

Document Version

Final published version

Licence

Dutch Copyright Act (Article 25fa)

Citation (APA)

Cavallo, A., Navarro, M., Segarra, S., & Isufi, E. (2025). Fair CoVariance Neural Networks. In B. D. Rao, I. Trancoso, G. Sharma, & N. B. Mehta (Eds.), *2025 IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2025 - Proceedings* IEEE. <https://doi.org/10.1109/ICASSP49660.2025.10888282>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

Fair CoVariance Neural Networks

Andrea Cavallo*, Madeline Navarro†, Santiago Segarra†, Elvin Isufi*

*Delft University of Technology, Delft, Netherlands

†Rice University, Houston, TX, USA

Abstract—Covariance-based data processing is widespread across signal processing and machine learning applications due to its ability to model data interconnectivities and dependencies. However, harmful biases in the data may become encoded in the sample covariance matrix and cause data-driven methods to treat different subpopulations unfairly. Existing works such as fair principal component analysis (PCA) mitigate these effects, but remain unstable in low sample regimes, which in turn may jeopardize the fairness goal. To address both biases and instability, we propose Fair coVariance Neural Networks (FVNNs), which perform graph convolutions on the covariance matrix for both fair and accurate predictions. Our FVNNs provide a flexible model compatible with several existing bias mitigation techniques. In particular, FVNNs allow for mitigating the bias in two ways: first, they operate on fair covariance estimates that remove biases from their principal components; second, they are trained in an end-to-end fashion via a fairness regularizer in the loss function so that the model parameters are tailored to solve the task directly in a fair manner. We prove that FVNNs are intrinsically fairer than analogous PCA approaches thanks to their stability in low sample regimes. We validate the robustness and fairness of our model on synthetic and real-world data, showcasing the flexibility of FVNNs along with the tradeoff between fair and accurate performance.

Index Terms—Covariance neural networks, fair machine learning, fair PCA

I. INTRODUCTION

Covariance-based learning has a long-standing history as an approach to conveniently model critical information about observed data, boasting success in several applications ranging from brain connectivity estimation [1], [2] to blind source separation [3], [4] and financial data analysis [5], [6]. For example, the covariance matrix is the foundation of principal component analysis (PCA) [7], the prevailing approach for summarizing high-dimensional data via dimension reduction. PCA exploits the eigenvectors of the covariance matrix, termed principal components (PCs), which denote primary directions of spatially distributed data. Beyond PCA, the theoretical and empirical advantages of graph neural networks (GNNs) [8]–[10] have led to the development of covariance neural networks (VNNs), where the covariance matrix is seen as the input graph for a GNN [11]. By spectral graph theory, VNNs can be viewed as an extension of PCA with learnable weights assigned to PCs [11, Theorem 1]; they are transferable across datasets [12] and effective in temporal and sparse settings [13], [14] and for applications to brain data [15]–[17]. Moreover, VNNs are provably stable to covariance estimation errors in low sample regimes [11, Theorem 2], while PCA-based data processing may encounter unexpected behavior if the estimated PCs differ greatly from the true ones [7].

This work was partially supported by the NSF under award CCF-2340481. Research was sponsored by the Army Research Office and was accomplished under Grant Number W911NF-17-S-0002. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Army or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein. Part of this work was supported by the TU Delft AI Labs programme, the NWO OTP GraSPA proposal #19497, and the NWO VENI proposal 222.032. Emails: a.cavallo@tudelft.nl, nav@rice.edu, segarra@rice.edu, e.isufi-1@tudelft.nl

These tools have shown great success in extracting rich information from correlated data. However, real-world data often contains harmful biases such as poor representation of certain communities, resulting in disparate treatment across subpopulations [18]. If relevant information is correlated with sensitive attributes, then PCs can encode these biases. Such PCs may yield inaccurate representations or increased discrepancies in treatment of different groups. For example, segregation may worsen if observed data experiences shifts in distribution across groups [19]. Additionally, representations of underrepresented subpopulations may be far more inaccurate, resulting in overdependence on majority groups [19], [20].

Fair learning methods promote unbiased treatment for data-driven tools, with recent interest emerging for fairness in all steps of data processing pipelines, from data representation to predictions [21]–[23]. While fairness for predictions is the most well-studied task [24], [25], attention of late has turned towards unbiased representation learning [26], as several real-world datasets exhibit preferential treatment due to unequal representation of different groups [20]. As the pervasiveness of high-dimensional data increases, recent works attempt dimensionality reduction while mitigating biases in data [19], [20], [27]. Fair variants of PCA have shown success in reducing biases in projected data, where the goal is either to (i) obtain group-agnostic projections for fair downstream tasks [19], [27], [28] or (ii) to encourage equitable representation accuracy across groups [20], [29]–[31]. Such methods are often inefficient to compute or return suboptimal solutions [32]–[35]. Additionally, the sensitivity of PCA to outliers and insufficient data remains a challenge since fair PCA approaches are still unstable to minor perturbations.

