the Nyström method or Random Fourier Features using $m \ll n$ components [13, 27, 58], the memory footprint drops from $O(n^2)$ to $O(nm)$. Correspondingly, the bottleneck time complexity reduces to $O(Tnm)$ or $O(nm^2)$ (where T is the number of solver iterations), allowing Condor to scale efficiently to large datasets without computing the exact dense matrices.

5 CAUSAL INTERPRETATION

CONDOR captures the residual dependence of the ranking on protected attributes, once task-relevant observables are accounted for. As anticipated in Section 1, to have an effective audit we need to adopt the formal language of causality [6, 20]. By combining the the conditional–independence test $Z \perp\!\!\!\perp R \mid X$ of CONDOR with an unconditional test $Z \perp\!\!\!\perp R$, while properly keeping in account the causal assumptions at the basis of our setting, we obtain the 2×2 matrix in the right-most panel of Figure 1. We next prove that the causal structures provided in the 2×2 matrix are correct.

First, let us observe that by the characteristics of our setting, we have the following causal assumptions:

(1) X **is a cause of** R, ($X \to R$): whatever algorithm has been used to produce R, it surely has seen as (part of the) input the task-relevant attributes X, as there is no other type of information available (besides the "forbidden" Z).
(2) R **cannot be a cause of** X ($R \not\to X$): whatever is the decision-making process, the output (R) cannot be a cause of the input (X), as the input *existed before* the output.
(3) Z **does not have any ancestors** ($\nexists U : U \to Z$): as often is the case, we assume that the demographic attributes such as gender or race are not determined by any cause.

Next, we recall that in a Structural Causal Model, which uses a DAG (directed acyclic graph) to represent the causal relations, the key notion of **directional separation** (see e.g., [23, 42]) determines that two variables are conditionally independent if all informational paths between them are "blocked". Conversely, they are dependent if there exists at least an "open path" between them.

DEFINITION 4 (d-SEPARATION). *Let $C \subseteq V$ be a set of variables we condition on. Two nodes $A, B \in V$ are said to be d-separated by C if every path (considered on the undirected version of the DAG) between them is "blocked" by the conditioning set C. A path P is blocked by C if it satisfies any of the following conditions:*

- ***Chain:*** *P contains a chain $A \cdots \leftarrow M \leftarrow \cdots B$ or $A \cdots \rightarrow M \rightarrow \cdots B$, and the middle node M is conditioned on (i.e., $M \in C$).*
- ***Fork:*** *P contains a fork $A \cdots \leftarrow M \rightarrow \cdots B$, and the middle node M is conditioned on (i.e., $M \in C$).*
- ***Collider:*** *P contains a collider $A \cdots \rightarrow M \leftarrow \cdots B$, such that the collision node M is* not *conditioned on ($M \notin C$) and no descendant of M is conditioned on (i.e., $Desc(M) \cap C = \emptyset$).*

We consider the following two statements:

S1: Z is unconditionally independent of R ($Z \perp\!\!\!\perp R$);
S2: Z is conditionally independent of R, given X ($Z \perp\!\!\!\perp R|X$).

Our analysis of the four scenarios hinges on which paths are open or blocked by default ($S1$) and which become open or blocked after conditioning on X ($S2$).

THEOREM 1. *Under the causal assumptions (1)-(3) above, based on the truthfulness of $S1$ and $S2$ the following four cases hold:*

(I) $S1 \wedge S2 \implies Z$ *is unrelated to* $X \to R$
(II) $S1 \wedge \neg S2 \implies$ *causally impossible*
(III) $\neg S1 \wedge S2 \implies Z \to X \to R$
(IV) $\neg S1 \wedge \neg S2 \implies Z \to R$

PROOF. We prove the four mutually exclusive cases.

(*I*) Satisfying S1 ($Z \perp\!\!\!\perp R$): For Z and R to be unconditionally independent, there must be no open paths between them. Given the constraints (1)-(3), the absence of open paths requires: that Z does not cause R directly ($Z \not\to R$), and that Z does not cause X ($Z \not\to X$), otherwise the path $Z \to X \to R$ would be open. Furthermore, no fork due to a confounder U ($Z \leftarrow U \rightarrow R$ or $Z \leftarrow U \rightarrow X$) is possible due to constraint (3) (U cannot be ancestor of Z).

Satisfying S2 ($Z \perp\!\!\!\perp R \mid X$): If no path exists between Z and R, then conditioning on any set of variables, including X, cannot create or open a path. Thus, independence holds trivially. Therefore, the only viable causal structure is when Z is isolated from $X \to R$.

(*II*) There is no open path between Z and R ($S1$), but conditioning on X creates an open path ($S2$). This requires X to be a collider on a path P between Z and R, or a descendant of such a collider.

- Case 1: X is the collider. The path is $Z \to X \leftarrow R$. This graph violates constraint (2) which states $R \not\to X$.
- Case 2: X is a descendant of the collider ($C \to X$). The path is $Z \to C \leftarrow R$ with $C \to X$. Adding constraint (1), $X \to R$, produces the structure $R \to C \to X \to R$, which is a causal cycle, violating the acyclicity assumption of causal structures.

Therefore, the case $S1 \wedge \neg S2$, is causally impossible.

(*III*) We seek a DAG where an open path exists between Z and R ($\neg$S1), such that it is blocked by conditioning on X (S2). This requires X to be a non-collider (chain or fork center) on every path. Only two structures are possible:

- Chain: $Z \to X \to R$, which satisfies all requirements;
- Fork center: $Z \leftarrow U \rightarrow X \rightarrow R$, which is not admissible due to constraint (3) (U cannot be ancestor of Z).

(*IV*) We seek a DAG where at least one path is open ($\neg S1$), and at least one path remains open after conditioning on X ($\neg S2$). This implies the existence of an unblocked path between Z and R bypassing X. Only three structures are, in principle, possible:

- Direct link bypass: $Z \to R$.
- Direct link bypass + chain: besides $Z \to R$ a second path $Z \to X \to R$ might exists. However,from an auditing standpoint, the second path is less relevant than the fact that $Z \to R$.
- Confounding bypass: $Z \to X \to R$ and $Z \leftarrow U \rightarrow R$ which is not viable due to constraint (3).

In both of the first two structures holds $Z \to R$, which proves case (*IV*), concluding our proof. □

Connection to Interventional Fairness. Our causal diagnostic framework is deeply connected to the concept of *interventional fairness* [47]. A system is considered interventionally fair if changing an individual's protected attribute does not affect their outcome, assuming all legitimate, task-relevant features are held constant. While formally proving this requires knowing the system's exact internal causal rules, CONDOR provides a practical, observational proxy. Under our causal assumptions, passing the CONDOR audit provides strong empirical evidence that an opaque ranking mechanism is interventionally fair, as it rules out direct discrimination.

6 EXPERIMENTAL SETTINGS

In this section, we describes the datasets (synthetic and real-world) and the baselines. Finally, we introduce the specific unconditional dependence test that we couple with CONDOR to provide the final causal interpretation.

6.1 Synthetic Data and Ranking Generation

We next present the process to generate synthetic data where we can control the influence of Z over X, and the influence of Z and X over $\mathbf{r}$. We create cohorts of size n, with task-relevant features $\mathbf{X} \in \mathbb{R}^{n \times d_o}$, protected features $\mathbf{Z} \in \mathbb{R}^{n \times d_p}$, and a ranking $\mathbf{r}$. The generation process proceeds as follows:

(1) Draw protected features Z: $\mathbf{Z}$ is drawn from a *Gaussian Cluster Generation* process: k_q cluster centers are picked in d_p-dimensional space, and n samples are drawn from Gaussian distributions centered at a randomly assigned cluster mean.

(2) Draw latent features T: Latent task factors $\mathbf{T} \in \mathbb{R}^{n \times d_o}$ are generated using the same Gaussian Cluster process described in step (1).

(3) Generate task-relevant features X: We generate $\mathbf{X}$ combining latent factors and protected attributes, controlled by an entanglement parameter $\gamma \in [0, 1]$:

$$\mathbf{X} \;=\; \gamma\,\mathbf{T} \;+\; (1-\gamma)\,\mathbf{Z}' \;+\; \boldsymbol{\varepsilon}_X\,,$$

where $\mathbf{Z}'$ is a projection of $\mathbf{Z}$ onto $\mathbb{R}^{n \times d_o}$, and where the i-th row of $\boldsymbol{\varepsilon}_X$ is $\boldsymbol{\varepsilon}_{X,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \Sigma_X)$. Here, $\gamma = 1$ implies $\mathbf{X} \perp \mathbf{Z}$ (independence), while $\gamma = 0$ implies $\mathbf{X}$ is fully determined by $\mathbf{Z}$.