Contributions. Motivated by the value of covariance-based learning and the challenge of removing biases in PCs, we propose *fair VNNs* (FVNNs) to exploit the advantages of VNNs for covariance-based learning in unfair settings. Exploiting the flexibility of VNNs, we introduce fairness by (i) estimating a fairer version of the covariance matrix and (ii) penalizing biases in the training loss. Given an unbiased covariance estimate, we prove that the natural stability of VNNs leads to fairer outcomes than fair PCA when groups follow different distributions. Moreover, tuning the weight of the loss penalty allows for flexible control of the tradeoff between fairness and accuracy, whereas PCA is performed separately from any downstream task. We summarize our contributions as follows.

- (i) We present FVNNs for fair covariance-based model predictions with reduced influence from biased correlations. Our model uses fair covariance matrices from transformed data while explicitly promoting unbiased predictions in end-to-end learning.
- (ii) We theoretically show that the inherent stability of VNNs promotes equitable treatment of different groups.
- (iii) We empirically validate the stability and flexibility of FVNNs on one synthetic and three real datasets with known biases on both classification and regression tasks.

II. PROBLEM STATEMENT

Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i, z_i)\}_{i=1}^T$ of T tuples, each consisting of features $\mathbf{x}_i \in \mathbb{R}^N$, a target $y_i \in \mathcal{Y}$, and a group label $z_i \in \{1, \dots, G\}$ for every $i \in \{1, \dots, T\}$. Depending on the task at hand, $\mathcal{Y}$ can be a set of discrete class labels or real-valued regression targets. Each group $g \in \{1, \dots, G\}$ is associated with a random vector $\mathbf{x}^{(g)}$ with mean $\boldsymbol{\mu}_g = \mathbb{E}[\mathbf{x}^{(g)}]$ and covariance matrix $\mathbf{C}_g = \mathbb{E}[(\mathbf{x}^{(g)} - \boldsymbol{\mu}_g)(\mathbf{x}^{(g)} - \boldsymbol{\mu}_g)^\top]$. If $z_i = g$, then sample i belongs to group g , and the feature vector $\mathbf{x}_i$ is an instantiation of $\mathbf{x}^{(g)}$. Furthermore, let $\mathbf{Z} \in \{0, 1\}^{T \times G}$ be the indicator matrix denoting group membership, where $Z_{ig} = 1$ if and only if $z_i = g$.

Our goal is to learn a mapping $\Phi : \mathbb{R}^N \rightarrow \mathcal{Y}$ using covariance information to predict targets $\mathbf{y} = \{y_i\}_{i=1}^T$ from features $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_T]^\top \in \mathbb{R}^{T \times N}$ such that prediction performance is not biased with respect to group membership $\mathbf{z} = \{z_i\}_{i=1}^T$. With the observed samples, we estimate the mean $\hat{\boldsymbol{\mu}} = \frac{1}{T} \mathbf{X}^\top \mathbf{1}_T$ and covariance matrix $\hat{\mathbf{C}} = \frac{1}{T} (\mathbf{X} - \mathbf{1}_T \hat{\boldsymbol{\mu}}^\top)^\top (\mathbf{X} - \mathbf{1}_T \hat{\boldsymbol{\mu}}^\top)$. Analogously, we define a data matrix for each group $\mathbf{X}_g$ containing the T_g samples in group g with the corresponding sample mean $\hat{\boldsymbol{\mu}}_g$ and covariance $\hat{\mathbf{C}}_g$. Finally, we let $\mathbf{y}_g$ collect the targets corresponding to group g .

We view features as nodes in a graph whose connectivity is described by the covariance matrix $\mathbf{C}$, data samples $\mathbf{X}$ as graph signals, and we let our model Φ be a VNN, which performs graph convolutions on the signals [11], followed by a readout layer. More formally, a VNN architecture stacks L VNN layers, each consisting of a graph convolutional covariance filter bank of size $F_{\text{in}} \times F_{\text{out}}$ followed by a point-wise nonlinearity σ . The covariance filter and the propagation rule for each layer $l = 1, \dots, L$ and parallel filter $f = 1, \dots, F_{\text{out}}$ are defined as

$$\mathbf{H}(\mathbf{C}) = \sum_{k=0}^K h_k \mathbf{C}^k \quad \text{and} \quad \mathbf{x}_f^l = \sigma \left(\sum_{j=1}^{F_{\text{in}}} \mathbf{H}^l(\mathbf{C}) \mathbf{x}_j^{l-1} \right),$$

where $\{h_k\}_{k=0}^K$ denotes the set of learnable filter coefficients. Let $\Phi(\mathbf{x}, \hat{\mathbf{C}}, \mathcal{H})$ be the VNN architecture followed by a readout layer for the downstream task, where $\mathbf{x}$ is the input feature vector, $\hat{\mathbf{C}}$ the input covariance matrix, and $\mathcal{H}$ collects the filter coefficients for all layers.

We aim to learn parameters $\mathcal{H}$ from $\mathcal{D}$ to yield predictions $\hat{y} = \Phi(\mathbf{x}, \hat{\mathbf{C}}, \mathcal{H})$ that are fair with respect to the group label z . While equitable outcomes are desirable, balancing treatment of different groups is also a necessity [20]. Indeed, the downside of popular fairness definitions such as demographic parity (DP) [36] and equality of odds (EO) [37] is that they focus solely on outcomes without considering if the model exhibits preferential learning for certain groups [38], [39]. Inaccurate predictions for an underrepresented group can lead the model to consider those samples irrelevant for subsequent training. In this case, we may satisfy DP or EO at the cost of neglecting certain subpopulations. Thus, we emphasize equitable attention in training, a goal well-suited to VNNs, which are robust to insufficient data.

We formalize imbalanced treatment across groups as the difference in prediction performance between each pair of groups,

$$\Delta \mathcal{L}(\mathbf{X}, \mathbf{y}, \mathbf{z}) := \sum_{g=1}^G \sum_{h>g} |\mathcal{L}(\mathbf{X}_g, \mathbf{y}_g, \Phi) - \mathcal{L}(\mathbf{X}_h, \mathbf{y}_h, \Phi)|, \quad (1)$$

where $\mathcal{L}(\mathbf{X}_g, \mathbf{y}_g, \Phi)$ denotes the loss function measuring the performance of model Φ on the data samples of group g . Not only is $\Delta \mathcal{L}$ in (1) analogous to the goal of equal reconstruction error for fair PCA, but it aligns with other notions of fairness such as bounded group loss [40].

III. METHODOLOGY

We promote fairness for FVNN predictions in two ways. First, as with some PCA-based approaches [27], [31], we consider a fair version of the sample covariance matrix using fair data preprocessing techniques. Second, since VNNs can be trained end-to-end for a downstream task, we explicitly encourage unbiased predictions by penalizing biases in the loss function to be minimized, allowing control of the trade-off between fairness and accuracy during training.

A. Fair covariance matrices

FVNNs are a general framework that can accommodate any fair covariance estimation technique. We exemplify them with two in particular that promote different goals. First, for $G = 2$ with one group poorly represented, we may consider a balanced covariance matrix estimate [31]

$$\hat{\mathbf{C}}_{\text{bal}} = \alpha \hat{\mathbf{C}} + (1 - \alpha)(\hat{\mathbf{C}}_h - \hat{\mathbf{C}}_g) = \alpha_g \hat{\mathbf{C}}_g + \alpha_h \hat{\mathbf{C}}_h, \quad (2)$$

where $\hat{\mathbf{C}}_h$ is the sample covariance of the disadvantaged group and $\hat{\mathbf{C}}_g$ the other, $\alpha \in [0, 1]$ is the balancing term, $\alpha_g = (\alpha T_g / T + \alpha - 1)$ and $\alpha_h = (\alpha T_h / T + 1 - \alpha)$. The estimate $\hat{\mathbf{C}}_{\text{bal}}$ was defined for fair PCA to yield equal reconstruction error across groups [32] as it interpolates between the original sample covariance $\hat{\mathbf{C}}$ and the attempt to reduce the discrepancy between the minority and majority covariances $\hat{\mathbf{C}}_h$ and $\hat{\mathbf{C}}_g$, respectively. Alternatively, we may wish to remove the dependence on groups in the covariance matrix to avoid predictions that are based on sensitive attributes. To this end, for $\beta \in [0, 1]$ we present

$$\hat{\mathbf{C}}_{\text{deb}} = \mathbf{X}^\top (\mathbf{I}_T + \beta \mathbf{Z} \mathbf{Z}^\top)^{-1} \mathbf{X} / T, \quad (3)$$

which essentially computes the sample covariance matrix for transformed data $(\mathbf{I}_T + \beta \mathbf{Z} \mathbf{Z}^\top)^{-1/2} \mathbf{X}$. While $\hat{\mathbf{C}}_{\text{deb}}$ was originally proposed for group-agnostic PCA projections [27], we consider the transformed covariance $\hat{\mathbf{C}}_{\text{deb}}$ for FVNNs which, unlike PCA, exploit all PCs of the data. Note that $\hat{\mathbf{C}}_{\text{bal}}$ in (2) is only defined for two groups, but extensions to more than two groups are possible [30], while $\hat{\mathbf{C}}_{\text{deb}}$ in (3) applies to any number of groups G .

B. Bias mitigation penalties

In addition to providing fairer data information, we can also encourage unbiased behavior by manipulating the training loss. We can formulate the training objective as

$$\min_{\mathcal{H}} \gamma \mathcal{L}(\mathbf{X}, \mathbf{y}, \Phi) + (1 - \gamma) \mathcal{R}(\mathbf{X}, \mathbf{y}, \mathbf{z}, \Phi), \quad (4)$$

where Φ is the VNN using $\hat{\mathbf{C}}_{\text{bal}}$ or $\hat{\mathbf{C}}_{\text{deb}}$, $\mathcal{L}$ is the task-specific loss function, $\mathcal{R}$ denotes a bias metric, and $\gamma \in [0, 1]$ is a balancing weight to tune between $\mathcal{L}$ and $\mathcal{R}$. The penalty $\mathcal{R}$ measures group-wise imbalance when using $\mathbf{X}$ and Φ to predict $\mathbf{y}$. The choice of $\mathcal{R}$ is entirely flexible, including popular bias metrics DP or EO. In our case, we let $\mathcal{R}$ be the group-wise imbalance in accuracy $\Delta \mathcal{L}$ in (1).