(4) Generate underlying scores s: We generate scores as a convex combination of task and protected components, controlled by an entanglement parameter $\beta \in [0, 1]$:

$$\mathbf{s} \;=\; (1-\beta)\,\mathbf{s}_X \;+\; \beta\,\mathbf{s}_Z \;+\; \boldsymbol{\varepsilon}_s, \quad \text{with} \quad \boldsymbol{\varepsilon}_s \sim \mathcal{N}(\mathbf{0},\, \sigma_s^2 I_n).$$

We use linear projections $\mathbf{s}_X = \mathbf{X}\mathbf{w}_X$ and $\mathbf{s}_Z = \mathbf{Z}\mathbf{w}_Z$ (with uniform weights $\mathbf{w}_X \in \mathbb{R}^{d_o}$ and $\mathbf{w}_Z \in \mathbb{R}^{d_p}$). β governs direct bias: $\beta = 0$ yields scores dependent only on $\mathbf{X}$, while $\beta = 1$ yields scores dependent only on $\mathbf{Z}$.

(5) Generate final ranking r: Individuals are sorted by $\mathbf{s}$, and ranks are normalized: $r_i := (\mathrm{pos}_i - 0.5)/n$.

In our experiments, γ and β vary across $[0, 1]$. Further specific parametrizations are detailed in Section 7.

6.2 Real Data

We use five publicly-available real-world datasets:

- ADULT: Individuals with demographic and work information [14] (individuals $n = 48842$, number of task-relevant features $d_o = 12$, number of protected features $d_p = 2$). Task: ranking for high income potential.
- COMPAS : Data on criminal defendants used to generate recidivism risk scores [3] ($n = 6172$, $d_o = 9$, $d_p = 2$). Task: ranking based on likelihood of recidivism.
- LAW SCHOOL: Individuals with law school admission records including LSAT and GPA [44] ($n = 20798$, $d_o = 2$, $d_p = 2$). Task: ranking for law school success.
- STUDENT PERFORMANCE: Students with demographic, social, and school-related information from two Portuguese schools [10] (individuals $n = 1044$, $d_o = 30$, $d_p = 2$). Task: ranking for academic performance.
- STRESS LEVEL: Individuals with smartphone usage behavior, sleep patterns, and productivity analytics [2] ($n = 50000$, $d_o = 9$, $d_p = 2$). Task: ranking for stress levels.

Real-world data allows us to test CONDOR in realistic scenarios where complex relationships between X and Z exist naturally in the data (beyond our control). We exploit this in two different ways.

Grey-Box Ranker. In Section 7.2, we assess performance with real-world feature distributions and correlations while still maintaining control over the influence of Z and X over $\mathbf{r}$. We first estimate one-sided scoring functions from the observed data,

$$\mathbf{s}_X := f_X(\mathbf{X}), \qquad \mathbf{s}_Z := f_Z(\mathbf{Z}).$$

When Y is binary, we take $\mathbf{s}_X = \widehat{\mathbb{P}}(Y = 1 \mid \mathbf{X})$ and $\mathbf{s}_Z = \widehat{\mathbb{P}}(Y = 1 \mid \mathbf{Z})$ from off-the-shelf probabilistic classifiers (e.g., logistic regression or XGBoost) trained with 5-fold cross-validation. We then mix the two components through a direct-influence knob $\beta \in [0, 1]$ and add noise,

$$\mathbf{s} \;=\; (1-\beta)\,\mathbf{s}_X \;+\; \beta\,\mathbf{s}_Z \;+\; \boldsymbol{\varepsilon}_s, \quad \boldsymbol{\varepsilon}_s \sim \mathcal{N}(\mathbf{0},\, \sigma_s^2 I_n),$$

and obtain the final ranking $\mathbf{r}$ by sorting $\mathbf{s}$ and normalizing the ranks via

$$r_i := (\mathrm{pos}_i - 0.5)/n.$$

This construction injects a controlled amount of Z-driven signal (through β) while preserving the empirical correlation structure within $\mathbf{X}$ and between $\mathbf{X}$ and $\mathbf{Z}$.

Black-Box Ranker. In Section 7.4, we use real-world features and employ as ranker directly the "target" feature provided in the dataset. This allows us to test CONDOR in a truly black-box setting where we ignore the inner logic that produced the ranking.

6.3 Baselines

We consider three baseline methods to test for conditional independence between R and Z given X.

KCI [65]. Kernel Conditional Independence estimates residual dependence between $\mathbf{Z}$ and $\mathbf{r}$ after regressing out $\mathbf{X}$ in the RKHS induced by a positive-definite kernel k_X. Defining

$$\mathbf{R}_X \;=\; \mathbf{I}_n - \mathbf{K}_X\big(\mathbf{K}_X + n\varepsilon\,\mathbf{I}_n\big)^{-1},$$

the residualized kernel matrices are $\widetilde{\mathbf{K}}_R = \mathbf{R}_X\mathbf{K}_R\mathbf{R}_X$ and $\widetilde{\mathbf{K}}_Z = \mathbf{R}_X\mathbf{K}_Z\mathbf{R}_X$. The empirical dependence is then measured by the Hilbert–Schmidt covariance:

$$\mathrm{KCI}_n(\mathbf{r}, \mathbf{z} \mid \mathbf{x}) \;=\; \frac{1}{(n-1)^2}\;\mathrm{tr}\big(\widetilde{\mathbf{K}}_R\,\mathbf{H}\,\widetilde{\mathbf{K}}_Z\,\mathbf{H}\big),$$

which is then normalized by the kernel self-variances. The p-value is then obtained generating the approximate null distribution with Monte Carlo simulation by computing a weighted sum of chi-squared samples [65]. We used the publicly available implementation from the Python library Causal-learn [66].

PDCOR [53]. Partial distance correlation instead operates on pairwise distances, projecting out the contribution of $\mathbf{X}$ at the level of distance matrices. For centered distance matrices

$\overline{\mathbf{D}}^{(r)}, \overline{\mathbf{D}}^{(z)}, \overline{\mathbf{D}}^{(x)}$—obtained by left- and right-multiplying each pairwise distance matrix by $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$—one computes

$$\textsc{pdCor}(\mathbf{r}, \mathbf{z} \mid \mathbf{x}) = \frac{\textsc{dCor}_{r,z} - \textsc{dCor}_{r,x}\,\textsc{dCor}_{z,x}}{\sqrt{(1 - \textsc{dCor}^2_{r,x})(1 - \textsc{dCor}^2_{z,x})}}, \tag{10}$$

where dCor is defined as in Equation 8, $\textsc{dCor}_{r,z} = \textsc{dCor}(\overline{\mathbf{D}}^{(r)}, \overline{\mathbf{D}}^{(z)})$, $\textsc{dCor}_{r,x} = \textsc{dCor}(\overline{\mathbf{D}}^{(r)}, \overline{\mathbf{D}}^{(x)})$, and $\textsc{dCor}_{z,x} = \textsc{dCor}(\overline{\mathbf{D}}^{(z)}, \overline{\mathbf{D}}^{(x)})$. This closed form yields a normalized measure in $[-1, 1]$, interpretable as the correlation between residualized distance-embeddings of $\mathbf{r}$ and $\mathbf{z}$. The p-value is then computed via a non-parametric permutation test, as provided in the Python library Hyppo [39].

CMI (Conditional Mutual Information) [11, 21, 32, 55]. Conditional Mutual Information (CMI) is an information–theoretic quantity that measures how much additional information the protected attributes $\mathbf{Z}$ provide about the ranking $\mathbf{r}$ once the observable attributes $\mathbf{X}$ are known. Formally, it is defined as

$$I(R; Z \mid X) = \mathbb{E}\left[\log \frac{p(r, z \mid x)}{p(r \mid x)\, p(z \mid x)}\right] \geq 0,$$

and equals zero if and only if R and Z are conditionally independent given X [11]. This makes CMI a natural population–level analogue of our auditing null $H_0 : R \perp\!\!\!\perp Z \mid X$. We used the publicly available Python library npeet [56] and we calibrate CMI via the Conditional-Randomization Test, as done with Condor, to obtain a p–value assessing evidence against H_0.

6.4 Unconditional Test

We adopt **HSIC** [26] as our test for **unconditional independence**, $R \perp\!\!\!\perp Z$. HSIC correspond to an unconditional version of KCI. In more details, let k_r, k_z be positive–definite kernels on the domains of r, z, again define the $n \times n$ Gram matrices $(K_r)_{ij} = k_r(r_i, r_j)$, $(K_z)_{ij} = k_z(z_i, z_j)$, then HSIC is defined as:

$$\text{HSIC}(\mathbf{r}, \mathbf{z}) = \frac{\text{tr}(K_r H K_z H)}{\sqrt{\text{tr}(K_r H K_r H)\,\text{tr}(K_z H K_z H)}} \in [0, 1].$$

The p-value is then computed as in [65], using the open source Python library Hyppo [39].