IV. THEORETICAL ANALYSIS

Fair covariance estimation approaches such as $\hat{\mathbf{C}}_{\text{bal}}$ and $\hat{\mathbf{C}}_{\text{deb}}$ are notoriously unstable in small data regimes or when the eigenvalues of the covariance matrix are close, that is, finite sample estimation errors may lead to significant differences in the estimated fair covariance matrix and, consequently, in their PCs [7]. This can lead to unfair behavior when samples from different groups are unbalanced or distributions differ across groups. In this context, the stability of VNNs to covariance estimation errors intrinsically improves fairness, as we discuss in the following theorem.

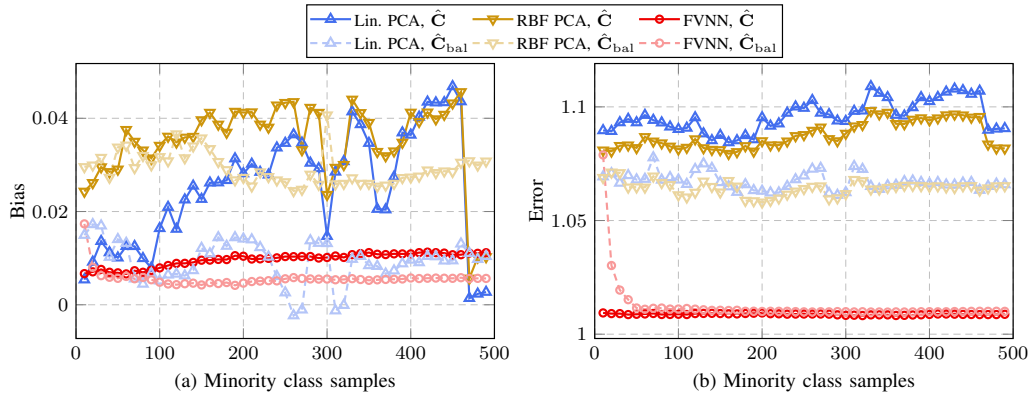


Fig. 1: Performance of PCA-based and FVNN models for a synthetic regression task. Each model is compared using the sample covariance $\hat{\mathbf{C}}$ and the balanced covariance $\hat{\mathbf{C}}_{\text{bal}}$. (a) Fairness measured as imbalance in sMAPE across groups. (b) Error measured as sMAPE. The legend is shared by both plots.

Definition 1 Let $h(\lambda)$ be the frequency response of the filter $\mathbf{H}(\mathbf{C})$, which evaluates the behavior of the filter in the spectral domain at eigenvalues λ of the covariance matrix [11]. The filter $\mathbf{H}(\mathbf{C})$ is Lipschitz with constant P if $|h(\lambda_i) - h(\lambda_j)| \leq P|\lambda_i - \lambda_j|$ for every eigenvalue pair $\lambda_i, \lambda_j, i \neq j$.

Theorem 1 Consider a covariance filter $\mathbf{H}(\mathbf{C})$ that is Lipschitz with constant P as per Def. 1.

First, consider two groups with true covariances $\mathbf{C}_g, \mathbf{C}_h$. We consider the fair covariance estimate $\hat{\mathbf{C}}_{\text{bal}}$ from (2) with T samples and we assume the fair true covariance to be $\mathbf{C} = \alpha_g \mathbf{C}_g + \alpha_h \mathbf{C}_h$ with α_g, α_h in (2). We express VNN stability as

$$\|\mathbf{H}(\mathbf{C}) - \mathbf{H}(\hat{\mathbf{C}}_{\text{bal}})\| \leq P\sqrt{N + 2N^2} \left(\mathcal{O}(T_g^{-1/2}) + \mathcal{O}(T_h^{-1/2}) \right). \quad (5)$$

Second, let $\mathbf{C}$ be any covariance matrix, but the observed data $\mathbf{X}$ is biased with covariance $\mathbb{E}[\mathbf{X}^\top (\mathbf{I}_T + \beta \mathbf{Z} \mathbf{Z}^\top) \mathbf{X}]$, where we let the mean $\boldsymbol{\mu}$ be zero for simplicity. Then, given the fair covariance estimate $\hat{\mathbf{C}}_{\text{deb}}$ in (3) using T samples, we write VNN stability as

$$\|\mathbf{H}(\mathbf{C}) - \mathbf{H}(\hat{\mathbf{C}}_{\text{deb}})\| \leq P\sqrt{N + 2N^2} \mathcal{O}(T^{-1/2}). \quad (6)$$

Proof: Following [14, Appendix C, Proposition 2], for two generic true and sample covariances $\mathbf{C}, \hat{\mathbf{C}}$, we can write

$$\|\mathbf{H}(\mathbf{C}) - \mathbf{H}(\hat{\mathbf{C}})\| \leq P\sqrt{N + 2N^2} \|\mathbf{E}\| + \mathcal{O}(\|\mathbf{E}\|^2), \quad (7)$$

where $\mathbf{E} = \mathbf{C} - \hat{\mathbf{C}}$ and $\|\mathbf{E}\|^2$ is negligible for T large enough.

For the first case, by the triangle inequality, we have

$$\|\hat{\mathbf{C}}_{\text{bal}} - \mathbf{C}\| \leq |\alpha_h| \|\hat{\mathbf{C}}_h - \mathbf{C}_h\| + |\alpha_g| \|\hat{\mathbf{C}}_g - \mathbf{C}_g\|, \quad (8)$$

where each term in the norm is the estimation error of a covariance matrix, which decreases as the inverse square root of the number of samples, that is, $\|\hat{\mathbf{C}}_{\text{bal}} - \mathbf{C}\| \leq \mathcal{O}(T_g^{-1/2}) + \mathcal{O}(T_h^{-1/2})$ with high probability [41, Theorem 5.6.1]. Replacing this in (7), we obtain the bound for $\hat{\mathbf{C}}_{\text{bal}}$.

For the second case, observe that the true covariance matrix $\mathbf{C}$ generates the unbiased samples $\mathbf{X}' = (\mathbf{I}_T + \beta \mathbf{Z} \mathbf{Z}^\top)^{-1/2} \mathbf{X}$, which are the transformed samples used to estimate $\hat{\mathbf{C}}_{\text{deb}}$. In this case, $\hat{\mathbf{C}}_{\text{deb}}$ is the classic sample covariance estimator for the samples $\mathbf{X}'$, so we have that $\|\hat{\mathbf{C}}_{\text{deb}} - \mathbf{C}\| \leq \mathcal{O}(T^{-1/2})$ [41, Theorem 5.6.1]. Combining this with (7) leads to the bound for $\hat{\mathbf{C}}_{\text{deb}}$. ■

Theorem 1 shows that a covariance filter $\mathbf{H}(\mathbf{C})$ and thus a VNN [11, Theorem 3] operating on a fair sample covariance estimate is stable to estimation errors. In particular, we may design the filter $\mathbf{H}(\mathbf{C})$ while considering the Lipschitz constant P , allowing us to control the tradeoff between stability and discriminability. On the contrary, PCA does not have this flexibility since its stability depends inversely on the smallest gap in covariance eigenvalues [13, Proposition 1]. Covariance estimation error may differ across groups if one group has fewer samples, but Theorem 1 shows that FVNNs achieve a more consistent behavior across groups compared to PCA and therefore intrinsically provide superior fairness.

V. NUMERICAL EVALUATION

A. Synthetic data

Experimental setup. To validate the impact of FVNN stability on fairness, we generate a synthetic dataset with $G = 2$ groups, where features are sampled corresponding to two different multivariate Gaussian distributions with different covariance matrices $\mathbf{C}_1$ and $\mathbf{C}_2$. We generate a regression target following the Friedman regression problem [42]. We let group 1 be the disadvantaged group, that is, the eigenvalues for $\mathbf{C}_1$ are closer than those of $\mathbf{C}_2$, rendering the estimation of $\mathbf{C}_1$ more difficult. While the training data is balanced between groups, in testing, we replace the sample covariance matrix used in training with an estimate obtained from testing data as the number of samples T_1 in group 1 increases from 1 to 500, while we fix $T_2 = 500$. We compute the test performance on the entire test set, that is, with $T_1 = T_2 = 500$. We compare FVNN performance, denoted “FVNN”, to SVM regression using PCA-projected features, where we apply both linear SVM, denoted “Lin. PCA”, and kernelized SVM, denoted “RBF PCA”. Moreover, for each method we apply the original sample covariance $\hat{\mathbf{C}}$ and the balanced covariance estimate $\hat{\mathbf{C}}_{\text{bal}}$ with $\alpha = 0.5$. We measure error $\mathcal{L}$ as the symmetric mean average percentage error (sMAPE), while the bias is measured as $\Delta \mathcal{L}$ as in (1).

Discussion. Fig. 1 shows that VNNs are significantly more stable both in terms of fairness and prediction performance compared to PCA-based variants, which are more susceptible to changes in the covariance matrix. Indeed, even a small number of new samples can yield significant changes in bias and error for Lin. PCA and RBF PCA, while FVNN returns smooth estimates as the sample ratio between groups varies. Thus, VNNs are more reliable for fair learning under noisy covariance estimates, particularly when one