7 EXPERIMENT RESULTS

The goal of our empirical assessment is to answer the following research questions:

RQ1: How accurately can Condor detect the influence of protected attributes Z on ranking $\mathbf{r}$ in a fully controlled setting, with varying levels of influence of Z over X, and of Z and X over $\mathbf{r}$?

RQ2: How does Condor performs when applied to real-world feature distributions with naturally occurring correlations? Furthermore, does it reliably avoid false positives when tested against spurious correlations (e.g., random noise as a protected attribute)?

RQ3: When paired with an unconditional independence test, can Condor effectively distinguish between the ideal scenario of no influence of Z on $\mathbf{r}$, indirect influence mediated by X, and direct influence?

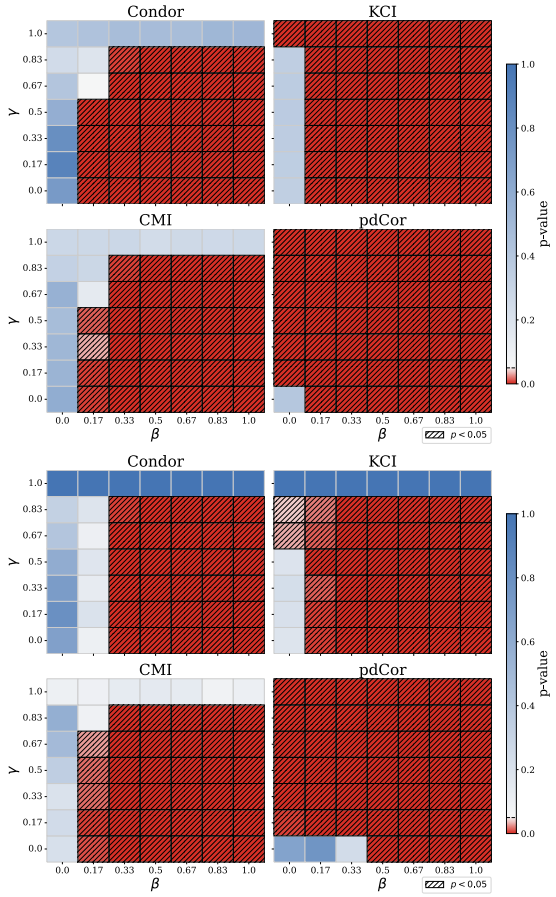

Figure 2: Heatmap of p-values on synthetic data (n=1000, k_x=k_t=5). Top four heatmaps: d_o=5, d_p=2. Bottom four heatmaps: d_o=10, d_p=5. Each panel corresponds to one scoring rule; axes report the influence of X on Z (γ) and of Z on the ranking (β), and the red/white/blue scale marks respectively strong evidence against, the $\alpha = 0.05$ boundary, and little evidence against $H_0 : R \perp\!\!\!\perp Z \mid X$.

RQ4: How does Condor work as an audit in a purely black-box setting?

RQ5: How does Condor compare to the baselines in terms of run-time?

7.1 Synthetic Data (RQ1)

We evaluate the performance of Condor and the baseline methods in detecting conditional dependence $H_0 : R \perp\!\!\!\perp Z \mid X$. We present the results as heatmaps of p-values, where red indicates strong evidence against H_0 ($p \leq 0.05$), blue indicates little evidence against H_0, and white corresponds to boundary values. Figure 2 show the p-value heatmaps for two synthetic data settings, with different numbers of protected and task features. The x-axis (β) controls the direct influence of the protected attribute Z on the ranking score S, while the y-axis (γ) controls the dependence between Z and the task-relevant features X.

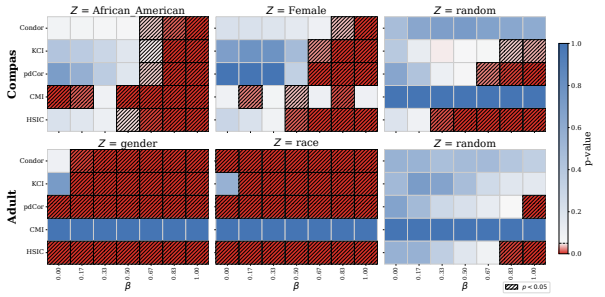

Figure 3: Compas and Adult datasets. Heatmap of p-values on Compas and Adult rankings. Each row corresponds to one dataset and each heatmap to a specific composition of the protected features Z. For each heatmap, the scoring rule is reported on the row and on the columns the influence of Z on the ranking (β), and the red/white/blue scale marks respectively strong evidence against, the $\alpha = 0.05$ boundary, and little evidence against $H_0 \colon R \perp\!\!\!\perp Z \mid X$.

The ideal audit tool should reject H_0 (show red) for $\beta > 0$, as this is when Z has a direct influence on the ranking R that is not mediated by X. The result should be insensitive to $\gamma < 1$, which only models the correlation between X and Z (a potential source of indirect discrimination, but not a violation of H_0). Furthermore, in the special case of $\gamma = 1$, which indicates that X contains fully Z (or all its information), a method should indicate the independence between R and Z given X (blue), since conditioning on X is removing all the information on Z.

As shown in the figure, Condor achieves this ideal behavior. It reliably shows high p-values (blue) in the first column where $\beta = 0$, correctly finding no evidence of residual dependence. For all columns where $\beta > 0$, it shows low p-values (red or white if close to the boundary), successfully detecting the injected direct bias, regardless of the value of γ.

KCI, instead, fails in the case of $\gamma = 1$ in the top heatmap of Figure 2 and also over-rejects the null when $\beta = 0$, in two cases in the bottom heatmap. CMI performs similarly to Condor on this synthetic-data setting, however, as we will see in the rest of this section, it does not perform as expected when dealing with the complex relationships of real data. Finally, pdCor over-reject the null hypothesis in all cases, except when $\beta = \gamma = 0$.

7.2 Real Data with Grey-Box Ranker (RQ2)

Figure 3 presents the results from the Compas and Adult datasets with generated rankings. Here, we only vary β, as the complex real-world relationship between X and Z is given by the datasets themselves. We test using different protected attributes Z for each dataset, and we also include a $Z = \textit{random}$ case as a sanity check, where the protected attribute Z is replaced with random noise and no dependence should be found.

For the protected attributes, Condor, KCI, and pdCor show similar results, in particular, it is reasonable and correct that with the increase of β the methods reject more and more the null hypothesis, indicating the conditional dependence between the protected attributes and the ranking. CMI, on the other hand, does not exhibit this desired behaviour. In the $Z = \textit{random}$ control case, Condor

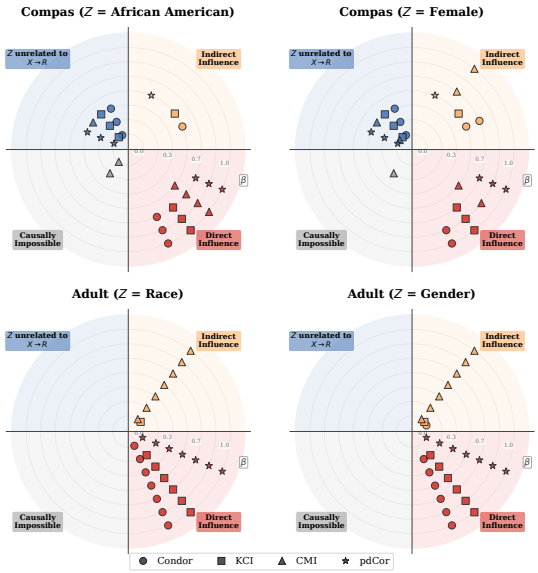

Figure 4: The four panels map the audit outcomes for the Compas and Adult datasets onto the 2×2 causal matrix introduced in Figure 1. Each marker represents an audited ranking, with concentric dotted lines indicating the strength of the injected direct bias (β). As the true bias β increases (moving outward from the center), Condor correctly tracks the transition from no influence ($\beta \approx 0$) to either "Direct Influence" or "Indirect Influence" ($\beta > 0$), successfully diagnosing the causal structure of the opaque ranker.

correctly reports high p-values (blue) for all β, confirming it does not find spurious correlations. KCI and pdCor, instead, fail and reject the null in this scenario.

7.3 Causal Interpretation (RQ3)

We now show how Condor can be combined with an unconditional test to understand the possible causal relationships between the protected attributes, the ranking, and the task features. As mentioned, we use HSIC as an unconditional test; however, in principle, any other unconditional test can be used.