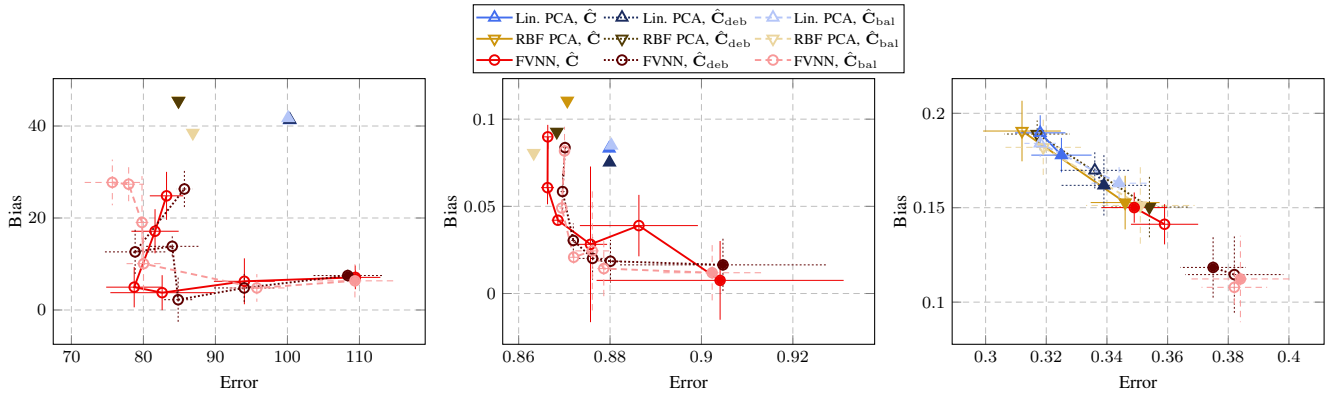


Fig. 2: Performance of PCA-based and FVNN models for real-world regression and classification tasks. For each plot, the y -axis denotes bias and the x -axis error. (a) Parkinson regression as the bias penalty weight γ increases. Results for PCA with $\hat{C}_{deb}$ are overlapped with those with $\hat{C}$, which are therefore not visible. (b) LSAC regression as the bias penalty weight γ increases. (c) German Credit classification. **FVNN** is shown with and without a bias penalty, while PCA-based models are shown with 10 and 30 PCs. The legend is shared by all three plots.

group is more difficult to estimate. Moreover, we not only observe improved bias in Fig. 1a when using $\hat{C}_{bal}$ in place of $\hat{C}$, but VNNs also outperform PCA-based models in terms of error with either covariance matrix estimate in Fig. 1b.

B. Real data - regression

Experimental setup. Next, we apply FVNNs for regression tasks using two real datasets with known biases. Parkinson [43] contains 5,875 records of 23 features for 42 patients with early-stage Parkinson’s disease. The objective is to predict, for each record, the *Unified Parkinson’s Disease Rating Scale score*, a continuous value measuring different aspects of Parkinson’s disease. The sensitive attribute is the sex of the patient (female 33%, male 67%). Law School Admission Council (LSAC) [44] contains 5 features for 22,407 law school students (22,368 without missing data). The target is the *Grade Point Average* and we use as sensitive attribute the race of students (white/Caucasian 88.2%, other 11.8%).

We again compare **FVNN** to **Lin. PCA** and **RBF PCA**, which respectively use linear and kernelized SVM for regression. For all three methods, we compare the sample covariance $\hat{C}$ with the fair estimates $\hat{C}_{bal}$ and $\hat{C}_{deb}$, where we select α and β through a grid search, along with the VNN size and number of PCs. Furthermore, we employ the bias penalty $\mathcal{R}$ as in (1) for the training loss in (4), where $\mathcal{L}$ denotes the mean squared error (MSE). We vary the penalty weight from $\gamma = 0.3$ (filled markers) to $\gamma = 1$. Figs. 2a and b show the average results and standard deviation over 5 trials.

Discussion. We observe that **FVNN** provides a significantly more flexible tool to control the fairness-accuracy tradeoff compared to the PCA-based models. Smaller values of γ lead to fairer solutions for **FVNN** at the expense of a higher regression error. For large enough γ , **FVNN** with any covariance matrix outperforms **Lin. PCA** and **RBF PCA** in both fairness and accuracy. Applying $\hat{C}_{bal}$ improves bias for **RBF PCA** but not **Lin. PCA**, while $\hat{C}_{deb}$ can decrease bias for both PCA methods on LSAC. This shows that fair covariance estimators may not lead to improved fairness for downstream tasks, calling for more flexible and powerful solutions. For our **FVNN** approach, the fair covariance matrices yield minor improvements in bias for LSAC and negligible differences for Parkinson. Thus, we show that real-world datasets may contain biases that cannot be reduced by data preprocessing, but FVNNs offer an effective approach.

C. Real data - classification

Experimental setup. Finally, we consider a classification task using a real-world German Credit dataset of 1000 individuals. The goal is to predict credit score as either good or bad for each individual given a set of 46 features, with sex (female 31%, male 69%) as the sensitive attribute. We compare **FVNN**, **Lin. PCA**, and **RBF PCA** for $\hat{C}$, $\hat{C}_{deb}$, and $\hat{C}_{bal}$ over 4 train-test splits. We again consider linear and kernelized SVM for classification. Fig. 2c shows the average performance in terms of error, that is, one minus accuracy, and imbalance in error between groups. For the bias penalty weight, we consider $\gamma = 1$, that is, with no penalty $\mathcal{R}$, and $\gamma = 0.25$, and we also show the number of PCs as 10 and 30, where the markers for $\gamma = 0$ and 10 PCs are filled in Fig. 2.

Discussion. All PCA-based methods increase bias while reducing error when more PCs are considered. In contrast, when we increase the influence of $\mathcal{R}$ by decreasing $\gamma = 1$ to $\gamma = 0.25$, **FVNN** improves the bias at the cost of increased error. Importantly, while we see less effect due to $\mathcal{R}$, we note that applying either fair covariance estimate $\hat{C}_{bal}$ or $\hat{C}_{deb}$ yields greater improvements on bias compared to regression. This shows that biases in real-world data may not be reduced by a single method, but our proposed **FVNN** model provides the flexibility to address multiple kinds of bias such as imbalanced representations or unfair outcomes.

VI. CONCLUSION

In this work, we proposed Fair coVariance Neural Networks (FVNNs), a fairness-aware graph convolutional neural network that operates on the covariance matrix of the data. FVNNs promote fairness in two ways: by employing a fair covariance matrix to remove biases in data before training and by adding a regularization term in the loss to penalize unfair performance across groups. We theoretically showed that FVNNs are intrinsically fairer than fair PCA techniques by building a connection between VNN stability and VNN performance on groups with different distributions. Furthermore, we empirically validated the efficiency of FVNNs in managing the tradeoff between prediction performance and fairness for multiple applications, showing them to be a significantly more flexible approach than fair PCA. In future work, we will expand on the effects of biased data and fair interventions for VNN performance beyond stability. Furthermore, we will address additional notions of fairness beyond balancing performance, such as DP and EO.