The empirical analysis is reported in Figure 4. In each figure, the four quadrants correspond to the four quadrants of the 2×2 matrix on Figure 1. Concentric circular lines represent different values of the parameter β, which controls the influence of Z on R: recall that $\beta = 0$ indicates that the ranking only depends on the task-relevant features X, while $\beta = 1$ indicates that the ranking depends only on the protected attributes Z. It is worth recalling that, for both datasets, this analysis is not directly on the datasets themselves, but it represents an audit of the ranking produced according to the procedure described in Section 4.2.

First, we can observe that the *causally impossible* remains rightfully unpopulated in all the experiments, with the exception of CMI which, however, already showed discordant behaviours in Section 7.2. In the Compas dataset, we observe that, for both settings of Z, when the impact of Z on the ranking is small ($\beta < 0.5$) all the methods correctly report the first quadrant (Z (unrelated to)

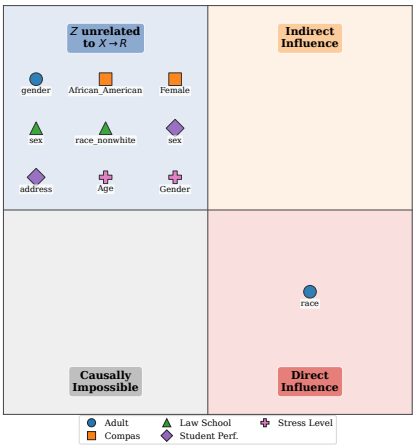

Figure 5: Causal interpretation of CONDOR's audit on fully black-box models.

$X \to R$). As β increases, all the methods start signaling $Z \to R$. When β is neither too strong nor too weak, we can observe some influence of Z on X, i.e., the quadrant $Z \to X \to R$. In the ADULT dataset, the influence of Z on the original binary label of the dataset is very strong, and thus it remains strong in the ranking we generated according to the procedure described in Section 4.2. That explains why all the methods are concordant in producing $Z \to R$, with again the exception of CMI.

7.4 Complete Black-Box Setting (RQ4)

We now apply CONDOR as an auditor on fully black box and fully empirical data, to reproduce a real life application example.

Figure 5 maps various protected attributes (Z) from five real-world datasets (ADULT, COMPAS, LAW SCHOOL, STUDENT PERFORMANCE, and STRESS LEVEL) onto the four quadrants of the causal interpretation matrix. Most protected attributes fall into the top-left quadrant, indicating no influence. Conversely, the *race* attribute in the ADULT dataset falls into the bottom-right quadrant, revealing a direct influence on the ranking.

7.5 Run-time Analysis (RQ5)

In Figure 6, we report the runtime of CONDOR and the baselines. The linear growth of CONDOR in the log-log plot confirms the polynomial runtime, w.r.t. the input size n, proven in Section 4.3. As expected, KCI which has the same complexity of CONDOR, shows the same asymptotic runtime. Also as expected, CMI is faster, while PDCOR is significantly slower.

8 CONCLUSIONS

In this work, we introduced CONDOR, a model-agnostic framework designed to audit opaque ranking systems for demographic bias. By bridging reproducing kernel Hilbert space (RKHS) residualization with distance correlation, CONDORrigorously isolates and quantifies the residual influence of protected attributes beyond legitimate, task-relevant features. Beyond its practical utility, our framework is backed by solid theoretical guarantees: we prove that the Protected Attribution Score ensures exact null alignment under conditional

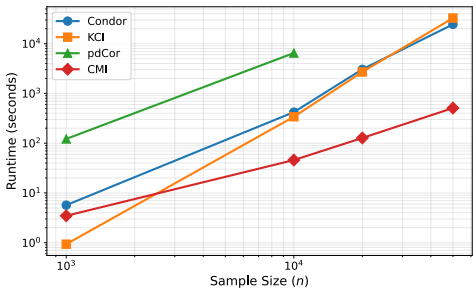

Figure 6: Runtime (in seconds) of each method as a function of the sample size n on synthetic data with $d_o = d_p = 5$, $\gamma = \beta = 0.5$, $K = 5$ folds and $B = 200$ permutations per fold. Both axes use a logarithmic scale. Experiments are run on a CPU Intel Xeon Gold 6248R, with 172 GB of RAM, running Ubuntu 24.04 LTS with Python 3.12.3.

independence and exhibits predictable local sensitivity to emerging biases. Futhermore, by pairing this conditional test with an unconditional independence baseline, we provide a comprehensive causal diagnostic tool capable of distinguishing between direct discrimination and indirect, feature-mediated influence.

Our empirical evaluation confirms that CONDOR is robust, natively accommodates mixed data types, and reliably outperforms established statistical baselines across both synthetic and real-world scenarios. Ultimately, CONDOR equips auditors with a practical, statistically sound forensic tool to hold black-box systems accountable. Moving forward, we plan to extend this framework toward prescriptive auditing, translating statistical bias signals into concrete, accountable rank interventions that restore conditional independence under explicit utility budgets, and to broaden its scope to encompass intersectional fairness guarantees.

A APPENDIX

THEOREM 2 (NULL ALIGNMENT AND SAMPLE CONVERGENCE). *Assume i.i.d. sampling of (R, Z, X). Let $\phi_R : \mathrm{R} \to \mathcal{H}_R$ and $\phi_Z : \mathrm{Z} \to \mathcal{H}_Z$ be feature maps into separable Hilbert spaces, and assume*

$$\mathbb{E}\big\|\phi_R(R)\big\|^4_{\mathcal{H}_R} < \infty, \qquad \mathbb{E}\big\|\phi_Z(Z)\big\|^4_{\mathcal{H}_Z} < \infty.$$

Define the population residual features

$$\widetilde{\phi}_R := \phi_R(R) - \mathbb{E}[\phi_R(R) \mid X], \qquad \widetilde{\phi}_Z := \phi_Z(Z) - \mathbb{E}[\phi_Z(Z) \mid X],$$

and adopt the normalization under which, for any Hilbert–valued random elements U, V with finite second moments,

$$\mathrm{DCov}^2_\star(U, V) = \|\mathrm{Cov}(U, V)\|^2_{\mathrm{HS}}.$$

If $Z \perp\!\!\!\perp R \mid X$, then $\mathrm{PAS}_\star := \mathrm{DCor}(\widetilde{\phi}_R, \widetilde{\phi}_Z) = 0$.

Moreover, for each sample size n suppose there exist measurable maps $f_{R,n}, f_{Z,n} : \mathrm{R} \times \mathrm{Z} \times \mathrm{X} \to \mathcal{H}_R, \mathcal{H}_Z$ such that, for i.i.d. copies $\{(R_i, Z_i, X_i)\}_{i=1}^n$ of (R, Z, X),

$$\widetilde{\phi}^{(i)}_R = f_{R,n}(R_i, Z_i, X_i), \qquad \widetilde{\phi}^{(i)}_Z = f_{Z,n}(R_i, Z_i, X_i), \qquad i = 1, \ldots, n,$$

and $f_{R,n}(R, Z, X) \to \widetilde{\phi}_R, \qquad f_{Z,n}(R, Z, X) \to \widetilde{\phi}_Z \quad$ in L_2.
Assume in addition that the fourth moments are uniformly bounded:

$$\sup_n \mathbb{E}\big\|f_{R,n}(R, Z, X)\big\|^4_{\mathcal{H}_R} < \infty, \qquad \sup_n \mathbb{E}\big\|f_{Z,n}(R, Z, X)\big\|^4_{\mathcal{H}_Z} < \infty.$$

Let $\widehat{PAS}$ be the usual sample distance correlation (U– or V–statistic) computed from $\{(\widetilde{\widetilde{\phi}}_R^{(i)}, \widetilde{\widetilde{\phi}}_Z^{(i)})\}_{i=1}^n$. Then, under $Z \perp\!\!\!\perp R \mid X$,

$$\widehat{PAS} \xrightarrow{\mathbb{P}} 0 \qquad \text{as } n \to \infty.$$

Proof. *Population.* Set $U := \phi_R(R) \in \mathcal{H}_R, \quad V := \phi_Z(Z) \in \mathcal{H}_Z$. By the fourth–moment assumption we have $\mathbb{E}\|U\|_{\mathcal{H}_R}^2 < \infty$ and $\mathbb{E}\|V\|_{\mathcal{H}_Z}^2 < \infty$, so the Bochner conditional expectations $\mathbb{E}[U \mid X]$ and $\mathbb{E}[V \mid X]$ and the tensor product $U \otimes V \in \mathcal{H}_R \otimes \mathcal{H}_Z$ are well defined. Define $\widetilde{\phi}_R := U - \mathbb{E}[U \mid X], \quad \widetilde{\phi}_Z := V - \mathbb{E}[V \mid X]$, so that $\mathbb{E}[\widetilde{\phi}_R \mid X] = \mathbb{E}[\widetilde{\phi}_Z \mid X] = 0$ a.s. The residual cross–covariance operator is

$$C_{RZ}^{\text{res}} := \mathbb{E}\Big[(U - \mathbb{E}[U \mid X]) \otimes (V - \mathbb{E}[V \mid X])\Big] \in \mathcal{H}_R \otimes \mathcal{H}_Z.$$