REFERENCES

- [1] A. Bessadok, M. A. Mahjoub, and I. Rekik, "Graph neural networks in network neuroscience," *IEEE Trans. Pattern Analysis and Mach. Intell.*, vol. 45, no. 5, pp. 5833–5848, 2022.
- [2] L. Qiao, H. Zhang, M. Kim, S. Teng, L. Zhang, and D. Shen, "Estimating functional brain networks by incorporating a modularity prior," *NeuroImage*, vol. 141, pp. 399–407, 2016.
- [3] A. Belouchrani, K. Abed-Meraim, J.-F. Cardoso, and E. Moulines, "A blind source separation technique using second-order statistics," *IEEE Trans. Signal Process.*, vol. 45, no. 2, pp. 434–444, 1997.
- [4] J. Nikunen and T. Virtanen, "Direction of arrival based spatial covariance model for blind sound source separation," *IEEE/ACM Trans. Audio, Speech, and Language Process.*, vol. 22, no. 3, pp. 727–739, 2014.
- [5] J. V. de Miranda Cardoso, J. Ying, and D. P. Palomar, "Algorithms for learning graphs in financial markets," *arXiv preprint arXiv:2012.15410*, 2020.
- [6] Y. Wang and T. Aste, "Network filtering of spatial-temporal GNN for multivariate time-series prediction," in *Intl. Conf. on AI in Finance*, p. 463–470, Association for Computing Machinery, 2022.
- [7] I. T. Jolliffe, *Principal Component Analysis*. Springer Verlag, 2002.
- [8] E. Isufi, F. Gama, D. I. Shuman, and S. Segarra, "Graph filters for signal processing and machine learning on graphs," *IEEE Trans. Signal Process.*, 2024.
- [9] F. Gama, E. Isufi, G. Leus, and A. Ribeiro, "Graphs, convolutions, and neural networks: From graph filters to graph neural networks," *IEEE Signal Process. Mag.*, vol. 37, no. 6, pp. 128–138, 2020.
- [10] Z. Gao, A. Prorok, and E. Isufi, "On the trade-off between stability and representational capacity in graph neural networks," *arXiv preprint arXiv:2312.02372*, 2023.
- [11] S. Sihag, G. Mateos, C. McMillan, and A. Ribeiro, "coVariance neural networks," *Advances in Neural Info. Process. Syst.*, vol. 35, pp. 17003–17016, 2022.
- [12] S. Sihag, G. Mateos, C. McMillan, and A. Ribeiro, "Transferability of covariance neural networks," *IEEE J. Sel. Topics Signal Process.*, vol. 18, no. 2, pp. 199–215, 2024.
- [13] A. Cavallo, M. Sabbaghi, and E. Isufi, "Spatiotemporal covariance neural networks," in *Joint European Conf. on Mach. Learn. and Knowledge Discovery in Databases*, pp. 18–34, Springer Nature Switzerland, 2024.
- [14] A. Cavallo, Z. Gao, and E. Isufi, "Sparse covariance neural networks," *arXiv preprint arXiv:2410.01669*, 2024.
- [15] S. Sihag, G. Mateos, C. McMillan, and A. Ribeiro, "Explainable brain age prediction using covariance neural networks," *Advances in Neural Info. Process. Syst.*, vol. 36, 2024.
- [16] S. Sihag, G. Mateos, and A. Ribeiro, "Towards a foundation model for brain age prediction using covariance neural networks," *arXiv preprint arXiv:2402.07684*, 2024.
- [17] S. Sihag, G. Mateos, and A. Ribeiro, "Explainable brain age gap prediction in neurodegenerative conditions using covariance neural networks," *arXiv preprint arXiv:2501.01510*, 2025.
- [18] A. Choudhchova, "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- [19] M. Olfat and A. Aswani, "Convex formulations for fair principal component analysis," *AAAI Conf. on Artif. Intell.*, vol. 33, no. 01, pp. 663–670, 2019.
- [20] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala, "The price of fair PCA: One extra dimension," in *Advances in Neural Info. Process. Syst.*, vol. 31, 2018.
- [21] M. Navarro, C. Little, G. I. Allen, and S. Segarra, "Data augmentation via subgroup mixup for improving fairness," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 7350–7354, 2024.
- [22] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *Intl. Conf. on Mach. Learn. (ICML)*, vol. 28 of *Proceedings of Machine Learning Research*, pp. 325–333, PMLR, 2013.
- [23] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Info. Syst.*, vol. 33, no. 1, pp. 1–33, 2012.
- [24] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2022.
- [25] S. Caton and C. Haas, "Fairness in machine learning: A survey," *ACM Computing Surveys*, vol. 56, no. 7, pp. 1–38, 2024.
- [26] J. Liu, Z. Li, Y. Yao, F. Xu, X. Ma, M. Xu, and H. Tong, "Fair representation learning: An alternative to mutual information," in *Intl. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, p. 1088–1097, 2022.
- [27] O. D. Kose and Y. Shen, "Fairness-aware dimensionality reduction," in *European Signal Process. Conf. (EUSIPCO)*, pp. 660–664, IEEE, 2023.
- [28] Z. Tan, S. Yeom, M. Fredrikson, and A. Talwalkar, "Learning fair representations for kernel models," in *Intl. Conf. Artif. Intell. Stat. (AISTATS)*, pp. 155–166, PMLR, 2020.
- [29] U. Tantipongpipat, S. Samadi, M. Singh, J. H. Morgenstern, and S. Vempala, "Multi-criteria dimensionality reduction with applications to fairness," in *Advances in Neural Info. Process. Syst.*, vol. 32, 2019.
- [30] M. M. Kamani, F. Haddadpour, R. Forsati, and M. Mahdavi, "Efficient fair principal component analysis," *Mach. Learn.*, vol. 111, no. 10, p. 3671–3702, 2022.
- [31] G. D. Pelegrina and L. T. Duarte, "A novel approach for fair principal component analysis based on eigendecomposition," *IEEE Trans. Artif. Intell.*, vol. 5, no. 3, pp. 1195–1206, 2024.
- [32] G. D. Pelegrina, R. D. Broto, L. T. Duarte, R. Attux, and J. M. Romano, "Analysis of trade-offs in fair principal component analysis based on multi-objective optimization," in *Intl. Joint Conf. on Neural Networks (IJCNN)*, pp. 1–8, IEEE, 2022.
- [33] J. Lee, G. Kim, M. Olfat, M. Hasegawa-Johnson, and C. D. Yoo, "Fast and efficient MMD-based fair PCA via optimization over Stiefel manifold," *AAAI Conf. on Artif. Intell.*, vol. 36, no. 7, pp. 7363–7371, 2022.
- [34] M. Kleindessner, M. Donini, C. Russell, and M. B. Zafar, "Efficient fair PCA for fair representation learning," in *Intl. Conf. Artif. Intell. Stat. (AISTATS)*, vol. 206 of *Proceedings of Machine Learning Research*, pp. 5250–5270, PMLR, 2023.
- [35] M. Xu, B. Jiang, W. Pu, Y.-F. Liu, and A. M.-C. So, "An efficient alternating Riemannian/projected gradient descent ascent algorithm for fair principal component analysis," in *IEEE Intl. Conf. Acoust., Speech and Signal Process. (ICASSP)*, pp. 7195–7199, IEEE, 2024.
- [36] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *Intl. Conf. on Knowledge Discovery and Data Mining (SIGKDD)*, pp. 259–268, 2015.
- [37] M. Hardt, E. Price, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *Advances in Neural Info. Process. Syst.*, vol. 29, 2016.
- [38] B. R. Baer, D. E. Gilbert, and M. T. Wells, "Fairness criteria through the lens of directed acyclic graphical models," *arXiv preprint arXiv:1906.11333*, 2019.
- [39] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," *arXiv preprint arXiv:1609.05807*, 2016.
- [40] A. Agarwal, M. Dudik, and Z. S. Wu, "Fair regression: Quantitative definitions and reduction-based algorithms," in *Intl. Conf. on Mach. Learn. (ICML)*, pp. 120–129, PMLR, 2019.
- [41] R. Vershynin, *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2018.
- [42] J. Friedman, "Multivariate adaptive regression splines," *The Ann. of Stat.*, vol. 19, no. 1, pp. 1–67, 1991.
- [43] A. Tsanas, M. Little, P. McSharry, and L. Ramig, "Accurate telemonitoring of Parkinson's disease progression by non-invasive speech tests," *Nature Precedings*, pp. 1–1, 2009.
- [44] L. F. Wightman, "LSAC national longitudinal bar passage study. LSAC research report series," 1998.