Expanding and using the tower property plus X–measurability of the conditional expectations,

$$\begin{aligned} C_{RZ}^{\text{res}} &= \mathbb{E}[U \otimes V] - \mathbb{E}\big[\mathbb{E}[U \mid X] \otimes V\big] - \mathbb{E}\big[U \otimes \mathbb{E}[V \mid X]\big] \\ &\quad + \mathbb{E}\big[\mathbb{E}[U \mid X] \otimes \mathbb{E}[V \mid X]\big] \\ &= \mathbb{E}[U \otimes V] - \mathbb{E}\big[\mathbb{E}[U \mid X] \otimes \mathbb{E}[V \mid X]\big]. \end{aligned}$$

Under $Z \perp\!\!\!\perp R \mid X$ we have, a.s.,

$$\mathbb{E}[U \otimes V \mid X] = \mathbb{E}[U \mid X] \otimes \mathbb{E}[V \mid X].$$

Taking expectations yields

$$\mathbb{E}[U \otimes V] = \mathbb{E}\big[\mathbb{E}[U \mid X] \otimes \mathbb{E}[V \mid X]\big],$$

and therefore $C_{RZ}^{\text{res}} = 0$. By the adopted normalization for distance covariance on Hilbert–valued arguments,

$$\mathrm{dCov}_\star^2(\widetilde{\phi}_R, \widetilde{\phi}_Z) = \|\mathrm{Cov}(\widetilde{\phi}_R, \widetilde{\phi}_Z)\|_{\text{HS}}^2 = \|C_{RZ}^{\text{res}}\|_{\text{HS}}^2 = 0.$$

Thus $\mathrm{PAS}_\star = \mathrm{dCor}(\widetilde{\phi}_R, \widetilde{\phi}_Z) = 0$, with the convention that $\mathrm{dCor} = 0$ if a distance variance vanishes.

Sample. Let $\{(R_i, Z_i, X_i)\}_{i=1}^n$ be i.i.d. copies of (R, Z, X) and define

$$U_n := f_{R,n}(R, Z, X), \qquad V_n := f_{Z,n}(R, Z, X).$$

By assumption, $U_n \to \widetilde{\phi}_R, \quad V_n \to \widetilde{\phi}_Z$ in L_2, so in particular $U_n \to \widetilde{\phi}_R$ and $V_n \to \widetilde{\phi}_Z$ in L_1, and

$$\sup_n \mathbb{E}\|U_n\|_{\mathcal{H}_R}^2 < \infty, \qquad \sup_n \mathbb{E}\|V_n\|_{\mathcal{H}_Z}^2 < \infty.$$

The same uniform fourth moment assumption implies

$$\sup_n \mathbb{E}\|U_n\|_{\mathcal{H}_R}^4 < \infty, \qquad \sup_n \mathbb{E}\|V_n\|_{\mathcal{H}_Z}^4 < \infty.$$

As in the population case, the covariance of Hilbert–valued variables is continuous with respect to L_2: from $U_n, V_n \to \widetilde{\phi}_R, \widetilde{\phi}_Z$ in L_2, it follows that $\mathrm{Cov}(U_n, V_n) \longrightarrow \mathrm{Cov}(\widetilde{\phi}_R, \widetilde{\phi}_Z)$ in Hilbert–Schmidt norm. By continuity of the norm,

$$\mathrm{dCov}_\star^2(U_n, V_n) = \|\mathrm{Cov}(U_n, V_n)\|_{\text{HS}}^2 \longrightarrow$$
$$\|\mathrm{Cov}(\widetilde{\phi}_R, \widetilde{\phi}_Z)\|_{\text{HS}}^2 = \mathrm{dCov}_\star^2(\widetilde{\phi}_R, \widetilde{\phi}_Z) = 0.$$

An analogous argument with $U_n = V_n$ yields the convergence of the distance variances associated with U_n and V_n towards those of $\widetilde{\phi}_R$ and $\widetilde{\phi}_Z$.

For each n, the pairs

$$\Big(\widetilde{\widetilde{\phi}}_R^{(i)}, \widetilde{\widetilde{\phi}}_Z^{(i)}\Big) = \Big(f_{R,n}(R_i, Z_i, X_i), f_{Z,n}(R_i, Z_i, X_i)\Big), \qquad i = 1, \dots, n,$$

are i.i.d. with the same distribution as (U_n, V_n). The sample distance covariance and the sample distance variances are finite–order U– or V–statistics constructed from these pairs via a kernel h which is a polynomial in the distances $\|U_n^{(i)} - U_n^{(j)}\|, \|V_n^{(i)} - V_n^{(j)}\|$. From the uniform fourth moment assumptions on U_n and V_n, it follows

$$\sup_n \mathbb{E}\big[h^2\big] < \infty,$$

i.e., the kernel has a second moment uniformly bounded in n. The usual variance bounds for U– and V–statistics then give

$$\mathrm{Var}\Big(\widehat{\mathrm{dCov}}_\star^2(\widetilde{\widetilde{\phi}}_R, \widetilde{\widetilde{\phi}}_Z)\Big) \le \frac{C}{n}, \qquad \mathrm{Var}\Big(\widehat{\mathrm{dVar}}_\star(\widetilde{\widetilde{\phi}}_\bullet)\Big) \le \frac{C'}{n},$$

for some constants C, C' independent of n and for $\bullet \in \{R, Z\}$. In particular,

$$\widehat{\mathrm{dCov}}_\star^2(\widetilde{\widetilde{\phi}}_R, \widetilde{\widetilde{\phi}}_Z) - \mathrm{dCov}_\star^2(U_n, V_n) \xrightarrow{\mathbb{P}} 0,$$
$$\widehat{\mathrm{dVar}}_\star(\widetilde{\widetilde{\phi}}_\bullet) - \mathrm{dVar}_\star(U_n) \xrightarrow{\mathbb{P}} 0,$$

and similarly for V_n.

Combining these convergences with those of the population functionals (two applications of Slutsky's theorem), we obtain

$$\widehat{\mathrm{dCov}}_\star^2(\widetilde{\widetilde{\phi}}_R, \widetilde{\widetilde{\phi}}_Z) \xrightarrow{\mathbb{P}} \mathrm{dCov}_\star^2(\widetilde{\phi}_R, \widetilde{\phi}_Z) = 0,$$
$$\widehat{\mathrm{dVar}}_\star(\widetilde{\widetilde{\phi}}_\bullet) \xrightarrow{\mathbb{P}} \mathrm{dVar}_\star(\widetilde{\phi}_\bullet),$$

for $\bullet \in \{R, Z\}$. If $\mathrm{dVar}_\star(\widetilde{\phi}_R)\,\mathrm{dVar}_\star(\widetilde{\phi}_Z) > 0$, the continuity of $(a, b, c) \mapsto a/\sqrt{bc}$ and the continuous mapping theorem imply

$$\widehat{\mathrm{PAS}} = \widehat{\mathrm{dCor}}(\widetilde{\widetilde{\phi}}_R, \widetilde{\widetilde{\phi}}_Z) \xrightarrow{\mathbb{P}} \mathrm{dCor}(\widetilde{\phi}_R, \widetilde{\phi}_Z) = 0.$$

If one of the limit distance variances is 0, we adopt the same convention at the sample level (setting $\widehat{\mathrm{dCor}} = 0$ when an estimated distance variance is zero), and the convergence to 0 remains valid. In all cases $\widehat{\mathrm{PAS}} \xrightarrow{\mathbb{P}} 0$ as $n \to \infty$. □

Theorem 3 (Quadratic sensitivity of squared PAS under vanishing signals). *Assume bounded kernels k_X, k_R, k_Z with finite second moments; k_R, k_Z are characteristic and distance-induced. Assume the local submodel $R_\delta = \varphi(g(X) + \delta\, h(Z) + \eta)$ is DQM at $\delta = 0$ with score S, and $\varphi \in C^1$ has strictly positive bounded derivative on the support. Assume realizability for the RKHS L_2 residualization: $m_R(x) := \mathbb{E}[\phi_R(R) \mid X = x] \in \overline{\mathcal{G}}_R$ and $m_Z(x) := \mathbb{E}[\phi_Z(Z) \mid X = x] \in \overline{\mathcal{G}}_Z$. Assume moreover $\eta \perp\!\!\!\perp Z \mid X$, $\mathbb{E}[\eta \mid X, Z] = 0$, $\mathrm{Var}(\eta \mid X, Z) > 0$ a.s., Z is nondegenerate, and $h(Z)$ is nonconstant given X on a set of positive $\mathbb{P}_X$–measure.*

Explicit non-degeneracy condition. *Let $\mathbb{P}_0$ be the null distribution (i.e., at $\delta = 0$), and define the residualized features under $\mathbb{P}_0$:*

$$\widetilde{\phi}_R^0 := \phi_R(R_0) - \mathbb{E}_0[\phi_R(R_0) \mid X], \qquad \widetilde{\phi}_Z^0 := \phi_Z(Z) - \mathbb{E}_0[\phi_Z(Z) \mid X].$$

Equivalently, $\mathbb{E}_0[\cdot \mid X]$ is the $L_2(\mathbb{P}_0)$ orthogonal projection onto the closed subspace of $\sigma(X)$-measurable Hilbert-valued functions. Set

$$A := \mathbb{E}_0\big[(S\,\widetilde{\phi}_R^0) \otimes \widetilde{\phi}_Z^0\big] + \mathbb{E}_0\big[\widetilde{\phi}_R^0 \otimes (S\,\widetilde{\phi}_Z^0)\big] \in \mathcal{H}_R \otimes \mathcal{H}_Z,$$

and assume

$$\|A\|_{\text{HS}} > 0 \qquad \Big(\textit{equivalently } \mathbb{E}_0\big[h_0(W_1, W_2)\,(S_1 + S_2)^2\big] > 0\Big),$$

where $W = (R, Z, X)$, $S_i := S(W_i)$, and

$$h_0(W_1, W_2) := \big\langle \widetilde{\phi}_R^0(W_1), \widetilde{\phi}_R^0(W_2) \big\rangle_{\mathcal{H}_R} \big\langle \widetilde{\phi}_Z^0(W_1), \widetilde{\phi}_Z^0(W_2) \big\rangle_{\mathcal{H}_Z}.$$

Then, after residualization w.r.t. X,

$$PAS^2_\star(\delta) = c\,\delta^2 + o(\delta^2), \qquad c > 0.$$

Proof. We work at the population level. In the local submodel, only R depends on δ; (Z, X) do not depend on δ.

For arbitrary δ, define the residualized features

$$\widetilde{\phi}_R(\delta) := \phi_R(R_\delta) - \mathbb{E}_\delta[\phi_R(R_\delta) \mid X],\ \widetilde{\phi}_Z := \phi_Z(Z) - \mathbb{E}_0[\phi_Z(Z) \mid X].$$

We also fix population-centering at the null: let $k^{\mathrm{c}}_R, k^{\mathrm{c}}_Z$ be the kernels centered under $\mathbb{P}_0$, so that $\mathbb{E}_0[\widetilde{\phi}_R(0)] = 0$ and $\mathbb{E}_0[\widetilde{\phi}_Z] = 0$.

Let the residual cross-covariance operator be

$$C^{\mathrm{res}}_{RZ}(\mathbb{P}_\delta) := \mathbb{E}_\delta\big[\widetilde{\phi}_R(\delta) \otimes \widetilde{\phi}_Z\big].$$

With distance-induced kernels, fix the normalization so that $c_d = 1$ in the dCov–RKHS equivalence:

$$\Psi(\mathbb{P}_\delta) := \mathrm{dCov}^2_\star\big(\widetilde{\phi}_R(\delta), \widetilde{\phi}_Z\big) = \big\|C^{\mathrm{res}}_{RZ}(\mathbb{P}_\delta)\big\|^2_{\mathrm{HS}}.$$

By realizability and null alignment, under $H_0 : Z \perp\!\!\!\perp R \mid X$ (i.e., $\delta = 0$) we have $C^{\mathrm{res}}_{RZ}(\mathbb{P}_0) = 0$; hence the sample counterpart is a symmetric order-2 U-statistic with first Hoeffding projection equal to zero (order-1 degeneracy). Define, for each δ, the order-2 kernel induced by the residualized features:

$$h_\delta(W_1, W_2) := \big\langle \widetilde{\phi}_R(\delta; W_1), \widetilde{\phi}_R(\delta; W_2)\big\rangle_{\mathcal{H}_R} \big\langle \widetilde{\phi}_Z(W_1), \widetilde{\phi}_Z(W_2)\big\rangle_{\mathcal{H}_Z},$$

so that $\Psi(\mathbb{P}_\delta) = \mathbb{E}_\delta[h_\delta(W_1, W_2)]$. By L_2-continuity of $\delta \mapsto \widetilde{\phi}_R(\delta)$ at 0 and boundedness of the kernels,

$$\mathbb{E}_\delta[h_\delta(W_1, W_2)] - \mathbb{E}_\delta[h_0(W_1, W_2)] = o(\delta^2).$$

Therefore, $\Psi(\mathbb{P}_\delta) = \mathbb{E}_\delta[h_0(W_1, W_2)] + o(\delta^2)$.

Consider the local parametric submodel $\{\mathbb{P}_\delta : \delta \in (-\delta_0, \delta_0)\}$ that is DQM at $\delta = 0$ with score $S(W)$. Applying the second-order Le Cam expansion to the linear functional $\delta \mapsto \mathbb{E}_\delta[h_0(W_1, W_2)]$ (with respect to the product measure $\mathbb{P}^{\otimes 2}_\delta$), order-1 degeneracy removes the first-order term and yields

$$\mathbb{E}_\delta[h_0(W_1, W_2)] = \frac{\delta^2}{2}\,\mathbb{E}_0\big[h_0(W_1, W_2)\,(S(W_1) + S(W_2))^2\big] + o(\delta^2).$$

Hence $\Psi(\mathbb{P}_\delta) = \frac{\delta^2}{2}\,\mathbb{E}_0\big[h_0(W_1, W_2)\,(S(W_1) + S(W_2))^2\big] + o(\delta^2)$. By Lemma 1, $\mathbb{E}_0\big[h_0(W_1, W_2)\,(S_1 + S_2)^2\big] = \frac{1}{2}\,\|A\|^2_{\mathrm{HS}}$. Therefore, $\Psi(\mathbb{P}_\delta) = \frac{1}{4}\,\delta^2\,\|A\|^2_{\mathrm{HS}} + o(\delta^2)$. By the explicit non-degeneracy condition, $\|A\|_{\mathrm{HS}} > 0$. Set $c_0 := \frac{1}{4}\,\|A\|^2_{\mathrm{HS}} > 0$, so

$$\Psi(\mathbb{P}_\delta) = c_0\,\delta^2 + o(\delta^2).$$

Since $\varphi \in C^1$ with strictly positive bounded derivative and moments are finite, $\delta \mapsto \phi_R(R_\delta)$ is L_2-continuous; consequently the residual distance variances

$$v_R(\delta) := \mathrm{dCov}_\star\big(\widetilde{\phi}_R(\delta), \widetilde{\phi}_R(\delta)\big), \qquad v_Z(\delta) := \mathrm{dCov}_\star\big(\widetilde{\phi}_Z, \widetilde{\phi}_Z\big)$$

are continuous at 0. Moreover, $v_Z(0) > 0$ because Z is nondegenerate, and $v_R(0) > 0$ under $\mathrm{Var}(\eta \mid X, Z) > 0$. By definition of PAS,

$$\mathrm{PAS}^2_\star(\delta) = \frac{\Psi(\mathbb{P}_\delta)}{v_R(\delta)\,v_Z(\delta)} = \frac{c_0}{v_R(0)\,v_Z(0)}\,\delta^2 + o(\delta^2).$$

Finally, set $c := c_0/\{v_R(0)\,v_Z(0)\} > 0$,

which completes the proof. □

Lemma 1 (Polarization identity for residual RKHS features). *Let $W = (R, Z, X) \sim \mathbb{P}_0$, and let $(\mathcal{H}_R, \langle\cdot,\cdot\rangle_{\mathcal{H}_R})$ and $(\mathcal{H}_Z, \langle\cdot,\cdot\rangle_{\mathcal{H}_Z})$ be separable Hilbert spaces. Consider measurable feature maps $\phi_R : \mathsf{R} \to \mathcal{H}_R$ and $\phi_Z : \mathsf{Z} \to \mathcal{H}_Z$ with $\mathbb{E}_0\|\phi_R(R)\|^2_{\mathcal{H}_R} < \infty$ and $\mathbb{E}_0\|\phi_Z(Z)\|^2_{\mathcal{H}_Z} < \infty$. Define*

$$\widetilde{\phi}^0_R := \phi_R(R) - \mathbb{E}_0[\phi_R(R) \mid X], \qquad \widetilde{\phi}^0_Z := \phi_Z(Z) - \mathbb{E}_0[\phi_Z(Z) \mid X].$$

Assume (null alignment) $\mathbb{E}_0[\widetilde{\phi}^0_R \otimes \widetilde{\phi}^0_Z] = 0$ in $\mathcal{H}_R \otimes \mathcal{H}_Z$. Let $S \in L_2(\mathbb{P}_0)$ be real-valued and assume the integrability condition

$$\mathbb{E}_0\big[(1 + S(W)^2)\,\|\widetilde{\phi}^0_R(W)\|_{\mathcal{H}_R}\,\|\widetilde{\phi}^0_Z(W)\|_{\mathcal{H}_Z}\big] < \infty.$$

For i.i.d. copies W_1, W_2 of W, define

$$h_0(W_1, W_2) := \big\langle \widetilde{\phi}^0_R(W_1), \widetilde{\phi}^0_R(W_2)\big\rangle_{\mathcal{H}_R} \big\langle \widetilde{\phi}^0_Z(W_1), \widetilde{\phi}^0_Z(W_2)\big\rangle_{\mathcal{H}_Z}.$$

Write $S_i := S(W_i)$ and $\widetilde{\phi}^{0,i}_\bullet := \widetilde{\phi}^0_\bullet(W_i)$. Then

$$\mathbb{E}_0\big[h_0(W_1, W_2)\,(S_1 + S_2)^2\big] = \tfrac{1}{2}\,\|A\|^2_{\mathrm{HS}},$$

where we use the canonical isometric identification $\mathcal{H}_R \otimes \mathcal{H}_Z \simeq \mathrm{HS}(\mathcal{H}_Z, \mathcal{H}_R)$, and $A := \mathbb{E}_0[(S\,\widetilde{\phi}^0_R) \otimes \widetilde{\phi}^0_Z] + \mathbb{E}_0[\widetilde{\phi}^0_R \otimes (S\,\widetilde{\phi}^0_Z)]$.

Proof. All expectations are under $\mathbb{P}_0$.

Integrability. By independence of (W_1, W_2) and Cauchy–Schwarz,

$$\begin{aligned}\mathbb{E}_0|h_0(W_1, W_2)| &\le \mathbb{E}_0\big[\|\widetilde{\phi}^{0,1}_R\|_{\mathcal{H}_R}\,\|\widetilde{\phi}^{0,1}_Z\|_{\mathcal{H}_Z}\,\|\widetilde{\phi}^{0,2}_R\|_{\mathcal{H}_R}\,\|\widetilde{\phi}^{0,2}_Z\|_{\mathcal{H}_Z}\big] \\ &= \Big(\mathbb{E}_0\big[\,\|\widetilde{\phi}^0_R\|_{\mathcal{H}_R}\,\|\widetilde{\phi}^0_Z\|_{\mathcal{H}_Z}\,\big]\Big)^2 \le \big(\mathbb{E}_0\|\widetilde{\phi}^0_R\|^2_{\mathcal{H}_R}\big)\big(\mathbb{E}_0\|\widetilde{\phi}^0_Z\|^2_{\mathcal{H}_Z}\big) < \infty,\end{aligned}$$

so $h_0 \in L_1$. Moreover, independence gives

$$\mathbb{E}_0\big[|h_0|\,S_1^2\big] \le \mathbb{E}_0\big[\,S^2\,\|\widetilde{\phi}^0_R\|_{\mathcal{H}_R}\,\|\widetilde{\phi}^0_Z\|_{\mathcal{H}_Z}\big]\,\mathbb{E}_0\big[\,\|\widetilde{\phi}^0_R\|_{\mathcal{H}_R}\,\|\widetilde{\phi}^0_Z\|_{\mathcal{H}_Z}\big] < \infty,$$

and similarly for S_2^2. Hence the tower property and Fubini–Tonelli (Bochner) apply. In particular,

$$T := \mathbb{E}_0[S\,\widetilde{\phi}^0_R \otimes \widetilde{\phi}^0_Z] \in \mathcal{H}_R \otimes \mathcal{H}_Z$$

is well defined.

Preparation. With the canonical identification,

$$h_0(W_1, W_2) = \big\langle \widetilde{\phi}^{0,1}_R \otimes \widetilde{\phi}^{0,1}_Z,\ \widetilde{\phi}^{0,2}_R \otimes \widetilde{\phi}^{0,2}_Z\big\rangle_{\mathrm{HS}}.$$

(i) First-order degeneracy. Using the previous identity and $\mathbb{E}_0[\widetilde{\phi}^0_R \otimes \widetilde{\phi}^0_Z] = 0$,

$$\mathbb{E}_0\big[h_0(W_1, W_2) \mid W_1\big] = \big\langle \widetilde{\phi}^{0,1}_R \otimes \widetilde{\phi}^{0,1}_Z,\ \mathbb{E}_0[\widetilde{\phi}^0_R \otimes \widetilde{\phi}^0_Z]\big\rangle_{\mathrm{HS}} = 0,$$

and symmetrically for $\mid W_2$. Therefore

$$\mathbb{E}_0[h_0\,S_1^2] = \mathbb{E}_0\big[\mathbb{E}_0[h_0 \mid W_1]\,S_1^2\big] = 0, \qquad \mathbb{E}_0[h_0\,S_2^2] = 0.$$

(ii) Cross term. By independence of (W_1, W_2) and Fubini–Tonelli (Bochner),

$$\mathbb{E}_0[h_0\,S_1S_2] = \big\langle \mathbb{E}_0[S\,\widetilde{\phi}^0_R \otimes \widetilde{\phi}^0_Z],\ \mathbb{E}_0[S\,\widetilde{\phi}^0_R \otimes \widetilde{\phi}^0_Z]\big\rangle_{\mathrm{HS}} = \|T\|^2_{\mathrm{HS}}.$$

(iii) Conclusion. Hence

$$\mathbb{E}_0\big[h_0(W_1, W_2)\,(S_1 + S_2)^2\big] = 2\,\mathbb{E}_0[h_0\,S_1S_2] = 2\,\|T\|^2_{\mathrm{HS}}.$$

Since S is scalar, $A = 2\,\mathbb{E}_0[S\,\widetilde{\phi}^0_R \otimes \widetilde{\phi}^0_Z] = 2T$, so $\|A\|^2_{\mathrm{HS}} = 4\,\|T\|^2_{\mathrm{HS}}$ and therefore

$$\mathbb{E}_0\big[h_0(W_1, W_2)\,(S_1 + S_2)^2\big] = 2\,\|T\|^2_{\mathrm{HS}} = \tfrac{1}{2}\,\|A\|^2_{\mathrm{HS}},$$

which completes the proof. □

REFERENCES

[1] Philip W Adler, Casey Falk, Sorelle A Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing black-box models for indirect influence. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 1–10.

[2] amar5693. 2024. Screen Time, Sleep and Stress Analysis Dataset. https://www.kaggle.com/datasets/amar5693/screen-time-sleep-and-stress-analysis-dataset. Accessed: 2026-02-27.

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And It's Biased Against Blacks. *ProPublica* (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] Abolfazl Asudeh, H. V. Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing Fair Ranking Schemes. In *SIGMOD*. 1259–1276.

[5] Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. 2018. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*. 405–414.

[6] Francesco Bonchi, Sara Hajian, Bud Mishra, and Daniele Ramazzotti. 2017. Exposing the probabilistic causal structure of discrimination. *Int. J. Data Sci. Anal.* 3, 1 (2017), 1–21.

[7] Emmanuel J. Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. 2018. Panning for gold: 'Model-X' knockoffs for high-dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80, 3 (2018), 551–577.

[8] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. 2018. Ranking with Fairness Constraints. In *ICALP*.

[9] John J. Cherian and Emmanuel J. Candès. 2024. Statistical Inference for Fairness Auditing. *Journal of Machine Learning Research* 25, 149 (2024), 1–49. https://www.jmlr.org/papers/v25/23-0739.html Earlier version: arXiv:2305.03712.

[10] P. Cortez and Alice Maria Gonçalves Silva. 2008. Using data mining to predict secondary school student performance. https://api.semanticscholar.org/CorpusID:16621299

[11] Thomas M. Cover and Joy A. Thomas. 2006. *Elements of Information Theory* (2nd ed.). Wiley, Hoboken, NJ.

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

[13] Petros Drineas and Michael W. Mahoney. 2005. On the Nyström Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *Journal of Machine Learning Research* 6 (2005), 2153–2175.

[14] Dheeru Dua and Casey Graff. 2019. Adult (Census Income) Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Adult University of California, Irvine. Donors: Barry Becker and Ron Kohavi.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. 214–226.

[16] Cynthia Dwork, Michael P. Kim, Omer Reingold, Guy N. Rothblum, and Gal Yona. 2019. Learning from Outcomes: Evidence-Based Rankings. In *FOCS*. 106–125.

[17] Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, New York.

[18] U.S. Equal Employment Opportunity Commission. [n.d.]. https://www.eeoc.gov/laws/index.cfm.

[19] European Commission Diversity Charters. [n.d.]. https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/tackling-discrimination/diversity-management/diversity-charters-eu-country_en.

[20] Sheila R Foster. 2004. Causation in antidiscrimination law: Beyond intent versus impact. *Hous. L. Rev.* 41 (2004), 1469.

[21] Sebastian Frenzel and Bernd Pompe. 2007. Partial Mutual Information for Coupling Analysis of Multivariate Time Series. *Physical Review Letters* 99, 20 (2007), 204101.

[22] David García-Soriano and Francesco Bonchi. 2021. Maxmin-Fair Ranking: Individual Fairness under Group-Fairness Constraints. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21)*. 436–446.

[23] Dan Geiger, Thomas Verma, and Judea Pearl. 1990. d-separation: From theorems to algorithms. In *Machine intelligence and pattern recognition*. Vol. 10. Elsevier, 139–148.

[24] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-Aware Ranking in Search & Recommendation Systems with Application to LinkedIn Talent Search. In *KDD*. 2221–2231.

[25] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. 2005. Measuring statistical dependence with Hilbert-Schmidt norms. In *International conference on algorithmic learning theory*. Springer, 63–77.

[26] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard Schölkopf, and Alex Smola. 2007. A kernel statistical test of independence. *Advances in neural information processing systems* 20 (2007).

[27] Nathan Halko, Per-Gunnar Martinsson, and Joel A. Tropp. 2011. Finding Structure with Randomness: Probabilistic Algorithms for Constructing Approximate Matrix Decompositions. *SIAM Rev.* 53, 2 (2011), 217–288.

[28] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*. 3315–3323.

[29] Magnus R. Hestenes and Eduard Stiefel. 1952. Methods of Conjugate Gradients for Solving Linear Systems. *J. Res. Nat. Bur. Standards* 49, 6 (1952), 409–436.

[30] H. V. Jagadish, Francesco Bonchi, Tina Eliassi-Rad, Lise Getoor, Krishna Gummadi, and Julia Stoyanovich. 2019. The Responsibility Challenge for Data. In *Proceedings of the 2019 International Conference on Management of Data* (Amsterdam, Netherlands) *(SIGMOD '19)*. 412–414.

[31] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446. https://doi.org/10.1145/582415.582418

[32] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E* 69, 6 (2004), 066138. https://doi.org/10.1103/PhysRevE.69.066138

[33] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E* 69, 6 (2004), 066138.

[34] Erwin Kreyszig. 1989. *Introductory Functional Analysis with Applications*. John Wiley & Sons.

[35] Erich L. Lehmann and Joseph P. Romano. 2005. *Testing Statistical Hypotheses* (3rd ed.). Springer.

[36] Shira Mitchell, Eric Potash, Solon Barocas, Alexander D'Amour, and Kristian Lum. 2021. Algorithmic Fairness: Choices, Assumptions, and Definitions. *Annual Review of Statistics and Its Application* 8, 1 (2021), 141–163.

[37] Harikrishna Narasimhan, Andrew Cotter, Maya R. Gupta, and Serena Wang. 2020. Pairwise Fairness for Ranking and Regression. In *AAAI*.

[38] Christopher C. Paige and Michael A. Saunders. 1975. Solution of Sparse Indefinite Systems of Linear Equations. *SIAM J. Numer. Anal.* 12, 4 (1975), 617–629.

[39] Sambit Panda, Satish Palaniappan, Junhao Xiong, Eric W Bridgeford, Ronak Mehta, Cencheng Shen, and Joshua T Vogelstein. 2019. hyppo: A multivariate hypothesis testing Python package. *arXiv preprint arXiv:1907.02088* (2019).

[40] Eliana Pastor and Francesco Bonchi. 2024. Intersectional fair ranking via subgroup divergence. *Data Min. Knowl. Discov.* 38, 4 (2024), 2186–2222.

[41] Gourab K. Patro, Lorenzo Porcaro, Laura Mitchell, Qiuyue Zhang, Meike Zehlike, and Nikhil Garg. 2022. Fair Ranking: A Critical Review, Challenges, and Future Directions. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. 1929–1942.

[42] Judea Pearl. 1995. Causal diagrams for empirical research. *Biometrika* 82, 4 (1995), 669–688.

[43] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. 2022. Fairness in rankings and recommendations: an overview. *The VLDB Journal* 31, 3 (2022), 431–458.

[44] Salvatore Ruggieri and Jose Manuel Alvarez. 2023. Counterfactual Situation Testing: Uncovering Discrimination under Fairness given the Difference. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '23)*. ACM, 1–11. https://doi.org/10.1145/3617694.3623222

[45] Yousef Saad. 2003. *Iterative Methods for Sparse Linear Systems* (2 ed.). Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

[46] Yuta Saito and Thorsten Joachims. 2022. Fair ranking as fair division: Impact-based individual fairness in ranking. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1514–1524.

[47] Babak Salimi, Luke Rodriguez, Bill Howe, and Dan Suciu. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*. ACM, 793–810.

[48] Bernhard Schölkopf and Alexander J. Smola. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

[49] John Shawe-Taylor and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge, UK.

[50] Ashudeep Singh and Thorsten Joachims. 2018. Fairness of Exposure in Rankings. In *KDD*. 2219–2228.

[51] Ashudeep Singh and Thorsten Joachims. 2019. Policy Learning for Fairness in Ranking. In *NeurIPS*. 5427–5437.

[52] Igor Stepin, Jose M Oramas, and Alejandro Coca-Catalina. 2021. Certifai: Counterfactual explanations for black-box models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1518–1527.

[53] Gábor J Székely and Maria L Rizzo. 2014. Partial distance correlation with methods for dissimilarities. *The Annals of Statistics* 42, 6 (2014), 2382–2412.

[54] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. 2007. Measuring and testing dependence by correlation of distances. *Annals of Statistics* 35, 6 (2007), 2769–2794.

[55] Martin Vejmelka and Milan Paluš. 2008. Inferring the Directionality of Coupling with Conditional Mutual Information. *Physical Review E* 77, 2 (2008), 026214.

[56] Greg Ver Steeg. 2000. Non-parametric entropy estimation toolbox (npeet). *Non-parametric entropy estimation toolbox (NPEET)* (2000).

[57] Xueqin Wang, Wenliang Pan, Wenhao Hu, Yuan Tian, and Heping Zhang. 2015. Conditional distance correlation. *J. Amer. Statist. Assoc.* 110, 512 (2015), 1726–1734.
[58] Christopher K. I. Williams and Matthias Seeger. 2001. Using the Nyström Method to Speed Up Kernel Machines. In *Advances in Neural Information Processing Systems*, Vol. 13. 682–688.
[59] Lei Xu, Maria Skoularidou, Alfredo Cuesta-Infante, and Kalyan Veeramachaneni. 2019. Modeling Tabular Data Using Conditional GAN. *arXiv preprint arXiv:1907.00503* (2019).
[60] Ke Yang, Vasilis Gkatzelis, and Julia Stoyanovich. 2019. Balanced Ranking with Diversity Constraints. In *IJCAI*. 6035–6042.
[61] Ke Yang and Julia Stoyanovich. 2017. Measuring Fairness in Ranked Outputs. In *SSDBM*. 22:1–22:6.
[62] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A Fair Top-k Ranking Algorithm. In *CIKM*. 1569–1578.
[63] Meike Zehlike, Philipp Hacker, and Emil Wiedemann. 2020. Matching code and law: achieving algorithmic fairness with optimal transport. *Data Mining and Knowledge Discovery* 34, 1 (2020), 163–200.
[64] Meike Zehlike, Ke Yang, and Julia Stoyanovich. 2022. Fairness in ranking, part i: Score-based ranking. *Comput. Surveys* 55, 6 (2022), 1–36.
[65] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2011. Kernel-based Conditional Independence Test and Application in Causal Discovery. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*. AUAI Press, Barcelona, Spain, 804–813.
[66] Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. 2024. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research* 25, 60 (2024), 1–8.